Conscious and Unconscious Perception: A Computational Theory

Donald W. Mathis and **Michael C. Mozer**
Department of Computer Science & Institute of Cognitive Science
University of Colorado at Boulder
Boulder, CO 80309-0430
mathis@cs.colorado.edu, mozer@cs.colorado.edu

Abstract

We propose a computational theory of consciousness and model data from three experiments in visual perception. The central idea of our theory is that the contents of consciousness correspond to temporally stable states in an interconnected network of specialized computational modules. Each module incorporates a relaxation search that is concerned with achieving semantically well-formed states. We claim that being an attractor of the relaxation search is a necessary condition for awareness. We show that the model provides sensible explanations for the results of three experiments, and makes testable predictions. The first experiment (Marcel, 1980) found that masked, ambiguous prime words facilitate lexical decision for targets related to either prime meaning, whereas consciously perceived primes facilitate only the meaning that is consistent with prior context. The second experiment (Fehrer and Raab, 1962) found that subjects can make detection responses in constant time to simple visual stimuli regardless of whether they are consciously perceived or masked by metacontrast and not consciously perceived. The third experiment (Levy and Pashler, 1996) found that visual word recognition accuracy is lower than baseline when an earlier speeded response was incorrect, and higher than baseline when the early response was correct, consistent with a causal relationship between conscious perception and subsequent processing.

Introduction

In recent years there has been a resurgence of interest in the scientific study of consciousness. Experimental approaches have studied subliminal perception (e.g., Greenwald et. al, 1995), implicit memory and learning (e.g., Hintzman, 1990), neuropsychological dissociations between knowledge and awareness (e.g., Shallice, 1988), and most recently, the problem of finding the *neural correlates* of consciousness (Crick & Koch, 1990). In contrast, we take a computational approach, asking the question: What happens differently in the brain when one is processing information consciously versus unconsciously? We propose a *computational correlate* of consciousness, as part of a cognitive architecture, and evaluate the theory by accounting for experimental data. The theory is motivated by several basic experimental findings, including: (1) conscious percepts are interpretations of stimulus information (Kanizsa, 1979; Kolers & Von Grunau, 1976; Warren, 1970); (2) people are aware of the *results* of higher cognitive processes, not the processes themselves (Nisbett & Wilson, 1977); (3) cortex appears to be functionally *modular* (Felleman & Van Essen, 1991), but consciousness appears to be distributed in the brain (e.g., Young & DeHaan, 1991).

The Stability Theory of Consciousness

We propose a *stability theory of consciousness* that is based on a few simple principles. First, the human cognitive architecture consists of a set of functionally specialized, interconnected computational modules. Each module functions as an associative memory in its domain, outputting the best-fitting interpretation of its input, subject to a set of constraints that define reasonable entities in the domain. For example, a visual object-recognition module might output interpretations consistent with constraints governing realizability in three-dimensional space. Second, the operation of a module is a two-stage process. Each module first maps its input to an initial output in a fast *mapping* process. Then a slower, iterative *relaxation search* process transforms this output to one that is *well-formed*—i.e., satisfies the domain constraints. Third, the central hypothesis is that *temporally stable states* enter consciousness. That is, the stable output of the relaxation search process of *any* module enters consciousness. This implies that there is no special "consciousness module"; the contents of consciousness are distributed amongst the modules.

We embody the theory in a connectionist model (Figure 1). The mapping process is implemented by a feedforward network, which has the desired property of producing an output quickly, without iteration. The relaxation search process is implemented by an attractor or constraint-satisfaction network. Attractors of the net are the well-formed states, i.e., interpretations produced by the module. Fully distributed attractor networks have been used for similar purposes (e.g., Hinton & Shallice, 1991), but for simplicity we employ a localist-attractor architecture with a layer of *state* units and a layer of radial basis function (RBF) units, one RBF unit per attractor. The state units receive input from the mapping network and from the RBF units, and are updated with an incremental activation rule

$$s_i(t+1) = h\Big(s_i(t) + \alpha e_i(t) - \beta r_0(t) + \gamma \sum_{j \in RBF} r_j(t) a_{ji}\Big)$$

where $s_i \in [-1,1]$ is the activity of the ith state unit, e_i is the input to that unit from the mapping net, r_j is the activity of RBF unit j, a_{ji} is the value of component i of the jth attractor state, α, β and γ are small positive constants, and $h(.)$ is a function that simply bounds activity between -1 and $+1$. The zero'th rbf unit, r_0, corresponds to a special *rest state* located at the origin of the state space, in which the system resides at the start of each simulation. The RBF units use the update rule:

$$r_j(t) = exp(-\|s(t) - a_j\|^2/\beta)$$

where β is the *width* of the RBF. We quantify stability as an exponentially-decaying time-average of the reciprocal of the speed of the state vector:

$$stability(T) = \sum_{t=0}^{T} (\lambda^{(T-t)} exp(-\|s(t) - s(t-1)\|))$$

where $0 < \lambda < 1$ is a parameter controlling the window of the time average. When this quantity exceeds a threshold, we say that the state has stabilized.

