Cisek, P. (1999) Beyond the computer metaphor: Behaviour as interaction. *Journal of Consciousness Studies,* **6**, No. 11-12, pp. 125-42.

BEYOND THE COMPUTER METAPHOR: BEHAVIOR AS INTERACTION

Paul Cisek

Dept. de Physiologie

Université de Montréal

Abstract

Behavior is often described as the computation of a response to a stimulus. This description is incomplete in an important way because it only examines what occurs between the reception of stimulus information and the generation of an action. Behavior is more correctly described as a control process where actions are performed in order to affect perceptions. This closed-loop nature of behavior is de-emphasized in modern discussions of brain function, leading to a number of artificial mysteries. A notable example is the "symbol grounding problem". When behavior is viewed as a control process, it is natural to explain how internal representations, even symbols, can have meaning for an organism, and how actions can be motivated by organic needs.

What's wrong with "computationalism"?

One can say that the general working assumption in brain science today is that ***the function of the brain is to convert stimuli into reactions***. This was explicitly stated over a hundred years ago by one of the most influential psychologists of all time, William James:

> *The whole neural organism, it will be remembered, is, physiologically considered, but a machine for converting stimuli into reactions (James, 1890, p. 372).*

The modern interpretation of this statement takes the form of what may be called the computer metaphor, or "computationalism"[1]. This doctrine describes brain function as the computation of behavioral responses from internal representations of stimuli and stored representations of past experience. In its broadest sense, computationalism may be crudely defined in terms of the following analogy: perception is like input, action is like output, and all the things in-between are like the information processing performed by computers.

Below, I argue that this description is incomplete in an important way, leading brain sciences toward apparent mysteries where none actually exist. This is not to say that computationalism is false, it is merely incomplete and can be easily extended toward a more productive description of brain function without giving up many of its accompanying concepts. Before I suggest how this can be done, I will first briefly look at some aspects of the history of the computational analogy. Where does it come from and how has it come to be the dominant viewpoint in neuroscience, psychology, and philosophy of mind?

I believe that the computer metaphor for the mind earned its popularity by providing convergent answers to several major questions which confronted psychology during the first half of the 20th Century. Below I identify five such questions. Most of these stem directly from the theological foundations of philosophy, and specifically, from the belief in a distinction between the body and the soul.

The belief in a non-physical soul that is separate from the body predates history, and has been the fundamental assumption underlying the vast majority of human philosophical thought. Today, most psychologists do not believe in a soul, but they work within scientific traditions which grew up in the context of that belief. The influence of this heritage can still be felt in modern psychology, carried within its jargon and even within its academic taxonomy. The 17th Century philosophical foundations of psychology held mind-body duality (derived from soul-body duality) as a central theme. This led to three of the questions which I want to discuss.

First, mind-body duality forced an architecture for discussing issues of behavior. The assumption of a non-physical *Mind* compelled philosophers to conceive of two interfaces between it and the world: *Perception* - which presented

---

[1] The term "computationalism" may be used to imply any one of a number of theoretical viewpoints which differ on many important issues. A distinction between symbolic computationalism and connectionist computationalism is an example. Here, I will use the term in the broadest sense, referring to any theory which describes brain function as the computation of a response to a stimulus.

the world to the mind; and *Action* - which played out the wishes of the mind upon the world. René Descartes (1596-1650) described these two processes as completely mechanistic, and in animals it was assumed that they bore the sole responsibility for producing even complex behavior. In Man, however, there existed a non-physical *Mind* which linked these two processes, allowing for free will, rational thought, and consciousness.

When dualism was eventually rejected, the concept of the non-physical *Mind* was replaced with a mechanistic concept of *Cognition*, but the shape of the architecture remained, leading to the *Perception-Cognition-Action* model of behavior (sometimes called the "sense-think-act" model). This model established the basic taxonomy within which psychologists classify their work – nearly every question in psychology is immediately labeled either as a perceptual, a cognitive, or a motor control question. For those interested in higher mental functions, the most fundamental question is how Cognition operates, i.e. how it converts perceptions into action plans (Question #1).

Second, mind-body duality led to the mind-body problem and to a series of movements vying to provide its solution. Although the mind was initially seen as separate from the physical world, many insisted that it could still be studied scientifically. Psychology as a science was founded upon this assumption, primarily through the "introspective" method of Wilhelm Wundt (1832-1920). A series of reactionary movements ensued. First, the introspective method was taken to an extreme by Titchener's *structuralism* – an attempt to discover the elements of consciousness through rigorous introspection by trained human adults. This approach came under a great deal of criticism, the most successful being Watson's *behaviorism* (Watson, 1913) – a rejection of all notions of internal states and an exclusive focus on observable data[2]. This went too far the other way, however, and psychology was left with a desire for a return to concepts of internal states but without resorting to dualism or to a method of pure introspection (Question #2).

An alternative approach to the mind-body problem was the idea that mental states are processes, that they are functional states of the brain. This view is now called *functionalism*, and it has gained great prominence in recent decades. Before functionalism could be truly accepted, however, it had to explain how can it be that mental phenomena are so qualitatively different than the physical brain phenomena which presumably make them happen (Question #3).

Two other issues deserve some discussion. During the 19th Century, studies of living organisms became separated into those addressing issues on the level of behavior and those addressing issues of bodily physiology. This was done for very practical reasons of scientific specialization, but it led to the growth of a huge conceptual gap between knowledge obtained within psychology and knowledge obtained through physiological studies of the nervous system. For over a hundred years, there was precious little interaction between the two. In the last few decades, a concerted effort had begun to bring these two disparate fields back together toward a unified science of behavior,

---

[2] There is a tendency to assume all "behaviorism" to be so extreme. This is not true – many to whom that label is applied did in fact discuss and study internal phenomena.

but by then the conceptual gap was very great. It seemed difficult to say how psychological phenomena could be explained with biological elements (Question #4).

Finally, the field of psychology can be said to have suffered for a long time from a kind of "physics envy". Many attempts at establishing a science of mind, starting with John Locke in the 17th Century, strove to discuss mental operation with the same kind of mathematical rigor that was found in physics. There was and is still a widespread attitude that any field aspiring to the status of a "real science" needs to develop a precise formalism for expressing its concepts. Psychology had always struggled in this regard, and has often been criticized by scientists from other fields. There was a great desire for a formalism, preferably a mathematical one, which could capture psychological phenomena (Question #5).

