Mexican Journal of Behavior Analysis
Monographic issue
Vol 22, 1996, pp. 161-195

Consciousness in computational theories of the mind

La conciencia en las teorías computacionales de la mente

Earl Hunt
The University of Washington

Abstract

Theories in cognitive psychology are based upon the idea that an adequate theory of human thought would amount to the design of a machine that is capable of mimicking human thought. Such a machine could be described at three levels; as a physical device, as an abstract computational device, or as a device that represented certain aspects of the environment in its internal structure. Only the first and the last of these levels can be observed directly, computational capacities must be inferred. However observations at the physical and representational level can serve to constrain theories at the computational level. Recent observations on how visual percepts are created from sensory stimlulation and observations on the influence on thought of information that cannot be recalled explicitly are particularly important. Our current knowledge of these constraints is sufficient to show that a computational level theory of thought must consider two types of computations, that roughly parallel the folk psychology distinction between conscious and unconscious thought. These two types of thought are associated with different computational capacities. Unconscious thought processes are attached to specific stimuli defined in a single sensory modality, while conscious thought processes can be attached to stimuli defined by multiple sensory signals and to abstract descriptions of situations.

Keywords: Attention, blackboard models, computational models of the mind, computer simulation, consciousness, implicit memory,philosophy of mind, philosophy of science, visual perception, visual search

Resumen

Las teorías en la psicología cognoscitiva se basan en la idea de que una teoría adecuada del pensamiento humano equivale al diseño de una máquina que es capaz de mimetizar el pensamiento humano. Tal máquina podría describirse en tres niveles: como un dispositivo físico, como un dispositivo computacional abstracto o como un dispositivo que representara ciertos aspectos del ambiente en su estructura interna. Solo el primero y el último de estos niveles pueden observarse directamente. Las capacidad computacionales deben ser inferidas. Sin embargo, las observaciones en los niveles físico y representacional pueden servir para restringir las teorías en el nivel computacional. Las observaciones recientes de como se crean los perceptos visuales a partir de la estimulación sensorial y las observaciones sobre la influencia de la información que no puede recordarse explícitamente en el pensamiento, son particularmente importantes. Nuestro conocimiento actual sobre estas restricciones es suficiente para mostrar que una teoría del pensamiento en el nivel computacional debe considerar dos tipos de computaciones, que de manera muy gruesa, reproducen la distinción de la psicología popular entre pensamiento conciente e inconciente. Estos dos tipos de pensamiento se asocian con diferentes capacidades computacionales. Los procesos inconcientes de pensamiento se vinculan con estímulos específicos definidos en una sola modalidad sensorial, mientras que los procesos concientes de pensamiento pueden ligarse a estímulos definidos por senales sensoriales múltiples y a las descripciones abstractas de situaciones.

Palabras clave: atención, modelos de tablero, modelos computacionales de la mente, simulación de computadora, conciencia, memoria implícita, filosofía de la mente, filosofía de la ciencia, percepción visual, búsqueda visual

Introductory remarks

We think we think. I believe that every human has the subjective experience of conscious thought, even though I agree with Descartes that this is a statement of faith on my part. The issue of consciousness has been tightly bound to discussions about the soul, free will, and morality, for we can only be held responsible for those things of which we are aware and can control. If I were to kick my physician, gratuitously, he would have grounds to complain. If I strike him while he is testing my patellar reflex that is his fault, he should have known better than to stand in front of my foot. In an engineering sense, though, both

actions are the result of information processing mechanisms that transduce stimulus input into response output. But just what is the difference? This is one of the great questions for psychology. Here I want to take a modest step forward, by considering just what the distinctions are likely to be between conscious and unconscious reasoning.

A discussion of consciousness is a somewhat unusual topic for a conference dedicated to the Behaviorist viewpoint, because within that viewpoint it is difficult to even talk about the question. However my viewpoint is not the Behaviorist one, it is a cognitive science viewpoint. Therefore I shall begin with some remarks about just how the cognitive science viewpoint differs from Behaviorism. Then I shall go into the main part of my discussion. First I review the distinctions between conscious and unconscious reasoning as seen by folk and scientific psychology. I then consider these distinctions from the point of view of two classes of computational theories of the mind. I will then close with some further remarks about both the conclusions that cognitive psychologists are reaching about consciousness and about the ways in which they reach these conclusions.

The cognitive psychology view

There are many cognitive psychologists, and even more cognitive scientists. Therefore it is somewhat presumptuous to claim that any particular view is the revealed one. This is particularly true because cognitive science coalesced from a number of different laboratories, each of which was interested in different aspects of cognition. As a result, although cognitive science certainly has some investigators who are more prominent than others (including one Nobel Laureate... Herbert Simon) the field is not dominated by the ideas of one or two individuals, in the sense that Behaviorism has been driven by the ideas of Watson and Skinner. Watson preceded Skinner, and Skinner was clearly influenced by and extended Watsons ideas. The acknowledged leaders in cognitive science include Naom Chomsky, a linguist, Simon, an economist, Marvin Minsky and Allen Newell, both primarily computer scientists, David Marr, a computer scientist interested in machine vision, a very different field than Minsky and Newells, Jerry Fodor and Zenon Pylyshyn, philosophers of mind, Allen Baddeley and Michael Posner, both trained as classic experimental psychologists, and Frank Rosenblatt, who is best described as having invented neural based computation. . These people are all roughly contemporary . All but Marr,

Newell, and Rosenblatt are active today, and these three would only be in their fifties or sixties were they still alive. Cognitive science was produced by a collision rather than an evolution of ideas. Nevertheless, there is considerable agreement about certain aspects of research on human behavior.

Cognitive scientists believe very strongly in what Chiessa (1996) has referred to as a mechanistic view of human behavior. One of the most important early works cognitive psychology, Newell and Simons (1972) *Human Problem Solving*, does not contain a single statistical test. Newell and Simons argument for ignoring this crutch that most psychologists lean on so heavily is that virtually no human action is ever random. When a mathematician takes a step in solving an equation or a chess player moves a piece (both topics studied in *Human Problem Solving*) concepts such as "increased *probability* of emission of the behavior "are simply fictions.[1] They go further, and argue that each move in either of these cases is caused by beliefs, i.e. mental events, inside the persons mind. This idea is not at all mysterious, it is developed directly from an avowedly materialistic view of the world. The following analogy is the sort of argument that cognitive scientists find appealing.

Imagine a science-fiction scenario, in which a scientist lands on a far away planet, completely populated by robots. The scientists task is to find out how the robots work. This task could be approached at three different levels. One could ask how the devices worked at the physical level. This would entail the study of resistors, transistors (or their analogs) and basic circuitry. A second approach would be to ask how the robots interacted with the world around them. This would entail the study of *representations*, what information the robots internalized about the external world and how these representations controlled the robots actions. Finally, an intermediate approach would be to attack the problem at the computational level, by asking what symbolic computations the robots were capable of, and how those symbolic computations were combined to construct representations. Of course, it would also be possible to ask how the symbolic computations were realized by the computers hardware. In Newells

1. Newell and Simons anathema for statistics is an extreme case. Other investigators, at other times, do use statistics in the analysis of data. However we regard this as an unfortunate necessity, required in order to reveal trends in data produced in situations where we do not have adequate knowledge of all mechanistic causal variables. In other words, we use statistics because we have to treat the environment and response, as we define them, as closed systems whereas in fact they are open systems responding to variables that we do not observe. Nevertheless, with the possible exception of people who study thinking at the level of the neurosciences, we do believe that mechanistic cause and effect relationships tie observed responses to both the observed and unobserved environmental variables. Thus our use of statistics is philosophically quite different from the use made by a scientist who believes that some sort of Heisenbergs law of uncertainty is operating at the biological, psychological, or social levels.

(1980) terminology, any thought, by computers, humans, or robots, involves symbolic manipulation achieved by a physical system. Therefore when we study thinking we are seeking to understand the operations of a *physical symbol system.*

This idea has been developed by a number of authors, so I will only attempt to present the basic ideas. (The interested reader should see Pylyshyn (1989) or Hunt (in press) for detailed discussions.) A central concept is the notion of a constraint. Obviously, the biological capacities of humans constrain the symbolic computations that they can do. Very few, if any, of us can derive logarithms without artifacts(e.g. pieces of paper) to record intermediate computations. Similarly, human symbolic capacities restrain the sorts of representations that people can have. Thus the computational aspect of thought, the aspect that deals with abstract symbol manipulating capacities, plays an extremely important role in theories of cognitive psychology. However symbol manipulation can never be observed directly. At the molar level, thinking has to be about something. At the behavioral level all we can do is to observe how a person or animal responds to a variety of environmental challenges, stated at the representational level, and infer symbolic capacity from observations at the representational level. At the biological level, all we can do is to observe brain activity in situations that are designed to call upon certain symbol manipulation activities, such as holding information in short term memory. The observations at both the biological and representational level are unlikely to define unique systems at the symbol manipulation level. However they can place constraints upon the ideas about symbol manipulation that it is worth having.

The following section describes how a mixture of observations at the biological and representational level have constrained cognitive theories at the symbol manipulation level. The gist of the presentation is that the constraints are sufficient to force the acceptance of a fairly complicated theory, in which a distinction is made between conscious and unconscious thought.

Folk Psychology and conscious reasoning

The terms "conscious" and "unconscious" thought are almost anathema to advocates of radical behaviorism. On the other hand, the distinction between conscious and unconscious thought is basic to folk psychology. There are even cultural conventions about the appropriateness of different types of thinking. We are supposed to know what we are doing when we make scientific analyses

and stock market decisions. Overt rational analysis is proscribed when it comes to matters of the heart! A (divorced) friend of mine once told me that at age 23, just after graduation from college, he married a nineteen year old woman with a thousand dollar a month bill at boutiques. In retrospect that was a dumb move, but in prospect he was supposed to decide on the basis of intuition, not financial prospects.

The folk distinction between intuition and rational analysis surfaced in a quite different way when the U.S. military reviewed its performance after the Persian Gulf war of 1990. General Merrill McPeak, then Chief of Staff of the Air Force, used the term "situational awareness" to describe certain actions and errors committed by military pilots. While General McPeak never provided a precise definition, he was evidently referring to the ability to maintain awareness of the current state of important variables that define a tactical situation, so that ones spontaneous reaction to an observation is modified by an awareness of the context in which actions must be taken. Aviation is a particularly good field for showing lapses of situational awareness, simply because of the dramatic consequences these lapses can have. In January of 1996 an American Airlines aircraft crashed into a mountain near Cali, Colombia because the flight ignored messages indicating where they were on their flight path, and turned the wrong way when they received radar signals showing that they were approaching the ground. As the example suggests, the concept of situational awareness is closely related to the human factors definition of slips and context errors (Reason,1990). But this defines the problem, rather than solving it. How do we keep those pieces of information in consciousness that have to be there?

The folk distinction between conscious and unconscious thought seems to revolve around three distinctions. First, conscious reasoning is, by definition, available to introspection. There is a sharp distinction drawn between reasoning that can be explained to third parties and reasoning that just happens, intuitively. Second, conscious reasoning can only deal with a limited number of variables. It is based on reaction to selected aspects of a situation rather than a reaction to an unanalyzed *Gestalt*. This distinction is captured by scoring systems in sports such as ice skating and dancing, where judges are supposed to distinguish between technical merit, which is publicly defined as the execution of certain moves, and artistic impression, which is based on the judges unanalyzed reaction to the overall performance. Finally, the folk definition assumes that some complex reasoning takes place outside of conscious awareness. Freudian psychodynamics have clearly been incorporated into the public mind.

Cognitive psychology and conscious-unconscious distinction

An examination of the scientific literature show that cognitive psychologists agree in part with the folk psychology notion of a conscious-unconscious distinction, but that there are important differences between the scientific and the folk psychology view.

Cognitive psychologists definitely agree that conscious reasoning deals with a limited number of variables at any one time. Millers (1956) classic paper on "The magic number seven, plus or minus two" pointed out that at most from five to nine objects of thought can be at the focus of our attention at one time. Subsequent research (e.g. Baddeley, 1986, 1992) has left no doubt that if anything this was an overestimate, and that our limited ability to attend to objects of thought over time is a significant limitation on our reasoning ability. It is also easy to show that our perception of the world is constrained by limited attention. Again, we can turn to a classic study; Sperlings (1960) demonstration that our visual system registers far more information than we consciously attend to. The early interpretation of Sperlings work was that the mechanisms that produce consciousness, a mental concept, read from a fading iconic buffer provided by the brain. A more modern view (Crick, 1994, Kosslyn & Koenig, 1992) is that the brain disassembles the visual sensory signal into motion, color, and form components, as an automatic process, and then reassembles those components that are associated with locations to which attention is directed. An observer becomes aware of a visual stimulus only when the reassembly process has been completed. It is constructive to review how this conclusion was reached, because the experiments suggesting the conclusion illustrate how cognitive psychology views the world.

Sperling (1960) showed observers an array consisting of three rows of letters. The array was displayed for less than half a second. People reported seeing only from three to six letters, usually starting at the top left of the array. Sperling then sounded an auditory cue *after* the visual stimulus had been removed. Different tones were used to signal observers that they should report from the top, middle, or bottom row. It turned out that observers could report from three to six letters from whatever row was indicated, even though the indication came after the external stimulus was removed. This was widely reported as evidence that observers registered many letters in a large visual buffer, which Neisser (1967) dubbed the *iconic store*. Information was then 'read into conscious awareness from the iconic store, rather than from the external stimulus. Subsequent experiments were conducted that indicated that the iconic

store seemed to last for about a second, and that it was located in the brain, rather than being a reflection of fading neural activity in the retina. However a number of objections were made to this idea. Rather than conducting a critical review, which Haber (1983) has already done admirably, let us look at the idea that is replacing the notion of an iconic store.

The new ideas are based largely on the results of a series of visual search experiments conducted by Anne Treisman and her colleagues (Treisman, 1988). In a visual search experiment observers are asked to find a target in an array containing the target and a number of distractor signals. The procedure is illustrated in Figure 1. The topmost panel shows the display used when the task is to find in X in a field of O 's. The middle panel shows the display required in a search for a BLUE object in a field of RED Xs and Os. (In the interest of saving on printing costs, shaded and bold black has been substituted for blue and red.) In both cases the time required to search for a target is only slightly increased by an increase in the number of distractors. The target just 'pops out at the observer. Treisman has referred to these sorts of visual searches as *feature* searches, the task is to record the presence or absence of a visual characteristic that the brain treats as a primitive feature.

The bottom panel in Figure 1 shows a quite different sort of search. In this case the target, a RED X, is defined by the conduction of two primitive features

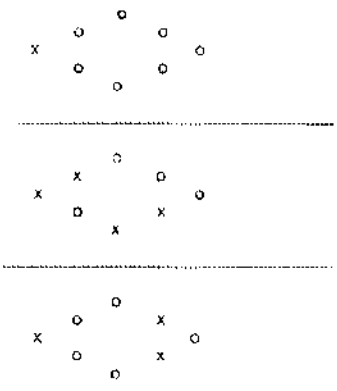

Figure 1. Feature and conjunction searches. In the top panel the task is to search for an X in a field of Os. In the middle panel the task is to search for a dark figure in a field of light figures. (The task was to search for a red object amidst blue objects in the original studies.) Both of these tasks are feature searches. The search time is almost independent of the number of items in the display. In the bottom panel the task is to search for a dark X in a field of figures that vary in form and shading. This is a feature search, and the time required is a linear function of the number of items in the display.

(Red color and diagonal lines). Distractors may possess one or the other of the two target features, but not both. In *conjunction search* tasks such as this one the time taken to locate a target is a sharply rising linear function of the number of distractors present.

Treisman (1988) offered a fairly complicated explanation. She proposed that the visual system consists of a number of separate location maps, some of which locate colors in positions and others locate primitive features (lines at various orientations, circles, etc.) at particular locations. Subsequent research (Treisman & Gormican, 1988) has produced a fairly large set of locatable features, including color, form, and motion. Treisman argued that detection of features in different maps is a rapid, process in which different features can be searched for in parallel. On the other hand reassembly of information into a perceivable percept requires that the brain determine that certain colors and forms all appear at the same location. Treisman argued that this is an effortful, time consuming process.

Treismans results, alone, do not compel us to reject the simpler iconic buffer model. Duncan and Humphreys (1988) present a model for searching an iconic buffer that could reproduce the behavioral differences between feature and conjunction searches. The literature, however, suggests strongly that cognitive psychologists are following Treismans approach rather than looking further at the iconic buffer. The reason is that the assembly-disassembly model, although not distinguishable from the iconic buffer model on behavioral grounds alone, is backed up by evidence from the neurosciences. Single unit recording studies in non-human animals have shown that there are indeed separate brain pathways for the transmission of information about color, form, location and motion (Livingstone & Hubel, 1988). These observations have been further confirmed by a variety of other neuroscientific studies, including studies of brain injured individuals (Kosslyn & Koenig, 1992). Crick (1994) has summed up our current knowledge by saying that we know that the brain takes the visual stimulus apart but we do not know how it is put back together again. We do know, though, that the assembly process is a demanding one. Although phenomenologically we think we see a large visual field, in fact our ability to attend is limited to detail is limited to a relatively small part of the visual field, even in situations in which there are no eye motions. We can be aware of only a few things at a time.