In the early part of the 20th Century, while psychology faced these five open questions (among others), several new concepts had emerged in seemingly unrelated fields. First, Alan Turing's pioneering work in machine theory resulted in a formal definition of "computation": According to Turing (1936), all computation is formally equivalent to the manipulation of symbols in a temporary buffer. Second, research aimed at the development of more efficient telephone communication resulted in a formal definition of "information": According to Shannon & Weaver (1949), the informational content of a signal is inversely related to the probability of that signal arising from randomness. These developments, along with others, launched computer science into what has become one of the most significant technological advancements of modern times.

As computers quickly grew in complexity and in functional sophistication, the potential similarity these machines had to the brain began to be widely recognized. Both received information from their external environment, and both acted upon this information in complex ways. Digital computers suddenly joined human brains as the only examples of systems capable of complex reasoning. The analogy between computers and brains was irresistible. Most importantly, once the brain was thought of as analogous to a computer, all of the five questions discussed above suddenly had answers:

First, the computer metaphor provided a candidate mechanism for how Cognition operates – it operates like a digital computer program, by manipulating internal representations according to some set of rules[3]. Second, it allowed discussion of non-dualist internal states – e.g. memories and temporary variables. Third, it provided an inspiring metaphor for functionalism – mental entities are like software while physical mechanisms are like hardware. This same metaphor also provided a quick way to bridge the gap between psychology and biology – psychological phenomena are the software running upon the biological hardware. Finally, it provided a mathematical formalism that gave psychology some long-desired rigor – the language of predicate logic and information theory.

All of these answers arrived on the scene within a short time of one another, carried along with the central idea that the brain is like a computer. Because this metaphor provided so many timely solutions, it was very eagerly embraced

[3] Connectionism is no exception here, albeit its representations are distributed and its rules are encoded in weight matrices and learning laws.

and tenaciously defended. The computer metaphor for the mind quickly found its way to become the official foundation of modern psychology.

Since digital computers were the only systems capable of complex reasoning whose operation was understood, it was not known which conclusions about their operation should be carried over to theories of the brain and which should not. At first, the analogy was taken quite literally, and notions of symbolic computation and serial processing were seen as inseparable from the concept of functionalism, and necessary ingredients to constructing any human-like intelligence. Meanwhile, the formal definitions of information used in computer science and communications technology found widespread use in psychological theory and practice as well. An example is the work of Paul Fitts, a psychophysicist who quantified the accuracy of human movements in terms of the bandwidth of the channels between perception and action (Fitts, 1954).

In recent decades, computationalism has become the basic axiom of most large-scale brain theories and the language in which students entering brain-related fields are taught to phrase their questions[4]. The task of brain science is often equated to answering the question of how the brain computes. Many debates remain, of course, about such issues as serial vs. parallel processing, analog vs. digital coding, and symbolic vs. non-symbolic representations, but these are all debated within the accepted metaphor: perception = input, action = output, and cognition = computation.

Despite its popularity, certain problems have plagued computationalism from the very beginning. Notable among these are questions of consciousness, emotion, motivation, and meaning. In this article I focus on the question of meaning. This, I will argue, is not a problem for the brain to solve through some dedicated "meaning assignment" mechanism. Instead, it is only a problem with our description of the brain – a symptom of the shortcomings of computationalism[5], and an argument for going beyond it.

The question of meaning is a central problem in the philosophy of mind. If the brain is doing computation, defined as a transformation of one representation into another, how does the brain know what these representations mean? To illustrate the problem, an analogy to computers is usually employed. For example, suppose there is a program which takes as input simple queries written in English and responds to them with either "yes", "no", or "don't know", based on a large database of facts. Such programs can be written today, and with sophisticated front-end parsing they may

---

[4] In time, criticisms of the analysis of behavior in terms of information processing came to be perceived as attacks upon the science of psychology as a whole (Still & Costall, 1991)! I hope that the reader will recognize that the modifications I recommend are not an attempt to disparage psychological science or to bring back behaviorism, but an attempt to move beyond some premature and limiting assumptions of the computer metaphor.

[5] The riddle of meaning is at least in part a symptom of a particularly inappropriate definition of "information" used by most psychologists and philosophers – the definition given by Shannon & Weaver (1949). It is an important irony that Shannon and Weaver were working on improving the transmission of information on telephone lines, and were openly not concerned with the semantic content of these transmissions. However, because no other definitions were widely recognized at the time, theirs became almost universally accepted as *the* definition of information. See Mingers (1996) for a review of a number of alternate definitions which attempt to bring semantics back into a theory of information.

give the illusion of understanding the questions. But the programmer, and anyone who knows how the machine works, will insist that the machine does not *understand* the queries – it only searches for keywords, analyzes tenses, applies prepared rules of response, etc. Let's suppose that in the future similar systems may be built, with larger databases and better parsers, that can answer more complex questions. Will they be able to understand? It seems not. It seems we'll just have more and more heuristics and programming tricks, but no actual understanding.

This apparent inability of computers to grasp the semantics of the symbols they manipulate has been discussed under a number of labels. The most famous example is Searle's Chinese Room Argument (Searle, 1980); but it is also known as "intrinsic meaning" and "the problem of intentionality" (Dennett, 1978). It underlies the so-called "frame problem" in artificial intelligence (McCarthy & Hayes, 1969). Stephan Harnad (1990) calls it "the symbol grounding problem". That is the label I will use here because I find his presentation of the problem to be the most clear. The question, as posed by Harnad, goes as follows: *"How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?"* (Harnad, 1990, p. 335)

Harnad suggests that the issue lies in symbolic vs. non-symbolic representations. He suggests that a symbol system, defined as a system which manipulates arbitrary tokens according to a set of explicit rules, is purely syntactic and thus cannot capture meaning. According to Harnad, the problem lies in the arbitrary nature of the assignment between the symbolic tokens and the objects or states of affairs that these tokens stand for. To ground the symbols, he proposes a hybrid system where the representations of the symbolic system are linked to two kinds of non-symbolic representations: *icons* which are analogs of the sensation patterns, and *categorical representations* which capture the invariant features of these icons. The fundamental symbols are arbitrary tokens assigned to the non-arbitrary patterns of the icons and categories, and higher-level symbols are composites of these. For example, the word "horse" is linked to all the images and all the categorical representations involved in the perception of horses, thereby being grounded.