This raises an interesting question about what we mean by "being aware." Being able to tell someone else that we have seen something sets a quite high criterion for awareness. One could design a robot in which this criterion had to be met before thought occurred. The robots ability to retrieve meanings previ-

ously associated with a stimulus would depend upon first reaching a stage of complete perception of the . Humans apparently do not work this way. There is now quite a bit of evidence indicating that showing that an external event can influence the activation of information in the brain even though the event is not processed to the point at which the observer can report it. Most of this evidence comes from studies of *subliminal perception.* In the better designed of these studies a word or letter string is exposed for a few milliseconds, followed by a pattern mask, then a word to be identified. An example would be the sequence

BARK

*&@#%)(&

DOG

The observers task is to identify the second word as a word or a non-word string, (e.g. KOG). If the identification time is speeded by the presentation of a semantically related first word then the first word is said to have *primed* the second word. Priming is quite easy to demonstrate if the interval between the first word and the mask (The *stimulus onset asynchrony,* SOA) is sufficiently long so that the first word is perceived. If SOAs are much less than 50 milliseconds observers will not report seeing the word, but priming may occur. Other effects depending upon semantic associations with the unseen word have also been observed (Fowler et al. 1981, Marcel, 1983). Such studies present some tricky technical problems, but by now enough of them have been met so that the general conclusion is not seriously in dispute. Stimuli that are not consciously perceived, in the sense that the individual can report having seen them, can bias the interpretation of a consciously perceived target stimulus that is presented within seconds or fractions of a second of the unperceived stimulus (Greenwald, 1992).

Subliminal perception phenomena show that there is a distinction between the processing of information from the external environment and our conscious awareness of that information. It can also be shown that information in long term memory can influence our current thinking, even though we are not able to report or manipulate these influences. There are numerous studies showing that information presented at time A can influence behavior at some subsequent time, B, even though that information cannot explicitly be recalled. The evidence from amnesic patients is particularly dramatic. Suppose an amnesic is presented with a list of words containing the word MOTOR. Subsequently the

patient may not be able to recall the list, but will show an elevated tendency to complete the word stem item MOT__ with MOTOR rather than an alternative, such as MOTEL (Richardson-Klahven & Bjork, 1988). Furthermore, this is not a special characteristic of amnesiacs. A closely related phenomenon can be shown in normal individuals, using a paradigm called *process dissociation* that has been developed by Jacoby and his colleagues (Jacoby, Toth, & Yonelina, 1993). In process dissociation studies people are shown a list of words and then asked to complete word stems, either using words that did appear on the list (inclusion condition) or without using words that appear on the list (exclusion condition). In spite of the instructions there is an elevated tendency to use the words on the list in the exclusion condition. This means that people are using words that they remember, in the sense of being influenced by them, even though they would not use the words if they remembered them in the sense of being consciously aware of those memories.

None of these findings would have surprised Freud, who certainly believed in the existence of unconscious cognitive processes. Many of Freuds ideas are now part of the popular wisdom, so folk psychology certainly finds the idea of unconscious thought appealing. If anything, Freud and folk psychologists would be disappointed at the simplicity of the unconscious processes that cognitive psychologists study. Freudian and popular beliefs assume that conscious processing can be quite involved, even to the point of leading a person to use one activity as a symbolic substitution for another, usually more primeval activity. Cognitive psychologists, in general, do not find any evidence for such complex unconscious processing. Instead all they see is an ability to sensitize certain aspects of information that are already in memory. For instance, Greenwald (1992) maintains that studies of "unconscious" perception have produced clear evidence that the semantics of single words can be processed, but that nothing more than this has been shown.

This attitude, of course, is very much at odds with the clinicians conclusions that repressed memories can control complex analyses of current situations. An extreme example is the argument, strongly believed by some clinicians (Bass & Davis, 1989) that *unremembered* childhood sexual abuse can interfere with present social adjustment. This argument has struck a chord in the public mind. Following dramatic recountings of cases in which therapists claimed to have uncovered memories of sexual repression that had lain dormant for years some state legislatures passed legislation exempting prosecutions based on recovered memories from the normal statute of limitation.[2] Experimental psychologists are, to put it mildly, extremely skeptical about such reports. Loftus (1993; Loftus

& Ketcham, 1994) has made the case against memory with exceptional skill and vigor. Because of the understandable emotionality of the topic the nature of the argument between the clinicians and the laboratory scientists should be stated clearly.

The argument revolves around two points; the criteria for evidence and the criteria for memory. Clinicians rely a great deal upon case reports. Many of these are situations in which they (the clinicians) have personally been involved in recovering an alleged memory. The circumstances of the recovery are often very emotional, for both clinician and client, and can result in clients stating that "now they understand." As part of folk psychology, the more vividly an event seems to be recalled, the more believable it is. Experimental psychologists are highly suspicious of such evidence. They point out that the recovered memories are usually uncorroborated by external evidence, and that the clinicians often have beliefs about repressed memories that, with the best will in the world, make the clinicians prone to accept possibly fictitious accounts. Going somewhat further, experimental psychologists have worried that the client-therapist setting, in which the client is supposed to be receptive to the suggestions of the therapists, will lead to the encouragement of report of a repressed memory if the *therapist* believes that such memory exists. A clients conformity to the therapists wishes might itself constitute an example of unconscious reasoning, a point that has been little explored. The general point remains. The clinical evidence offered for complex unconscious reasoning is often simply rejected by laboratory psychologists as being, in the legal sense, incompetent. This does not mean that the experimental psychologists are saying that the clinicians are incompetent *as therapists.* It simply means that when a clinician testifies that, say, a thirty year old client was raped as a child, the clinician is testifying to something he or she cannot possibly know.

The clinicians respond that their testimony is based upon the behavior of their clients, which they can observe, augmented by a theory of how that behavior came to be. Here an analogy may be useful, using evidence that is every bit as emotional as a clinical case report. On January 8, 1996, after a NATO enforced truce had been declared in the Bosnian war, a civilian bus was blown up in the Muslim area of Sarajevo. NATO forces announced that the bus had been destroyed by a rocket fired from a particular apartment in a Serbian occupied region of the city. NATO observers had not seen the rocket fired. They examined the cavity created by the explosion, and traced the flight path of the rocket backwards, using ballistic calculations based on Newtonian physics. The logic of the NATO artillerymen exactly parallels the logic of the clinicians who

testify about repressed memories; observations in the present augmented by a theory of how the present state might come to be lead to a conclusion about events in the past. The difference is that there is rather more evidence for the accuracy of Newtonian dynamics than Freudian psychodynamics, so the artillerymen have a stronger case than the clinicians.

This example illustrates the major contention of the current paper. The distinction between conscious and unconscious reasoning cannot be made without a commitment to some general theory about how thinking proceeds. The remainder of this paper is a consideration of how our present computational theories of the mind stand on the issue of conscious vs. unconscious thought. Two classes of computational theories will be considered, *blackboard models* and *connectionist models*. There are numerous submodels within each of these theoretical approaches, and the approaches themselves can be reconciled somewhat. Nevertheless, I think that the approaches are sufficiently distinct and sufficiently broad so that they cover most of the theories in present day cognitive psychology.

Blackboard models described

A great many current psychological theories can be fit under the rubric *blackboard models* or *production system* models of mental action. Production system modeling was first introduced into psychology by Allen Newell, Herbert Simon, and their collaborators at Carnegie-Mellon university (Newell & Simon, 1972; see the history given by Neches, Langley, & Klahr, 1987). The modeling technique has been widely adopted by throughout both artificial intelligence and cognitive psychology . As blackboard models were explicitly developed using the organization of computing systems as a metaphor for mental action, I will first describe the models in computational terms, and then present psychological interpretations and motivations for adopting particular varieties of blackboard models over others.