In summary, Harnad is proposing that we solve the symbol grounding problem by backing up out of the premature analogy, made during the beginnings of Artificial Intelligence, that all thought is like symbolic logic. Though I believe that this is moving in the right direction, I suggest that we need to back out further. We need to step back all the way out of the computer metaphor and to consider whether there is a better alternative description of what it is that the brain does.

In the following section, I outline an alternative metaphor for describing the function of the brain. Those who believe that "information-processing" already captures this function adequately might question the utility of searching for an alternative. I ask these readers to bear with me.

Behavior as control

In an attempt to develop a new metaphor, we must first break free from the preconceptions that our current one forces us into. This is not easy, but it may be possible if we step back from modern philosophical debates for a moment and

consider issues which at first might appear unrelated. This lets us develop a discussion that is not filtered through the lens of the current metaphor. Later, we can return toward the problems of interest and examine them from a novel perspective. Once this has been done, the reader may decide which viewpoint offers a more parsimonious account of the phenomena that both try to explain.

We begin with a fundamental premise: *The brain evolved*. This is accepted as fact by almost everyone, but its implications for philosophy are rarely acknowledged. The evolution of a biological system such as the brain is not merely a source of riddles for biologists to ponder. It is also a rich source of constraints for anyone theorizing about how the brain functions and about what it does. It is a source of insight that is too often overlooked[6].

An evolutionarily sound theory of brain function is not merely one which explains the selective advantage offered by some proposed brain mechanism. Lots of mechanisms may be advantageous. What is more useful toward the development and evaluation of brain theories is a plausible story of how a given mechanism may have evolved through a sequence of functional elaborations. The consideration of such a sequence offers powerful guidance toward the formulation of a theory, because the phylogenetic heritage of a species greatly constrains the kinds of mechanisms that may have evolved. Therefore, we should expect to gain insight into the abilities of modern brains by considering the requirements faced by primitive brains, and the sequence of evolutionary changes by which these primitive brains evolved into modern brains.

A contemplation of the most fundamental functional structure of behavior can start all the way back at the humble beginnings of life. The earliest entities deserving of the term "living" were self-sustaining chemical systems called "autocatalytic sets". There are various theories of how these systems came into existence (Eigen & Schuster, 1979; Kauffman, 1993), developed the genetic code (Crick, Brenner, Klug, & Pieczenik, 1976; Bedian, 1982), and enfolded themselves with membranes (Fox, 1965). Much of the story of how organic molecules organized into cells is not understood. However, although they differ on many important issues, all theories of early life agree that living systems took an active part in ensuring that the conditions required for their proper operation were met. This means that any changes in critical variables such as nutrient concentration, temperature, pH, etc. have to be corrected and brought back within an acceptable range. This is a fundamental task for any living system if it is to remain living. Mechanisms which keep variables within a certain range are usually called "homeostatic" mechanisms.

Biochemical homeostasis often works through chemical reactions where the compounds whose concentration is to be controlled affect their own rates of production and/or breakdown. For example, compound A might be a catalyst

---

[6] In fact, it is often argued that an account of evolution should be secondary: "any reasonable way to go about finding out how a mechanism evolved would be first to find out how the mechanism works, and then worry about how it evolved" (Crick, 1994). I respectfully disagree. Mechanisms generated by evolution are products of a long sequence of modifications and elaborations, all of these performed within living organisms. Because the modified organisms have to continue living, evolution does not have full freedom to redesign their internal mechanisms. Consequently, the modern form of these mechanisms is strongly constrained by their ancestral forms. Our theories should be similarly constrained. Therefore, an understanding of the fundamental architecture of the brain can greatly benefit from an understanding of the kinds of behaviors and mechanisms present at the time when that fundamental architecture was being laid down.

(a chemical activator) for a reaction BA which leads to the breakdown of A. If another reaction PA produces A at a constant rate, then the two reactions operating together will cause the concentration of A to equilibrate at some constant level. Any fluctuations in the concentration of A will cause the relative rates of PA and BA to change so that the concentration of A is brought back to this equilibrium. Because such a "negative feedback" mechanism exploits reliable properties of chemistry, it is likely to be discovered by the blind processes of variation and selection.

Next, consider a slightly different scenario. Compound B cannot be produced by the organism but must be absorbed from the environment. Suppose that compound B inhibits some cascade of chemical reactions WF which causes the waving of a flagellum (a hair-like appendage used for locomotion). If the concentration of B drops below some threshold level, the locomotion mechanism is released into action and the organism begins to swim randomly. Under the assumption that compound B is non-uniformly distributed in the environment, this motion is likely to improve the situation by bringing the creature to a site of higher concentration of compound B. Once such a site is reached, enough of compound B is absorbed to again inhibit reaction WF and motion ceases. This simple mechanism acts to maintain the concentration of B within some acceptable range, just as the purely chemical mechanism for controlling the concentration of A did above[7].

The second kind of homeostasis should not be any more surprising than the first. If evolution can exploit reliable properties of biochemistry, then it should also be able to exploit reliable properties of geometry and statistics. That the second kind of homeostasis involves a mechanism which effectively extends its action past the membrane, moving the organism through the environment, does not make it do something other than homeostasis. Both kinds of mechanisms ultimately serve similar functions – they maintain the conditions necessary for life to continue. They may be described as *control mechanisms*.[8]

As evolution produced increasingly more complex organisms, the mechanisms of control developed more sophisticated and more convoluted solutions to their respective tasks. Mechanisms controlling internal variables such as body temperature or osmolarity evolved by exploiting the consistent properties of chemistry, physics, fluid dynamics, etc. Today we call these "physiology". Mechanisms whose control extends out through the environment had to exploit consistent properties of that environment. These properties include statistics of nutrient distributions, Euclidean geometry, Newtonian mechanics, etc. Today we call such mechanisms "behavior". In both cases, the

---

[7] A functionally analogous mechanism is employed by modern woodlice. The mechanism operates under a simple rule – move more slowly when humidity increases – resulting in a concentration of woodlice in damp regions where they won't dry out. The bacteria *Escherichia coli* use a mechanism only slightly more sophisticated to find food. Their locomotion system increases turning rates when the nutrient concentration decreases, and thus they tend to move up the nutrient gradient. This mechanism, called *klinokinesis*, exploits the reliable fact that in the world of *E. coli*, food sources are usually surrounded by a chemical gradient with a local peak (Koshland, 1980).