A blackboard model consists of two parts; the *blackboard* and the *production memory* (Figure 2). The blackboard contains data structures describing what is going on at this moment, while the production memory contains rules specifying what actions to take when particular situations occur on the blackboard. These rules are stated as *productions*. They take the form *pattern - action* , which should be read as "if the appropriate pattern appears on the blackboard, take the following action. " Productions are organized into *production systems* that pro-

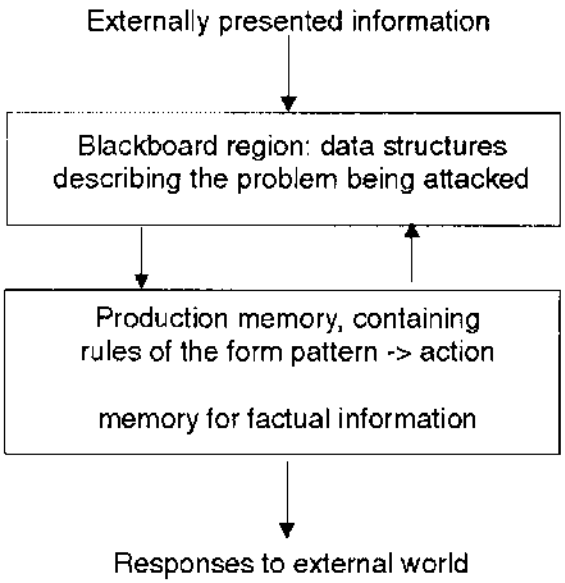

Figure 2. An expansion of the Hunt & Lansman (1986) blackboard model.

duce coherent actions by passing notes to each other via the blackboard. The flavor of the approach can be grasped by the following rules for driving an automobile:

(1.1) Red light —> stop
(1.2) Green light —> go
(1.3) Yellow light AND goal to be cautious —> stop
(1.4) Yellow light AND no goal to be cautious —> go.

In this example the rules refer to signals that appear on the blackboard. In psychological terms, this means the perception of a red (or green or yellow) light, not the sensation of one. In computational terms, it means that the data structure "Light color = red (green, yellow) " must appear on the blackboard.

Rules (1.3) and (1.4) also require that the blackboard contain a data structure indicating the goal to be realized. For instance, suppose that production memory contains the rule

(1.5) Police car observed —> place goal to be cautious on blackboard.

Rules (1.3) to (1.5), in combination, permit alteration of behavior depending upon context.

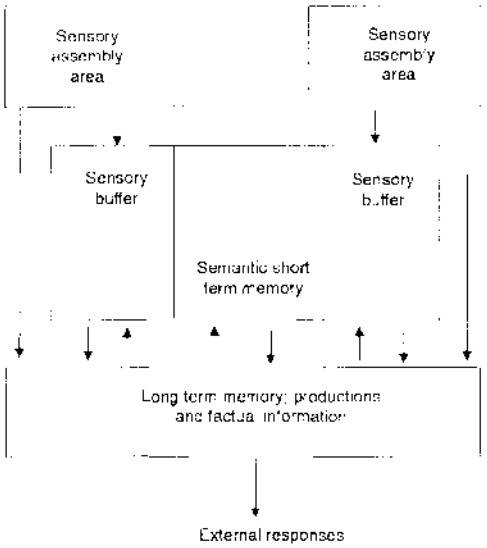

Figure 3. A threshold element as defined in connectionist models.

Blackboard models of thought are very widespread in psychology, even though the terminology used sometimes disguises them. For instance, the buffer models proposed by Atkinson and Shiffrin (1969) and Raajimakers & Shiffrin (1981) can be thought of as a specialized blackboard models developed to explain the activity exhibited in laboratory studies of simple memory situations. The blackboard idea is much more general. In fact, some of the first uses of blackboard models in psychology dealt with such complex topics as chess playing and algebra problem solving (Newell & Simon, 1972). While these applications display the power of the modeling technique the processes involved are so obviously tied to conscious problem solving that they do not illuminate any study of the special privileges of conscious and unconscious thought. In problem solving studies the participant is assumed to have read the problem. When we contrast conscious and unconscious reasoning we have to have some idea of what happens when the participant hasnt read the problem... or at least, does not think he has.

To deal with this issue I will propose an amplified version of a blackboard model that Marcy Lansman and I introduced in an attempt to encompass both complicated problem solving and some laboratory experiments on attention within a single model (Hunt & Lansman, 1986). An expansion of this model is shown in Figure 3. Let us step through its details.

In the (amplified) Hunt-Lansman model the blackboard is divided into three separate regions, visual and auditory buffers, and a more general "semantic" working memory area. These are shown in the middle box in Figure 3, and collectively constitute "working memory" as I shall refer to it. The distinction between semantic memory and the sensory buffers is almost identical to Baddeleys (1986) distinction between working memory and the echoic and visual-spatial scratch pads. Note that the buffers receive information either from the sensory system or from long term memory. The semantic region, by contrast, receives information only from long term (production) memory. The sensory buffers are preceded by sensory assembly areas. These areas are driven only by the sensory system itself, but can address long term memory. Thus the assembly areas do not act as a memory that can be maintained over time. The result is that the visual and auditory buffers within the blackboard contain sensory information, as modified by the long term system, and the semantic part of the blackboard contains a meaningful interpretation of that information.

The amplified Hunt-Lansman model differs from most Artificial Intelligence models, but is similar to many psychological models, in the functioning of long term memory. As in all production systems, the productions in long term memory essentially look at working memory, so that their actions will be taken when their pattern is detected in the blackboard area. *Detection* is thought of as a continuous rather than a discrete event. A pattern is detected when some data structure in working memory is close enough to the ideal pattern in the production so that the match between data structure and pattern exceeds the patterns current threshold. This brings us to the next point; how are thresholds set?

Each pattern is assumed to have a resting threshold determined by the patterns frequency of activation. In addition, productions are tied together, on the basis of previous histories of having fired in near simultaneity. This means that activation can passed from one production to related productions without necessarily passing through working memory. Similar ideas are contained in many models of this sort, particularly the work of John Anderson and his colleagues on their ACT* model (Anderson, 1983; 1993) I do not want to argue over the minor, and probably undecidable, details of differences between proposals such as ACT*, Baddeleys conceptualization of working memory, and my own notions, as presented here. The important things to stress are the common characteristics of the entire class of models.

Blackboard models provide for two routes for mapping a stimulus from input to output. One can be thought of as a note-passing method, in which patterns are passed from one production to another, using the blackboard as a

bulletin board, just as its name implies. The second is an activation passing method, in which activation flows through the network of productions. Activation passing is entirely a function of long term memory. The two systems are interdependent. If a pattern receives enough activation, via long term memory network mechanisms, so that its threshold is exceeded, then the associated action will be taken. A new data structure will be placed on the blackboard, altering the pattern of activation of the productions in long term memory, leading to new passages of activation, etc., etc. The model also allows for situations in which activation is passed from unit to unit on a *sub rosa* basis, without any pattern matching from the blackboard. From the viewpoint of an individual production, activation is activation. It does not matter whether it comes from a pattern match driven by the blackboard or whether it is comes from another production in long term memory.

I argue that what we conventionally refer to as "conscious thought" is computationally equivalent to the passage of information through the blackboard area, while unconscious thought is equivalent to computations that take place through the network of activations in long term memory. This is not an original idea; Baars (1988) has made a similar proposal. What I want to do here is to explore the implications of this idea for some recent studies of "unconscious vs. conscious thought," including a consideration of how this proposal might be modified in the light of developments of connectionist models.

The conscious-unconscious debate in terms of the blackboard model

Blackboard models provide a clearcut answer to General McPeaks problem; situational awareness is defined by the contents of the blackboard. In somewhat more scientific terms, context is defined by the data structures that are on the blackboard. Since a data structure may have been placed on the blackboard some time previously, this means that the effective stimuli for action are perceptions and interpretations, integrated over time. Since the blackboard is a limited respository we would expect context to be limited. However by directing attention to key aspects of a situation, and by rehearsal of key pieces of information, a problem solver can maintain situational awareness of appropriate contextual cues. This also allows for the common observation that situational awareness can be lost if irrelevant but attention grabbing distract a person from paying attention to the crucial aspects of the environment.

While a network of activated elements can maintain some context, the picture of what has happened that can be stored in a semantic network is much less clear than the sort of picture that can be stored on the blackboard. In particular, activation patterns do not provide any way for one production system, established by experience A, to communicate with a second system, established by experience B, unless the two systems have been linked in the past. On the other hand, if the stimuli originally associated with experience A appear in a new context, C, a combination of A-C cues in working memory may be sufficiently similar to the cues of experience B to arouse the second production system. Thus the blackboard provides a way for rapid shifts in thinking that are characteristic of human problem solving characterized by insight or restructuring; the "Eureka" sorts of problems that appear to be particularly difficult to explain by appeal to associationist principles alone.

Preconscious processing occurs when stimuli that do not reach conscious awareness (at least at the level of reportability) influence subsequence processing of perceivable stimuli. The ability to simulate this sort of phenomena is clearly built into the blackboard model of Figure 3, since information in the sensory assembly channels can contact long term memory, and hence influence the current level of activation of productions, without necessarily ever reaching the stage of conscious action, in a sensory buffer.

A similar argument can be used to account for a common perceptual process, imaging and its influences on perception of stimuli that are being sensed. The argument given here is that any response to a perceived stimulus will be a response to a data structure in a temporary buffer that may have been produced either by sensory input or by activation or long term memory systems or, as is most likely the case, by both these sources of information. Imaging studies and neuropsychological research on selected cases of brain injury have shown that there indeed are areas in the visual system, at least, which have this characteristic. These centers can be driven either by sensory input or by activation of long term memory when a person is asked to do an imagery task. As would be expected, parallel evidence exists showing that the act of imaging can either interfere with or enhance perception, depending upon whether or not the nature of the image enhances or detracts from a search for key features of the percept (Kosslyn, 1994).

In order to deal with the phenomena associated with implicit memory the basic model has to be augmented by learning mechanisms. This has been done by a variety of authors interested in blackboard models (Anderson, 1993, Hunt & Lansman, 1986; Newell, 1990). The first learning mechanism is directed at

computations involving the blackboard itself. Consider the situation just after some problem has been solved, using conscious reasoning mechanisms. At that time the blackboard will contain a description of the problem solving technique used and a record of the context in which the problem occurred. I will assume a copying mechanism that writes into long term memory whatever information is needed to associate successful problem solving methods with the context, *as it is represented in working memory at the time*. The qualification is important. The information stored in memory will be determined by the problem solvers analysis of the problem, not by the analysis of an external observer (or teacher). This could account for the discouraging specificity of much school learning. On the other hand, if a person was given experience with situations in which the same responses were required in multiple situation a single problem solving method could become associated with a variety of contexts. The resulting memory system could mimic the behavior of a system that was organized toward dealing with the world in terms of rules for abstract classes (Estes, 1987). so knowledge of classes of situations where particular problem solving methods are appropriate could be produced by a memory system that only stored specific experiences... providing that it stored enough of them.

Anderson (1983) and others have observed that there has to be room for a second type of learning that is associated with actions in long term memory alone, without the intervention of a blackboard copying device. Aristotles idea of associationism has its merits! It is probably the case that an activation-passing link will be established between two productions, A and B, whenever activity in the first production (A) is a consistent predictor of activity in a second production (B). This would produce highly specific, production to production, learning, and would not generalize. On the other hand, such learning would be sensitive to the specific cues, and so would provide a capacity for the execution of stereotyped response sequences almost autonomously from the analysis of information on the blackboard. It may be that a good deal of our learned motor skills, such as maintaining ones balance on a bicycle, eventually reach this state.

Now let us turn to the problems raised by the implicit memory studies. The studies cited earlier on stem completion and process dissociation indicate could easily be mimicked by a model in which information was presented, reached the blackboard, but did not pass through the copying mechanism. To be specific, suppose that a person is presented with the list of words DOG BAR MOTOR CHECK BLUE, and that for some reason MOTOR, although clearly perceived, does not pass through the copying mechanism. This means that MOTOR would not be associated with the context of the list. On the other hand, since MOTOR

was recognized at the time of presentation, the memory trace of the word MOTOR should be in a (temporarily) heightened state of activation at the end of learning. Therefore if MOT___ is presented as a stem completion item shortly after the list containing MOTOR had been presented, MOTOR would be a more probable response than usual until the trace activation decayed to its resting level. Note that this would be true in both the inclusion and exclusion conditions of a process dissociation experiment (see above), and so could provide an explanation of an important class of implicit memory studies.

The blackboard model provides a functional explanation for a number of retrospective amnesic syndromes, such as Korsakoffs disease or the effects of damage to the hippocampus. The interpretation is that in these cases what is lost is the ability to copy information from the blackboard into long term memory. This would produce an individual who could follow arguments that were within the time frame provided by working memory itself (e.g. could understand a sentence), and who could produce the sorts of phenomena that are usually associated with implicit memory. However the same individual could not perform tasks requiring the association of an event or object with a specific context. This would essentially disable the individuals ability to construct an autobiographic record.

Finally, how much intelligence will a blackboard model allow to unconscious processing?

A good argument can be made out that intelligent reasoning is closely linked to the ability to make discriminations between different situations. If we assume that the defined situation is what is on the blackboard, conscious problem solving can be thought of as the ability to take actions conditional upon the context currently held in the blackboard. It follows that good problem solvers, elegant language users, and good learners should be the people who have "big blackboards," usually operationalized as fairly large working memories. There is excellent evidence that this is the case. For instance, Kyllonen & Christal (1990) have shown that people who do well on working memory tests are superior at solving electrical circuit analysis problems. Dark & Benbow (1994) and Hunt, Frost, & Lunneborg (1973) have shown that short term memory superiority is associated with arithmetic facility, especially if the material to be remembered is itself numerical and, presumably, efficiently coded by those facile at mathematical problems. In experimental paradigms it has been shown that anything that distracts individuals from maintaining a proper context in working memory will distract from learning of sequences of actions or comprehending sentences (Nissen & Bullemer, 1987).

Problem solving that does not rely on blackboard information... in the model, "unconscious" problem solving... will be much more limited, since it relies on stimulus - response sequences that must be triggered by reoccurrence of the stimulus that was presented when the original learning occurred. For instance, semantic responses to a word could become associated with auditory or visual presentations on an automatic basis, because such responses are consistent in the world. On the other hand, there would normally be no opportunity to automate responses to a sentence, because sentences are typically understood only in context. This is the theoretical support for Greenwalds (1992) contention that the unconscious can respond to single words, but not very much more.

An interesting point is that the effect depends upon the context dependent interpretation of sentences, and not sentence length, per se, that limits non-conscious processing here. It should be possible to elicit unconscious responding to a multiple word string in cases where the "sentence" is actually a long word that, once begun, is always finished the same way. An example would be utterances such as "I pledge allegiance to the ..." However these are not too common. It may well be that Greenwalds conclusion that the unconscious deals only with single words is true, not because of the way the unconscious is, but because of the way that the world is.

The role of connectionism

Blackboard models were constructed using the idea of computing as a metaphor for the mind. It has been claimed that a computing metaphor is necessary (Pylsyhyn, 1989) and, contradictorily, that a computing metaphor is only one of several possible metaphors (Rumelhart, 1989). An alternative to the computational metaphor is the development of *connectionist* or *neural net* models of thought. In considering what a connectionist model is it is useful to review, briefly, what a model is.

Any reproducible human behavior can be thought of as a mapping from a set of stimuli to a set of possible responses. In spite of the terminology, this observation does not return us to the stimulus-response associationism of the pre 1960 era. The terms "set of stimuli" and "set of responses" are simply psychological jargon for the mathematicians terms *domain* and *range* of a function. Thus we can think of any precisely defined mathematical model of behav-

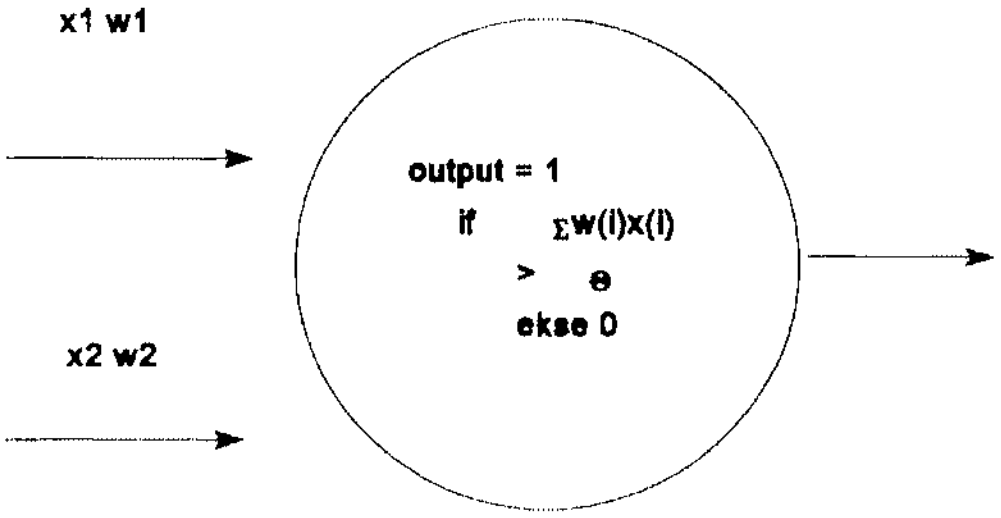

Figure 4. A connectionist model that computes the XOR function. Input is presented on the left, and output on the right.

ior for finite stimulus-response functions as a mapping from a vector defining the stimulus to a vector defining the response.

Now consider the element shown in Figure 4. Mathematically, this is simply an element that receives a set of weighted inputs from other, similar elements and computes a single response that is a non-decreasing function of the weighted sum of the inputs. A simple example is the threshold function,

(1) output = 1 if the weighted input threshold value Q, zero otherwise.