[8] The term "homeostasis" implies some constant goal-state, but we need not be too devoted to that implication. Much of the activity of living creatures is anything but constant. For that reason, I stay away from the term "homeostatic mechanism" and prefer the more general term "control mechanism", implying only that control over a variable is maintained to keep it within a desirable range. How that range changes may be determined by various factors, including other, higher-level control mechanisms (Powers, 1973). Furthermore, it should not be assumed that a mechanism which exerts control over some state necessarily involves an explicit representation of the goal state (consider the examples described in footnote 7).

functional architecture takes the form of a negative feedback loop, central to which is the measurement of some vital variable. Fluctuations in the measured value of this variable outside of some "desired range" initiate mechanisms whose purpose is to bring the variable back into the desired range. These mechanisms may be direct, as in the chemical homeostasis example, or involve indirect causes and effects as in the example of klinokinesis (see footnote 7).

The alternative "control metaphor" being developed here may now be stated explicitly: ***the function of the brain is to exert control over the organism's state within its environment***.

This is not a novel proposal. Over a hundred years ago, John Dewey made essentially the same point I am making now. Dewey (1896) argued that the concept of stimulus-response is insufficient as a unifying principle in psychology because it only mentions part of the behavioral picture:

> *What we have is a circuit, not an arc or broken segment of a circle. This circuit is more truly termed organic than reflex, because the motor response determines the stimulus, just as truly as sensory stimulus determines movement. (Dewey, 1896, p. 363)*

The concept of stimulus-response, or reflex arc, focuses attention only on the events leading from the detection of stimulation to the execution of an action, and leads one to ignore the results of that action which necessarily cause new patterns of stimulation. "The reflex arc theory… gives us one disjointed part of a process as if it were the whole" (Dewey, 1896, p. 370). Surely, nobody has denied Dewey's observation that actions and perceptions mutually affect each other. And yet, the history of brain sciences in the 20th century suggests that this observation has largely been ignored.

But it wasn't ignored completely. The notion of behavior as control was fundamental in the early work on cybernetics (Rosenblueth, Wiener, & Bigelow, 1943), inspired in part by the theory and practice of engineering feedback systems like Watt's classic centrifugal steam governor. In fact, the word "cybernetics" was originally intended to imply a control system (Wiener, 1958), though that aspect of its meaning seems to have been neglected in the last few decades. Ashby's (1965) "Design for a Brain" elaborates these ideas into a theory of control systems capable of maintaining homeostasis and adapting their "sensorimotor" architecture through rudimentary learning.

The feedback nature of behavior has often been discussed within psychology as well. The relationship between physiology and behavior described above had been eloquently discussed by Jean Piaget (1967), whose seminal work on sensorimotor development was highly influenced by cybernetics. William Powers' (1973) book "Behavior: The Control of Perception" outlines a model of a behavioral control hierarchy spanning everything from simple reflexes to social interactions – this work has become the foundation of an entire psychological movement called Perceptual Control Theory (Bourbon, 1995). There is also, of course, James Gibson's "ecological psychology" (Gibson, 1979), about which more will be said below.

Many other theoretical schools of thought share the foundation of the control metaphor, from the theory of "autopoiesis" (Maturana & Varela, 1980), to the branch of AI research often termed "situated robotics" (Brooks, 1991; Mataric, 1992; Harvey, Husbands, & Cliff, 1993). It is beginning to be rediscovered in philosophy as well (Adams & Mele, 1989; Van Gelder, 1995). To my knowledge, Hendriks-Jansen (1996) provides the best synthesis of related ideas from numerous disciplines.

While in some fields the control metaphor is often perceived as something novel and revolutionary, in others it has been a founding concept for decades. Feedback control has long been used to describe the physiological operation of the body (Cannon, 1932; Schmidt-Nielsen, 1990), including the function of the autonomic nervous system (Dodd & Role, 1991). Only the central nervous system has been considered different, described within the concept of stimulus-response, and only by psychology, neuroscience, AI, and philosophy. In ethology, the study of *animal* behavior, the feedback control nature of behavior has been a foundation for years (Hinde, 1966; McFarland, 1971; Manning & Dawkins, 1992).

Why then has the control system metaphor been so neglected within mainstream psychology? Several possible reasons come to mind: 1) The experimental methodology in psychology deliberately prevents the response from affecting the stimulus in an effort to quantify the stimulus-response function. This is appropriate for the development of controlled experiments, but can be detrimental when it spills over into the interpretation of those experiments. 2) There is excess homage paid to the skin, and the structural organization of behavior (from receptors to effectors) is mistaken to be its functional organization (from input to output). 3) The behavior of modern humans is so sophisticated that most actions that we tend to contemplate are performed for very long-range goals, where the ultimate control structure is more difficult to appreciate. Thus, because many traditions of brain science began by looking at human behavior, they were not likely to see the control structure therein. 4) Systems with linear cause and effect are much more familiar and easier to grasp than the dynamical interactions present in systems with a closed loop structure. 5) Interdisciplinary boundaries have split the behavioral loop across several distinct sets of scientific literature, making it difficult to study by any single person. The study of advanced behavior has become allocated among numerous scientific disciplines, none of which is given the mandate of putting it all together. Even philosophy has only recently begun to look into brain science and biology for insight into mental function. 6) Finally, the various attempts to reintroduce the control metaphor into brain theory must themselves take some of the blame. In an attempt to establish themselves as distinct entities, many of the movements listed above described themselves as revolutionary viewpoints that redefine the very foundations of scientific psychology (e.g. Gibson, 1979) or Artificial Intelligence research (e.g. Brooks, 1991). Much criticism was leveled at mainstream theories, resulting in impassioned defenses. And during such defenses, the new movements were portrayed as already familiar and discredited viewpoints (usually as variants of the most extreme form of behaviorism (c.f. Ullman, 1980)) and thus quickly rejected. But it is not true that these movements redefine psychology – they merely present novel perspectives on existing data, data which continues to be relevant to the study of behavior.