Functions can be chosen so that the unit in Figure 4 is at least a rough mathematical model of the input-output behavior of neurons. Now suppose that we designate a subset of these units as input units, and a second subset as output units. Providing no limit is placed on the number of units that intervene between input and output units, it is fairly easy to show that every possible function between input and output can be computed by at least one connectionist network. For instance, Figure 5 shows one (of several) connectionist networks that can compute the EXCLUSIVE OR function. The assertion that at least one connectionist network exists to compute each computable function, when proven, is a statement about the existence of a solution. Finding the appropriate network can be quite difficult. What is even more difficult is finding a network

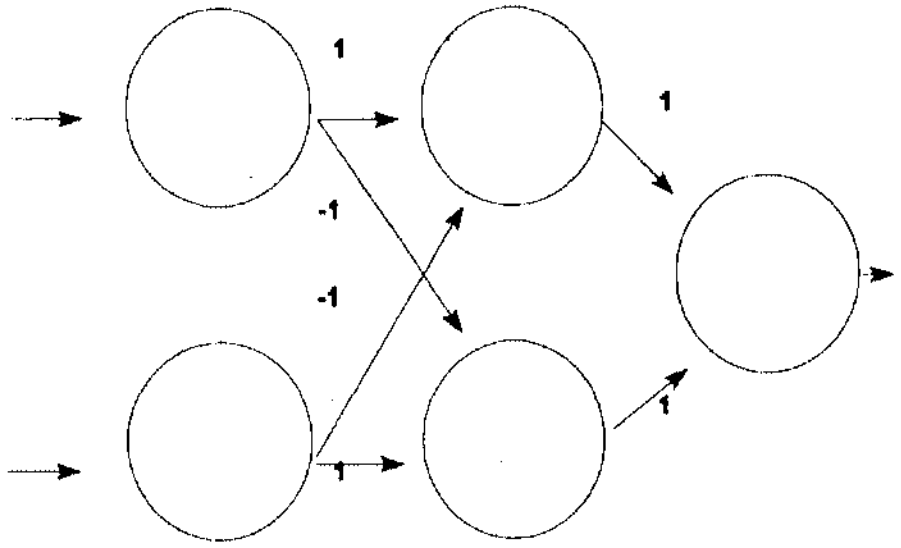

Figure 5.

that computes the required function and that satisfies some constraint on its organization that is at least analogous to the way that neurons are connected to each other in the brain. Nevertheless, advocates of connectionist modeling hope that this can be done. If they are correct, we are close to making a substantial step toward a materialist model of the mind.

Where do connectionist models stand on the question of conscious versus unconscious reasoning? Greenwald (1992) claims that connectionist models can be used to mimic the conscious-unconscious distinction. Here is a somewhat simplified summary of his argument.

When an input vector is offered to a connectionist network one of three things can happen. The input may be so weak that no interior units respond, i.e. nothing happens and there is no alteration of the network. Let x be the level of activity in the stimulus (input) unit, and let θ_1 be the threshold that must be exceeded to fire at least one interior unit. Next, suppose that we designate some subset of the output elements as "conscious report" elements, i.e. elements that must produce output if the network is said to mimic some criterion for consciousness, such as introspective self report. Let θ_2 be the level of activity in the input units that must be exceeded for at least one of the conscious report elements to be activated. Note that both θ_1 and θ_2 would vary over time, as a

function of the level of activity of elements in the network just prior to presentation of the stimulus of interest. Three things can happen.

(2.1) x_1_____. The stimulus is not recorded and has no effect on the network.

(2.2) x_2. The stimulus is perceived consciously. Network memory may or may not be altered.

(2.3) $_1x_2$. The stimulus is not perceived consciously but network activity and possibly alteration does occur.

This interpretation opens the door for all manner of conscious and unconscious processing just because it defines a machine with a functional architecture identical to the blackboard model. Greenwalds "consciousness producing" elements could be though of as defining working memory. If these elements contain, amongst themselves, reverberating loops they have the capability to maintain a context. Therefore the state of the working memory elements, including their definition of context, can produce output that could be used as by a memory writing function to retain memory of stimuli presented in a particular temporal and spatial context. If this interpretation of connectionism is accepted connectionist models of the conscious and unconscious thought simply inherit all the properties already asserted for blackboard models.

Viewed this way connectionism is not a new species of psychological theory. Rather, it is a proposal for a particular way of implementing a blackboard model. This, indeed, is the way that I regard connectionism. Not everyone agrees with me. John Anderson (1993), for instance, has argued that working memory should be thought of as the totality of network activation levels that are above a particular threshold. Two differences between this Andersons view of working memory and the one I have presented are relevant for the distinction between conscious and unconscious thinking.

Andersons view does not assign a privileged role to working memory elements. Since potentially any element in a connectionist network may participate in working memory functions, it follows that there cannot be any learning mechanism that depends upon working memory, unless that function is determined by the strength of activation rather than by its locus in the network. This assumption is difficult to reconcile with observations about amnesia. Korsakoffs syndrome and hippocampal patients lose the ability to record conscious experiences but retain an ability for implicit memory. Is it reasonable to assume that one can lose the ability to record strongly activated patterns of neural activity while retaining the ability to store weakly activated ones?

Andersons approach is incompatible with the idea that the functional blackboard is a place in the brain. In his approach working memory is a moving target, that could potentially be anywhere, depending upon the pattern of activation at the time. This is usually presented as a strength of the connectionist approach, but I believe that it is actually a weakness. The term *Parallel Distributed Processing (PDP)* has been used as an alternative to *connectionism* or *neural network modeling*. I believe, though, that this flies in the face of recent evidence from the cognitive neurosciences. Visual and auditory imaging are associated with specific locations in the brain that are also associated with perception of external stimuli . The frontal and prefrontal regions seem to be heavily involved whenever a task involves working memory. We know this both on the basis of imaging studies (Posner & Raichle, 1994) and animal studies involving controlled ablations (Goldman-Rakic, 1992). In fact, working memory tasks are very difficult for animals other than primates, and primates alone are characterized by substantial frontal lobes. (The bottlenosed dolphin, *T. Truncatus*, is a puzzling exception to this statement. It displays what appear to be short term memory abilities, but it does not have a large frontal lobe (Herman, Richards, & Wolz, 1984). Finally, it has long been argued that the inability to maintain situational awareness characteristic of attention deficit syndrome are associated with minor damage to the forebrain region (Das, Kirby, & Jarman, 1979).

From a strictly functionist viewpoint, whether one wants to approach conscious versus unconscious processes from a blackboard or connectionist approach is entirely a matter of convenience. The two are equivalent, because appropriate connectionist models will simply instantiate the functional architecture described in a blackboard model. The neuroscience evidence for a locus for short term memory is interesting, but it does not discriminate between connectionist and rule-based blackboard approaches. Perhaps cognitive psychologists should appeal to computer analogies once again. In computer science some people experiment with the design of circuits, others experiment with the design of modules. Neither approach is right or wrong, they are just useful for different things. The choice between rule based or connectionist models may well be guided by the same principle.

The computational properties of conscious thought

The argument presented thus far has centered on what a computer scientist would call "system architecture" issues. A case has been made for two types of

mental computation, one based on specialized processors and another governed by information that is 'on the blackboard, 'in working memory, or located in some other functionally equivalent module. The next step is to ask what differences there are between the sorts of computations that can be done with information that is or is not conscious. I shall argue that conscious thought permits three types of computations that are not available to non-conscious processes; computations based on a combination of sensory codes, computations that are conditional upon the temporal context of the situation, and computations that are directed at thought itself.

Baars (1988) has developed the first point, that conscious computations can receive input from numerous sources, in considerable detail. While acknowledging his priority, I will present a somewhat different argument. At the most primitive level, this property of consciousness is closely related to Treismans observation that conjunction searches, which require people to look for two different primitive visual features, appear to be conscious and are certainly attention demanding. (See above for details.) At a more general level, the argument is that only conscious computations can utilize information from different sense modalities. Somewhat surprisingly, I have found little information on this issue, but the assumption does follow from the theory.