This essay suggests that the control metaphor is a better way of describing brain function than the computer metaphor. One advantage, of particular interest to philosophy of mind, is that it provides a simple answer to the question of meaning. Briefly, rather than viewing behavior as "producing the *right response* given a stimulus", we should view it as "producing the response that results in the *right stimulus*". These statements seem pretty similar at first, but there is a crucial difference. While the first viewpoint has a difficult time deciding what is "right", the second does not:

Animals have physiological demands which inherently distinguish some input (in the sense of "what the animal perceives as its current situation") as "desirable", and other input as "undesirable". A full stomach is preferred over an empty one; a state of safety is preferred over the presence of an attacking predator. This distinction gives *motivation* to animal behavior – actions are performed in order to approach desirable input and avoid undesirable input. It also gives *meaning* to their perceptions – some perceptions are cues describing favorable situations, others are warnings describing unfavorable ones which must be avoided. The search for desirable input imposes functional design requirements on nervous systems that are quite different from the functional design requirements for input-output devices such as computers. In this sense, computers make poor metaphors for brains. For computers *there is no notion of desirable input within the computing system*, and hence there is the riddle of meaning, a.k.a. the symbol grounding problem.

Re-examining the problem of meaning

Most philosophical inquiries into meaning begin by contemplating the meaning of words and symbols. This is undoubtedly due to the influence that computer analogies and "language of thought" theories (Fodor, 1975) have historically had over the field. As discussed above, a few decades ago the paradigm of symbolic logic had been perceived as the only mechanistic explanation of complex behavior available as an alternative to the emptiness of dualism and the ineffectiveness of behaviorism. It thus defined the default premises for modern philosophy of mind. With that foundation, modern philosophical thought revolves around questions of how symbols acquire their meaning[9]. To repeat: "*How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?*" (Harnad, 1990).

It is usually assumed that if we can understand the meaning of symbolic tokens, then the meaning of non-symbolic representations (like sensorimotor schemata) will follow trivially. It is also usually assumed that the meaning of perceptual representations can be understood in isolation. For example, Harnad (1990) states that "motor skills will not be explicitly considered here. It is assumed that the relevant features of the sensory story … will generalize to the motor story." (Harnad, 1990, footnote 12). In an attempt to limit the scope of his analysis, Harnad isolates himself from considering the behavioral control loop which, I argue, is where the answer to his problem lies.

---

[9] One might suggest that we can avoid the symbol grounding problem by abandoning symbols in favor of connectionist representations. However, the problem exists for representations in general, be they symbolic or non-symbolic, as long as their role in controlling behavior is neglected.

Such conceptual isolation is a large part of the reason why meaning appears mysterious. As discussed above, traditional disciplinary boundaries and the need for specialization in science have divided the large problem of behavior into smaller sub-problems such as Perception, Cognition, and Action. These disciplinary boundaries then spill over into large-scale brain theories, yielding a model of the brain with distinct modules separated by putative internal representations. The Perceptual module is separated from the Cognitive module by an internal unified representation of the external world, and the Cognitive module is separated from the Action module by a representation of the motor plan. These boundaries help to limit the volume of literature that scientists are confronted with, but at the same time they isolate them from potential insights. Those who study the mechanisms of perception and movement control seldom contemplate "philosophical" issues like meaning. Meaning is left for those who study cognition. And those who study cognition start with a perceptual representation and usually phrase the question of meaning in terms of how meaningful symbolic labels can be attached to that representation, and how symbols can be *about* the things they refer to.

The symbol grounding problem has things backwards. Meaning comes long before symbols in both phylogeny (evolutionary history of a species) and ontogeny (developmental history of an individual). Animals interacted with their environment in meaningful ways millions of years before they started using symbols. Children learn to interact with their world well before they begin to label their perceptions. The invention of symbols, in both phylogeny and ontogeny, is merely an elaboration of existing mechanisms for behavioral control.

Again, let's step back from the philosophical debate surrounding meaning and consider issues that are more fundamental from a biological perspective.

In order to survive, organisms have to take an active part in controlling their situation and keeping it within desirable states. For an organism to exert control over its environment, there must exist predictable relationships between an action and the resulting stimulation ("motor-sensory" relationships). As discussed above, both physiology and behavior can function adaptively only if there exist reliable properties in the organism's niche which may be exploited toward its survival. Biochemical control exploits reliable properties of chemistry, diffusion, fluid dynamics, etc., while behavioral control exploits reliable properties of statistics, geometry, rules of optics, etc. With a simple behavioral control mechanism such as klinokinesis[7], it is easy to see how the system makes use of the statistics of nutrient distribution. With more complex behaviors, the properties of the niche which are exploited by the organism (i.e. brought within its control loop) are more subtle.

Gibson (1979) referred to these reliable properties of the niche as "affordances". Affordances are opportunities for action for a particular organism. For example, a mouse-sized hole affords shelter for a mouse, and it does so whether or not a mouse is perceiving it. Thus, an affordance is objective in the sense that it is fully specified by externally observable physical reality, but subjective in the sense of being dependent on the behavior of a particular kind of organism. For a mouse the hole affords shelter, while for a cat it instead affords a potential source of meat. Gibson

(1979) suggested that perception of the world is based upon perception of affordances, of recognizing the features of the environment which specify behaviorally relevant interaction. An animal's ecological niche is defined by what its habitat affords.

These relationships between animals and their habitats may be considered precursors to meaning. They are properties of the environment which make adaptive control possible and which guide that control. They make it possible for an organism to establish a behavioral control loop which can be used to approach favorable situations and avoid unfavorable ones. Because these properties tend to come packaged along with semi-permanent physical objects, we can speak of the "meaning" that these objects have to the organism in question. However, the crucial point is that *the "meaning of an object" is secondary to the meaning of the interactions which the object makes possible*.