The second point is that if a stimulus is consciously perceived the response to it can be conditional upon the current, temporally defined context. This has two important consequences. One is that conscious computations can be made conditional upon the goal that the actor is pursuing. This is a particularly important part of computer models of higher mental processes. In fact, Newell (1990) argued that new and innovative problem solving procedures will only be developed when a person realizes that his or her current attacks on a problem are not producing progress toward a goal. The second consequence is that conscious processes can produce redescriptions of problems in terms of abstractions that make sense. This is extremely important in human reasoning, because it appears to be the basis of transfer of training from one situation to another. In studies of animal learning generalization is almost uniformly based upon generalization of some perceivable aspect of the physical stimulus; size, color, or, in a few cases, perceivable relationships between elements, such as teaching an animal to approach the largest of two objects. In human education transfer is supposed to be based upon abstract redescriptions of objects. For instance, in teaching statistics to psychology students the instructor is not trying to train the students to analyze the data from observations on college students, babies, or rats. The instructor is trying to train students to apply abstract concepts such as

"contrasts based on two independent samples." Instructors are somewhat successful, even though students remain better able to apply statistical models to the areas in which they have seen examples than to other areas where the models apply (Fong & Nisbett, 1991). Even more dramatic examples of the use of abstract reasoning can be seen when we compare 'real experts (which certainly does not include students in psychological statistics classes!) to novices. Problems that elementary students see as one of a block sticking on an incline is seen by the expert physicist as a problem in balance of forces (Chi, Feltovich & Glaser, 1981).

The ability to reformulate the surface description of a problem into an abstract one requires that a person manipulate the objects of his or her own thought. This can only be done consciously, one has to be aware of what one is thinking about, and how one is thinking about it. This sort of reflective thought has been termed *metacognition.* While metacognition is a form of introspection, it can be studied without falling into the deep methodological problems that plagued the introspectionists of the 19th century. Two sources of evidence are important. One is the fact that introspective self-reports of thinking processes can be reliably, albeit not perfectly, related to publicly observable behavior. For instance, students who report redescribing and amplifying upon homework problem exercises appear, by objective criteria, to learn more from their lessons than people who simply work through the example (Chi et al., 1989). A number of other examples can be given to show that self-report of ones thoughts, although sometimes of questionable accuracy, does enter into orderly relationships with subsequent behavior (Ericsson & Simon, 1984) Since perfection is not a characteristic of measurement in psychology, it seems churlish to rule out such evidence on the grounds that ones metacognitive processes are not always perfect. Also, reliable and objective methods have been developed to assess the relationship between self report of a mental process and objective evidence of the execution of that process. As an example, there is a reliable correlation between peoples predictions as to whether or not they will be able to recall a fact they have memorized at some time in the future and their actual ability to do so. The size of this correlation can be altered by altering the extent to which the mental processes that are reported upon when the prediction is made resemble the processes that will be required when recall is actually attempted (Nelson, 1996).

Problems and objections

The basic thesis of this paper is that humans are capable of two distinct types of information processing, conscious and unconscious thought. Each of these thoughts is presumed to have distinct computational characteristics, with unconscious thought being quite limited. I will now consider some objections to these assertions.

Language phenomena present a major challenge to the claim that the unconscious is dumb. Syntactic analysis is particularly troubling. People perform quite involved syntactical analyses without being aware of them. Indeed, native speakers may be unable to articulate the rules of syntactic analysis that they are following. Fodor (1983) and most linguistics argue that this is because language is processed by specialized information-processing modules that function independently of the rest of the mind. If this is so, then the case of non-conscious syntactical analysis can possibly be ignored as a special property of our language mechanism. Recent research, however, suggests that linguistic analysis may be accomplished by general mechanisms, and hence be less modular than was once thought (MacDonald, Pearlmutter, & Seidenberg, 1994; Tannenhaus et al., 1995). If syntax analysis is done by a general mechanism, does that mechanism provide an ability for other brands of complex but non-conscious thought? We simply do not know.

There are reports of cases in which people have apparently performed quite complex reasoning, are apparently unaware of having done so, and yet have their subsequent actions influenced by the reasoning that they cannot recall. Kihlstrom, Barnhardt, & Tataryn (1992), in commenting on Greenwalds paper, cited studies of hypnosis and multiple personality syndrome as being particularly perplexing for anyone who wants to argue that the unconscious is necessarily unintelligent. It is important to realize, though, that in these cases information that clearly was conscious *in certain contexts* (the appropriate hypnotic state or appropriate dominate personality) is what was subjected to the complicated processing. Some of the results of that complicated processing leak through to influence conscious processing in other contexts. The issues are "what produces the walls between contexts? and "how much leaks through?" That will be a topic for further research.

Some philosophers of science are concerned about the cognitive science approach. Their concerns focus on two things; the apparent multiplication of intervening variables presumed to exist between stimulus and response and, more generally, the cognitive psychologists apparent lack of concern for the

philosophers of science characterize as tightly argued, well constructed theoretical statements. Both these concerns are real, and deserve a response.

Since its inception the cognitive psychology movement, and cognitive science in general, has been characterized by a proliferation of different ideas about the sorts of theoretical constructs needed to characterize the human mind. Philosophers of science are not alone in voicing their complaints. William Estes, one of the major developers of mathematical models of learning and memory, has complained that cognitive theories seem to undergo major expansion quite without any accompanying empirical data indicating a need for revision (Estes, 1991) . In some cases the revisions seem more driven by a desire to incorporate new concepts, thus showing that a theory is developing, than to incorporate new data. For instance, in his more recent work Anderson (1986, 1993) makes frequent references to a distinction between mental actions that are interpreted or compiled. These are terms borrowed from computer science, where the interpretation-compilation distinction is precisely defined. It is not at all clear what the distinction is supposed to mean in psychology.

Such free-wheeling borrowing of terms and ideas from one field to another is, to put it mildly, frowned upon by philosophers of science who wish words to be used precisely . To an outsider modern cognitive psychologists may seem to be following the philosophy of Humpty Dumpty rather than Wittgenstein; words are required to mean whatever the speaker wants them to mean in the context in which they are uttered. While there is some truth to these charges, there are defenses against them.

Philosophers of science and cognitive psychologists both agree that theories should be precise and that theories should serve as heuristics to suggest further investigations. These two goals are not always in agreement. Consider the current topic, consciousness. My argument has been that consciousness is characterized as one of two types of computation that the brain effects. There is an implicit assumption that the mind itself is nothing more than the computational capabilities of the brain. Philosophers can rightly object that in everyday discourse *consciousness* and *mind* carry more meaning than this. Harzem (1996) argues that these terms cannot refer to topics in science, precisely because they have been co-opted by humanists. Harzem worries that if an attempt is made to do science using humanistic terms scientists will begin with precisely defined operations, draw inferences from experiments using these operations, and then smuggle further meaning into the experiments by slipping from the precise usage of terms like consciousness and mind *as defined in the experimental setting* into the less precise way in which humanists use these terms. The charge is not

trivial. Explaining the conditions under which one does or does not see letter strings that are exposed for only fifty milliseconds is not the same as explaining consciousness, in the humanists terms. A philosopher of science might prefer terms like Computation1 and Computation2 instead of conscious and unconscious processing. Why do cognitive psychologists reject this argument? The answer to this question goes to the heart of the cognitive science movement, and can serve as a summary for the paper.

Summary: the cognitive psychology approach to consciousness

At the beginning of the cognitive revolution in Psychology, Newell, Shaw, and Simon (1958) stated that a theory of human thought should be equivalent to a design for a machine that was capable of thought. This idea has gone through numerous transformations. Cognitive psychologists today are far more concerned than they were forty years ago that any machine they design be at least consistent with modern findings concerning the physical and computational properties of the brain. Nevertheless, the idea that the business of cognitive psychology is to understand human thought by designing a thinking machine is still very much alive. For that reason, and because we want the machine to resemble what we know of the brain, it is extremely important to cognitive psychologists that we communicate with other relevant sciences, ranging from the neurosciences to economics. In order to do this we cannot develop a science, or a philosophy of science, that is based on a highly specialized, arcane vocabulary of technical terms. We have to talk about thought in ways that will match the way that others talk about thought. To do otherwise would leave us solely talking to each other, and we want to and must talk to the related sciences. While cognitive psychologists are virtually all thorough materialists at heart, in their daily life they find it useful to use dualistic terms such as consciousness and mind.

In particular, the distinction between conscious and unconscious thought can be made precise by step by step explorations of the different sorts of computational capabilities that are associated with different operational definitions of consciousness. I have presented some of the key experiments here, and pointed out that they lead to the conclusion that unconscious thought is limited to highly specialized, context free information processing, while conscious thought can yield much more flexible reasoning. The fact that there is a parallel between behavioral evidence of things that we are willing to define as conscious

thought and the activities of certain centers in the brain does not strike us as a confusion between definitions based on location and definitions based on function, it strikes us as a very important piece of evidence to be used to guide our construction of a machine that models the brain and the mind. As these parallels between types of thinking and brain activity are further developed it is quite likely that cognitive psychologists will both further sharpen the distinction between conscious and unconscious thought and will defend this distinction as one that has real meaning in brain operations, rather than being an abstract entity in a psychological theory.