For example, consider a child learning to control her arms. At first, she hardly even distinguishes the meaningless image of the hand on her retina from its background. Not knowing how to control the visual input, the infant at first sends random commands to the muscles. With time, her brain progressively comes to correlate certain random commands with certain motions on her retina – discovering the motor-sensory relationship. This correlation is possible only because the biomechanics of the arm and the laws of optics ensure consistency between the motor commands and the visual motion. After some time, the infant can produce desired visual motions of the hand-shaped retinal image by calling up the appropriate motor commands. She can control her hand. The resulting system might be described as involving representations and transformations between them (control theorists may describe these as pseudoinverse control). But these representations do not require meaning to be somehow *assigned* to them; their meaning is their role in the behavioral control of the arm.[10]

Such behavioral control itself establishes a new domain of consistent motor-sensory relationships, which can be used as the building blocks of higher-order behavioral control. For example, once control over the arm has been learned, it can be used to grasp food and thus to control satiation of hunger. The child's use of her arm for such higher-order behavioral control constitutes her understanding of the meaning of the arm. This is a non-symbolic, pragmatic kind of understanding, which only much later leads to the formation of such concepts as "arm" or "self" (Piaget, 1954).

In more sophisticated examples of sensorimotor learning, a child might discover that touching part of her environment produces interesting noises. With time the child may distinguish the rattle that lies next to her from the rest of the crib, discovering it as the source of the noises because other parts of the world seem to be irrelevant to these noises. She may learn to grasp the rattle independently of grasping other objects, and learn to shake it this way and that to produce desired sounds. She might thus discover that the rattle affords "noisemaking". She might even come up with some kind of crude internal symbol for the rattle. Again, must we worry how meaning gets

[10] This is reminiscent of Millikan's (1989) suggestion that the function of a representation is defined not by how it is produced, but by how it is used.

"assigned" to this symbol? The symbol is shorthand notation for "thing I can grasp and shake and make noises with". The symbol is constructed much later than the sensorimotor strategy which grounds it.

In summary, symbols are merely shorthand notation for elements of behavioral control strategies. The symbol "rattle" is learned by the child in the context of representations which already have meaning for her by virtue of their role in the behavioral control of her perception. There is no grounding problem because the symbol is only constructed after its meaning has already been established by the affordances that come packaged with the object to which it refers[11].

The symbol grounding problem has been turned upside-down: The question we're left with is not how meaning is assigned to the symbol "rattle", but how the symbol is assigned to meaningful interactions with a rattle. How does a shorthand representation emerge within a system of sensorimotor control strategies? What regularities in the environment make the establishment of symbols in the organism possible? – Perhaps one example is the fact that the affordances of an external object tend to remain attached to it as a set. For what behavioral abilities is such a shorthand representation useful? – One likely use is in predicting events; perhaps another is when constructive imitation among social animals evolves into purposeful teaching (Bullock, 1987). And at what phylogenetic stage of behavioral sophistication are these abilities observed? These are the kinds of questions that research on symbolic thought can productively pursue once it moves beyond the riddle of meaning.

One of the major reasons why meaning may appear as a riddle is its presence in communication. When humans talk over a telephone line, somehow there is meaningful information passed from one to the other, even though nothing more than arbitrary electrical impulses are being transmitted between them. This article is another example of a set of arbitrary symbols which somehow convey meaningful information. How can this be?

Let's consider communication from the evolutionary perspective which has been used throughout. In order to survive, an organism must exert control over its situation, and this control can only be achieved if there exist reliable relationships between actions and their results. While laws of chemistry create opportunity for biochemical homeostasis, laws of geometry, optics, etc. create opportunity for behavioral interaction. A behavioral control loop can be constructed wherever consistencies exist in the environment. And of course, such consistencies also exist in the behavior of other animals.

Animals respond in complex but predictable ways to various stimuli. When a threatening posture is assumed by one crayfish, another will either back off or respond with its own threat posture. This establishes a domain of interaction between the two creatures where each can attempt to control the other into conceding submission. Often, no actual

---

[11] In a sense, what we have is similar to Harnad's (1990) proposal that the arbitrary symbols such as the word "rattle" are linked to non-arbitrary and non-symbolic representations. However, these non-symbolic representations are not "iconic" or "categorical" representations of perceptual entities, but rather the elements of sensorimotor control mechanisms which operate by making use of consistent properties in the environment.

fighting needs to take place before the dominance hierarchy is achieved. It may be said that the dominant crayfish has successfully exerted control over its opponent. The behavioral control loop can thus extend out through other creatures.

The threatening posture of the crayfish may be called a "signal". It demonstrates a threat without having to do any actual physical damage. Other kinds of primitive signals in animal behavior include the eye-spots of certain moths, the white tails of fleeing deer, the exposed teeth of aggressive baboons, etc. All of these are used for exerting control over other individuals, whether they be of the same or of different species.

In social animals, such signaling protocols have been developed to great sophistication. These, it seems, deserve the label of "communication". While the earliest symbols (such as the threat posture of crayfish) closely resemble the state of affairs (battle) which they signal, over time the symbol forms may diverge into arbitrary variations. For example, the dance which bees use to communicate the direction and distance of a food source is quite different than the movements needed to arrive at the food source (Frisch, 1967). The dance may be said to involve arbitrary symbolic tokens, in Harnad's (1990) sense. But these tokens are grounded because they play the role of mediating the control that the dancing bee has over the other bees.

Thus, like physiology and behavior, communication is also an extension of control: one which encompasses other creatures in the environment. To describe communication merely as "transmission of information" is incomplete. While information is indeed transferred in communication, to miss the control purpose of the transmission is a fatal oversight in an attempt to understand the meaning of the communiqué.

The same case can be made for the meaning of words. As the system for communication grows in complexity and acquires syntax, the meaning of its elements derives from the behavioral control goals of the speaker. In humans, this ability has developed most impressively and meanings have been built upon meanings until even abstract concepts can be expressed. Nevertheless, most communication, including this article, is an attempt at *persuasion*.

Conclusions

A car may be described as a device for converting chemical energy into kinetic energy. This description is not false, and could serve as the foundation for a scientific study of cars. Such study could lead to theories of how parts of the car contribute to its role of energy conversion (for example, the drive-shaft and axle system may be viewed as a coordinate transformation of kinetic energy) but it would fail to provide a complete picture of the car's function. In order to understand the purpose of such things as the steering wheel, one has to understand that the function of a car is to transport people. Energy conversion is merely a useful means toward that end.