Thinking of conscious and unconscious thought as particular types of computations, guided by certain aspects of the system architecture of the mind, helps us understand both the intricacies of human thought and, in some cases, actually to design man-machine systems that rely on one or the other type of thinking at different points. To cognitive scientists this is the important criteria for a theory. We ask "Do these approaches help us get on with our work, in both theoretical and applied domains?" rather than asking "Do our investigations conform to those methods of inquiry that are approved of by philosophers of science?"

To conclude, theoretical mechanisms are important. We ought to be suspicious of assertions about facts when those facts have no explanation within the range of well justified theories of the appropriate phenomena. This does not mean that we want to exercise total theoretical arrogance. It is certainly true that the absolutely most interesting facts are those that demand a re-ordering of our theories. However such findings are not so common as romanticists of science would have us believe. Assertions about distinctions between conscious and unconscious mental processing are assertions about the computational capabilities of different states of the mind-brain system. Reconciling these assertions with computational theories of the mind is an essential step in understanding the data.

References

Anderson, J.R. (1983) *The Architecture of Cognition*. Cambridge, MA. Harvard U. Press.

Anderson, J.R. (1986) Knowledge Compilation: The General Learning Mechanism. in R. S. Michalski, J.G. Carbonell, & T.M. Mitchell (eds.) *Machine Learning II*. Los Altos, CA. Morgan Kaufmann.

Anderson, J. R.(1993) *Rules of the Mind* Hillsdale, NJ. L. Erlbaum Assoc.

Atkinson, R. C. & Shiffrin, R. M. (1968) Human Memory: A proposed system and its control processes. in K.W. Spence & J.T. Spence (dds.) *The Psychology of Learning and Motivation: Advances in Research and Theory.* (Vol.2). New York: Academic Press.

Baars, B.J. (1989) *A Cognitive Theory of Consciousness.* Cambridge: Cambridge U. Press.

Baddeley, A.D. (1986) *Working Memory.* Oxford: Oxford U. Press.

Baddeley, A.D. (1992) Working Memory. *Science, 255*, 556-559.

Bass, E. & Davis, L. (1989) *The Courage to Heal: A Guide for Women Survivors of Child Sexual Abuse* New York: Harper Collins.

Chi, M.T.H., Bassok, M., Lewis, M.W., Reinman, P. & Glaser, R. (1989) Self explanation: How students study and use examples in learning to solve problems. *Cognitive Science, 13* (2) 145-182.

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981) Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science 5* (2) 121-152.

Chiessa, M. (1996) Cause, explanation, and theory in the science of behavior. Paper delivered at the 4th Biennial Conference on the Science of Behavior. Chapala, Mexico. Feb. 1996.

Crick, F. (1994) *The astonishing hypothesis.* New York: Scribner.

Dark, V.J. & Benbow, C.P. (1994) Type of stimulus mediates the relationship between working memory performance and type of processing *Intelligence, 19 (3)* 337-358.

Das, J.P., Kirby, J. & Jarman, R.F. (1979) *Simultaneous and Successive Cognitive Processes.* New York: Academic Press.

Ericsson, K.A. & Simon, H.A. (1984) *Protocol Analysis: Verbal Reports as Data.* Englewood Cliffs, NJ: Prentice-Hall.

Estes, W.K. (1987) Array models for category learning. *Cognitive Psychology, 18 (4)* 500-549.

Estes, W.K. (1991) Cognitive architectures from the standpoint of an experimental psychologist. *Annual Review of Psychology. 42*, 1-28.

Fodor, J.A. (1983) *The modularity of the mind.* Cambridge, MA. MIT Press.

Fong, G.T., & Nisbett, R.E. (1991) Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General, 120*, 34-45.

Fowler, C.A., Wolford, G., Slade, R. & Tassinary, L. (1981) Lexical access with and without awareness. *Journal of Experimental Psychology: General. 110,* 341-362 .

Goldman-Rakic, P.S. (1988) Topography of cognition. Parallel distributed networks in primate association cortex. *Annual Review of Neuroscience,* 11, 137-156.

Greenwald, A.G. (1992) New Look 3: Unconscious cognition reclaimed. *American Psychologist, 47 (6)* 766-779.

Harzem, P. (1996) The craft of understanding the mind. Why it cannot be a science. Paper delivered at the 4th Biennial Conference on the Science of Behavior. Chapala, Mexico. Feb. 1996.

Herman, L.M., Richards, D.G. & Wolz, J.P. (1984) Comprehension of sentences by bottlenosed dolphins. *Cognition, 16,* 129-219.

Haber, R.N. (1983) The impending demise of the icon. A critique of the concept of iconic storage in visual information processing. *Behavioral and Brain Sciences. 6,* 1-54.

Hunt, E. (in press). What is a theory of thought? In R.J. Sternberg, Ed. *The concept of cognition.* Cambridge, MA. MIT Press.

Hunt, E. & Lansman, M. (1986) A Unified Model of Attention and Problem Solving. *Psychological Review, 93,* 446-461.

Jacoby, L.L., Toth, J.P., & Yonelinas, A.P. (1993) Separating conscious and unconscious influences on memory: measuring recollections. *Journal of Experimental Psychology: General, 122 (2)* 139-154.

Kihlstrom, J.F., Barnhardt, T.M., & Tataryn, D.J. (1992) The psychological unconscious: Found, lost, and regained. *American Psychologist, 47 (6)* 788-791.

Kosslyn, S.M. (1994) *Image and mind.* Cambridge, MA. MIT Press.

Kosslyn, S.M. & Koenig, O. (1992) *Wet Mind: The new cognitive neuroscience.* New York: Free Press.

Kyllonen, P.C. & Christal, R.E. (1990) Reasoning ability is (little more than) working memory capacity ?! *Intelligence, 14,* 389-433.

Livingstone, M. & Hubel, D. (1988) Segregation of form, color, movement, and depth; Anatomy, physiology, & perception. *Science, 240* 740-749.

Loftus, E.F. (1993) The reality of repressed memories. *American Pschologist, 48,* 518-537.

Loftus, E.F. & Ketcham, K. (1994) *The myth of repressed memory: false memories and allegations of sexual abuse.* New York: St. Martins Press.

Loftus, E.F. & Klinger, M.R. (1992) Is the unconscious smart or dumb? *American Psychologist, 47 (6)* 761-765.

Marcel, A.J. (1983) Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology, 15,* 197-237.

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

McClelland, J.L. (1979) On the time-relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287-330.

McDonald, M.C., Pearlmutter, N.J. & Seidenberg, M.S. (1994) Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101 (4)* 676-703.

Neches, R., Langley, P. & Klahr, D. (1987) Learning, Development, and Production Systems. in Klahr, D., Langley, P., & Neches, R. (Eds) *Production system models of learning and development.* Cambridge MA. MIT Press (pp.1-53).

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

Nelson, T.O. (1996) Consciousness and Metacognition. *American Psychologist, 51* (2) 102-116.

Newell, A. (1980) Physical Symbol Systems. *Cognitive Science, 4,* 135-183.

Newell, A. (1990) *Unified Theories of Cognition*. Cambridge, MA. Harvard U. Press.

Newell, A., Shaw, J.C., and Simon, H.A. (1958). Elements of a theory of human problem solving. *Psychological Review, 65* (3) 151-166.

Newell, A. & Simon, H.A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ. Prentice-Hall.

Nissen, M.J. & Bullemer P. (1987) Attentional Requirements of Learning: Evidence from performance measures. *Cognitive Psychology, 19,* 1-32.

Posner, M.I. & Raichle, M.E. (1994) *Images of Mind.* San Francisco: Freeman.

Pylyshyn, Z.W. (1989) Computing in Cognitive Science. in Posner, M.I. (Ed.) *Foundations of Cognitive Science*. Cambridge, MA. MIT Press, pg. 51-91.

Raaijmakers, J.G. & Shiffrin, R.M. (1981) Search of associative memory. *Psychological Review 88 (2)* 93-134.

Reason, J. (1990) *Human Error.* Cambridge: Cambridge U. Press.

Richardson-Klahaven, A. & Bjork, R.A. (1988). Measures of memory. *Annual Review of Psychology, 39,* 475-543.

Rumelhart, D.E. (1989) The architecture of the mind: A connectionist approach. in Posner, M.I. (Ed.) *Foundations of Cognitive Science*. Cambridge, MA. MIT Press, pg.133-159.

Sperling, G. (1960) The information available in brief visual presentations. *Psychological Monographs 74*, (Whole No. 11).

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M. & Sedivy, J.C. (1995) Integration of visual and linguistic information in spoken language comprehension. *Science 268* 1632-1634.

Treisman, A. (1988) Features and objects: The Fourteenth Bartlett Memorial Lecture. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology. 40A* (2) 201-237.

Treisman, A. & Gormican, S. (1988) Feature analysis in early vision: Evidence from search assymetries. *Psychological Review, 95* (1) 15-48.