Likewise, the "processing" of "information" is merely a useful means toward the goals of behavior. The brain is not merely an input-output device; it is a control system which exerts control over the organism's state in the environment. This task is accomplished through the exploitation of regularities within the environment that define reliable rules for interaction. A viewpoint which isolates the behaving organism from its domain of interaction, as

done by the computer metaphor, misses the importance that such regularities have for the establishment of effective control strategies, and the meaning they have within the system. Such a view has forced most modern attempts to understand behavior to place all explanatory burden upon mechanisms within the skin, ignoring the contribution that the environment can make toward the guidance of behavior.

One may argue that any control system is a special case of an input-output system: one where the outputs feed back to affect the inputs. This is true. In fact, it means that the control metaphor is *more precise* than the input-output metaphor. To describe brains as input-output devices is like describing cars as energy-conversion systems without adding something that distinguishes them from chainsaws and chloroplasts. Because the control metaphor is a more precise description, it better constrains the task of explaining behavior. Control problems, being a subset of input-output problems, present a smaller search-space of possibilities. To step outside that subset is to risk spending time on questions which are not of immediate relevance, even if they apply to some other members of the general set. For example, questions aimed at systems with unbounded time and unbounded memory, such as idealized Turing Machines, may be of limited utility for studies of the biological brain.

The historical interdependence in the development of both functionalism and computationalism have resulted in the two concepts often being viewed as inseparable. However, they are not, and several prominent criticisms aimed at functionalism are really criticisms of computationalism. Notably, Searle's Chinese Room Argument (Searle, 1980) is an argument against the possibility of true understanding within a system that performs input-to-output computations, but as discussed above, that's not a good analogy for what brains do. Brains interact with a world, controlling their situation by performing actions that result in desirable input. They do this by discovering and exploiting the reliable rules of output-to-input transformation that are made possible by the external world – *this is what a pragmatic understanding of the world equates to*. Any system which is capable of performing such control over the environment will perforce contain internal representations that have meaning to it. For Searle's room, and for computer systems which merely accept information, there is no notion of desirable input, no utilization of opportunities for interaction with the world, and thus no understanding of the external world or its projection onto receptor surfaces.

In summary, what I am advocating here comes down to correcting two crucial mistakes that psychology has been led to make: 1) severing the behaving organism from its environment; and 2) decomposing behavior into Perception, Cognition, and Action modules which are then studied in isolation. Both of these are very old ideas, but they have become particularly entrenched in mainstream psychological thought with the development of computationalism.

To reject these two mistakes by viewing the brain as a control system does not, however, invalidate the progress made under computationalism. For example, the idea of lawful transformations of internal patterns of activity is still useful. We can still refer to this as the "processing" of "representations" as long as we focus on the pragmatic value these representations have for the task of control and not dwell solely on how they may describe the world. The idea

of specialized brain subsystems is also perfectly reasonable, as long as we delimit these subsystems for functional reasons and not because of conceptual traditions. Finally, many existing models of brain systems, developed in the context of the computer metaphor, are equally compatible with a high-level view of the brain as a control system – this is particularly true of connectionist models.

Making the change in perspective from viewing the brain as an input-output device to viewing it as a control system also leads to a number of important conceptual shifts. A major one is a shift from an emphasis on representations to an emphasis on behaviors, from the analysis of serial stages of processing to an analysis of parallel control streams. This lets one avoid some classic problems in philosophy of mind. First, as discussed above, once neural representations are viewed within the context of the behavioral control to which they contribute, their meaning is not a mystery. Second, motivated action is also not a mystery – when an animal's physiological state no longer meets its internal demands (like a growing hunger), action is generated so as to bring it to a more satisfying state. Third, once one no longer assumes the presence of a complete internal representation of the external world, many forms of the "binding problem" are no longer difficult. When environmental regularities are allowed to take part in behavior, they can give it coherence without need of explicit internal mechanisms for binding perceptual entities together (Cisek & Turgeon, 1999).

Finally, the shift away from serial representations leads one to reconsider some classic notions concerning consciousness. Much of psychological theory in the last century has been developed in the context of philosophical viewpoints on the mind-body problem. Because dualism and its variations have thrived at least until the 1950's, they had a great deal of influence on the foundations of psychology. This dualist backdrop led many psychologists to assume that the brain, somewhere within it, presents a model of the world for the mind to observe. Dualism called for a central representation, and computationalism provided that in the form of the internal world model upon which cognition presumably applies its computations. Thus, there has existed for a long time a symbiotic relationship between dualism, a philosophical stance, and computationalism, a psychological viewpoint. And although dualism itself has largely been discredited, some if its influences, such as the assumption of a unified internal world model, remain with us still. Deconstructing that assumption by moving beyond computationalism will have profound effects on what we imagine that the neural correlates of consciousness might be. Perhaps the shift will help us develop a more functional concept of consciousness than the currently prevalent dualistic one, freeing us from the persistence of the so-called "hard problem".

The last three decades of brain science have witnessed a progressive backing away from several premature assumptions based on the computer analogy. It was quickly obvious that serial searches among a combinatorial set of possibilities cannot be the way that a human brain reasons, even if today such a process can be made fast enough to beat a chess grand-master. Neuroscience research has made it clear that the brain operates with large numbers of noisy elements working in parallel rather than with a single powerful CPU. Human memory appears to be stored in a distributed manner rather than in the sequential addresses of computer memory. The re-emergence of connectionism

has questioned the notion that symbolic logic is the only form of computation to be considered. Artificial life research has demonstrated that robots can be built without accurate sensors and explicit internal representations (Brooks, 1991; Mataric, 1992), and even that such robots can be developed through simulated evolution (Beer & Gallagher, 1992). All these developments demonstrate weaknesses of the aging analogy that brains are like computers, and motivate us to take a few steps back away from some of the assumptions it originally generated.

This article argues that one more step needs to be taken. We need to step back from the input-output metaphor of computationalism and ask what *kind* of information processing the brain does, and *what is its purpose*? The answer, suggested numerous times throughout the last hundred years, is that the brain is exerting control over its environment. It does so by constructing behavioral control circuits which functionally extend outside of the body, making use of consistent properties of the environment including the behavior of other organisms. These circuits and the control they allow are the very reason for having a brain. To understand them, we must move beyond the input-output processing emphasized by computationalism and recognize the closed control-loop structure that is the foundation of behavior.

Acknowledgements:

I am grateful to Peter Cariani for his valuable advice on this essay. Supported by a fellowship from the National Institutes of Health (F32 NS10354-02).

References

Adams, F., & Mele, A. (1989). The role of intention in intentional action. *Canadian Journal of Philosophy*, *19,* 511-531.

Ashby, W.R. (1965). *Design for a Brain: The Origin of Adaptive Behaviour*. London: Chapman and Hall.

Bedian, V. (1982). The possible role of assignment catalysis in the origin of the genetic code. *Origins of Life*, *12,* 181

Beer, R.D., & Gallagher, J.C. (1992). Evolving dynamical neural networks for adaptive behavior. *Adaptive Behavior*, *1,* 91-122.

Bourbon, W.T. (1995). Perceptual Control Theory. In H.L. Roitblat & J.-A. Meyer (Eds.), *Comparative Approaches to Cognitive Science*. (pp. 151-172). Cambridge: MIT Press.

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

Bullock, D. (1987). Socializing the theory of intellectual development. In M. Chapman & R.A. Dixon (Eds.), *Meaning and the growth of understanding: Wittgenstein's significance for developmental psychology*. (pp. 187-218). New York: Springer-Verlag.

Cannon, W.B. (1932). *The Wisdom of the Body*. New York: Norton.

Cisek, P., & Turgeon, M. (in press). 'Binding through the fovea', a tale of perception in the service of action. *Psyche*.

Crick, F.H.C. (1994). Interview with Jane Clark. *Journal of Consciousness Studies*, *1,* 10-17.

Crick, F.H.C., Brenner, S., Klug, A., & Pieczenik, G. (1976). A speculation on the origin of protein synthesis. *Origins of Life*, *7,* 389

Dennett, D.C. (1978). Current issues in the philosophy of mind. *American Philosophical Quarterly*, *15,* 249-261.

Dewey, J. (1896). The reflex arc concept in psychology. *Psychological Review*, *3,* 357-370.

Dodd, J., & Role, L.W. (1991). The autonomic nervous system. In E.R. Kandel, J.H. Schwartz, & T.M. Jessell (Eds.), *Principles of Neural Science*. (pp. 761-776). New York: Elsevier.

Eigen, M., & Schuster, P. (1979). *The Hypercycle - a principle of natural self-organization*. Heidelberg: Springer-Verlag.

Fitts, P.M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47,* 381-391.

Fodor, J.A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fox, S.W. (1965). Simulated natural experiments in spontaneous organization of morphological units from protenoid. In S.W. Fox (Ed.), *The Origins of Prebiological Systems*. New York: Academic Press.

Frisch, K.v. (1967). *The Dance Language and Orientation of Bees*. Cambridge, MA: The Belknap Press of Harvard University Press.

Gibson, J.J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42,* 335-346.

Harvey, I., Husbands, P., & Cliff, D.T. (1993). Issues in evolutionary robotics. In J.-A. Meyer, H.L. Roitblat, & S. Wilson (Eds.), *Proceedings of the Second Conference on Simulation of Adaptive Behavior*. Cambridge, MA: MIT Press.

Hendriks-Jansen, H. (1996). *Catching Ourselves in the Act: Situated Activity, Interactive Emergence, Evolution, and Human Thought*. Cambridge, MA: MIT Press.

Hinde, R.A. (1966). *Animal Behaviour: A Synthesis of Ethology and Comparative Psychology*. New York: McGraw-Hill Book Company.

James, W. (1890). *The Principles of Psychology*. New York: Holt.

Kauffman, S.A. (1993). *The Origins of Order: Self-organization and selection in evolution*. New York: Oxford University Press.

Koshland, D.E. (1980). *Behavioral Chemotaxis as a Model Behavioral System*. New York: Raven Press.

Manning, A., & Dawkins, M.S. (1992). *An Introduction to Animal Behavior*. Cambridge: Cambridge University Press.

Mataric, M.J. (1992). Integration of representation into goal-driven behavior-based robots. *IEEE Transactions on Robotics and Automation*, *8,* 304-312.

Maturana, H.R., & Varela, F.J. (1980). *Autopoiesis and cognition: the realization of the living*. Boston: D. Reidel.

McCarthy, J., & Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (Eds.), *Machine Intelligence IV*. Edinburgh: Edinburgh University Press.

McFarland, D.J. (1971). *Feedback Mechanisms in Animal Behaviour*. New York: Academic Press.

Millikan, R.G. (1989). Biosemantics. *The Journal of Philosophy*, *86,* 281-297.

Mingers, J.C. (1996). An evaluation of theories of information with regard to the semantic and pragmatic aspects of information systems. *Systems Practice*, *9,* 187-209.

Piaget, J. (1954). *The Construction of Reality in the Child*. New York: Basic Books.

Piaget, J. (1967). *Biologie et Connaissance: Essai sur les Relations Entre les Régulation Organiques et les Processus Cognitifs*. Paris: Editions Gallimard.

Powers, W.T. (1973). *Behavior: The Control of Perception*. New York: Aldine Publishing Company.

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, *10,* 18-24.

Schmidt-Nielsen, K. (1990). *Animal Physiology: Adaptation and Environment*. Cambridge: Cambridge University Press.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3,* 417-457.

Shannon, C., & Weaver, W. (1949). *The Mathematical Theory of Information*. Urbana: University of Illinois Press.

Still, A., & Costall, A. (1991). *Against Cognitivism: Alternative Foundations for Cognitive Psychology*. Hemel Hempstead: Harvester Whitesheaf.

Turing, A.M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, Series 2*, *42,* 230-265.

Ullman, S. (1980). Against direct perception. *Behavioral and Brain Sciences*, *3,* 373-415.

Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *91,* 345-381.

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

Wiener, N. (1958). *Cybernetics, or control and communication in the animal and the machine*. Paris: Hermann.