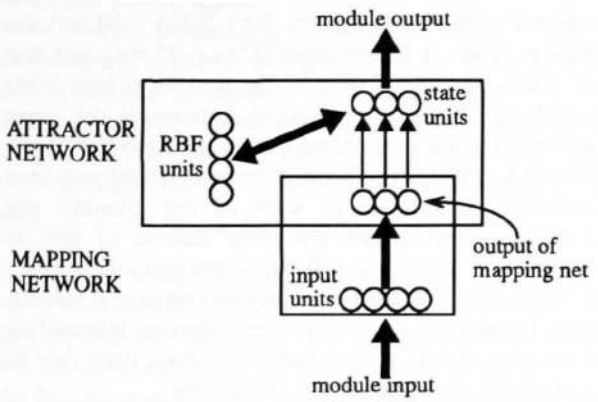

Figure 1: Architecture of a module. Heavy lines denote full connectivity between units in two layers. Thin lines denote copying of activity.

Modeling Experimental Data

Subliminal Semantic Priming & Lexical Ambiguity

Marcel (1980) showed subjects a sequence of three words—the *context*, *prime* and *target*—and instructed them to perform lexical decision on the target. The prime was ambiguous, with two unrelated meanings. There were three experimental conditions relevant to our theory. In the *congruent* condition, the context word was semantically related to one meaning of the prime, and the target word was related to that same meaning. The *incongruent* condition is the same except the target is related to the *other* meaning of the prime. In the *unassociated* condition, the context and target words are unrelated to either meaning of the prime. The prime was either presented for 500 ms (the *conscious* condition) or for roughly 10 ms and backward-masked to prevent awareness (the *unconscious* condition).

Table 1: Results for simulation of Marcel (1980)

	human data		simulation	
	conscious	unconsc.	conscious	unconsc.
congruent	499 ms	511 ms	44.6 cycles	46.0 cycles
incongruent	547 ms	520 ms	47.9 cycles	46.8 cycles
unassociated	541 ms	548 ms	48.0 cycles	47.9 cycles

When the prime is consciously perceived, Marcel found facilitation for the congruent condition relative to the unassociated condition, but not for the incongruent condition. However, if the prime is not consciously perceived, there is facilitation in both the congruent and incongruent cases. Thus, in the unconscious case, both meanings of the prime are available to facilitate targets, but in the conscious case, only the meaning consistent with the prior context is available. These results show a correlation between the onset of awareness and the selection of one meaning.

Modeling Word Reading, Ambiguity Resolution, Masking, and Priming. Before describing our account of the data, we explain how we model word reading and priming in general. A single module is used to map orthography to meaning (Hinton & Shallice, 1991). Attractor states in this module correspond to meanings of known words, and the rest state is used to represent the state "no meaningful input." The time-course of processing is as follows: With the relaxation net initialized to the rest state, an input is presented, and activity flows through the mapping net, providing input to the state units. This input causes the state units to move into the attractor representing the meaning of the input word. Simple one-word disambiguation occurs as follows: We assume that ambiguous word patterns are associated with multiple attractors, and that the mapping network outputs a semantic pattern that falls between them in semantic space. In the process of settling to one attractor, the relaxation net will effectively *select* one of the meanings.

We model the effect of backward masking by removing input from the word recognition module. If an input is presented briefly, the relaxation net will settle back to the rest state. Since stable (and non-rest) states correspond to conscious perception in our theory, this would correspond to no conscious perception of the input.

To model priming, we adopt a variant of the method used by Becker et al. (1993) of *strengthening* attractors that the system visits during processing. After the network settles, we increase the β (width) parameter of each RBF

unit in rough proportion to its activity:

$$\Delta\beta_i = eligibility_i / \sum_{j \in RBF} eligibility_j$$

where

$$eligibility_i = \sum_t exp[-\|s(t) - a_i\|^2] .$$

The eligibility of each RBF is normalized to prevent the amount of priming to increase without bound for long exposures.

Simulation. For simplicity, we simulated only the attractor network of the module. Twenty attractor states representing different semantic concepts were randomly chosen. These states were corners of a 50-dimensional hypercube. Semantic relatedness was achieved via overlap in the state vectors. The overlap between semantically unrelated states was 25 bits, and between semantically related states 48 bits. The twenty attractors were grouped into four classes of five patterns. We refer to two patterns in the same class as *neighbors*.

A sequence of three patterns representing the context, prime and target words was presented to the network via external input to the state units. Noise was added by randomly removing inputs with probability 0.03. After each pattern was processed, attractors were strengthened according to the rule above. The network was allowed to settle to an attractor for each input, except in the unconscious prime condition, in which the input was removed after 5 cycles, and with this short stimulus exposure the network always settled back to the rest state.

Table 1 shows the average results obtained from several repetitions of the simulation using different sets of random attractor states. The results qualitatively matched the experimental results: in the conscious condition, statistically significant facilitation was observed only for congruent contexts, whereas in the unconscious condition, significant facilitation was observed for both congruent and incongruent contexts. The reason for this is as follows: When the prime is processed in the conscious condition, the attractor net settles to the attractor representing the meaning of the prime that is related to the context (because the context attractor and related attractors have been strengthened by presentation of the context). This attractor is strengthened much more than the attractor representing the other meaning of the prime (due to the normalization of eligibility), resulting in facilitation of only congruent targets. In contrast, in the unconscious condition, both meanings of the prime are transiently activated before the state vector settles back to the rest state; consequently, both attractors are strengthened, resulting in facilitation of congruent and incongruent targets.

Discussion. The Marcel experiment showed a correlation between the onset of awareness and selection of a word's meaning. We modeled this correlation, but our theory makes a stronger claim: selection is a necessary precursor to awareness. This is because the state stabilizes only after selection, and awareness requires stability.

One interesting prediction arises from our model. Although both meanings of the prime were facilitated equally in the unconscious condition, this is not absolutely necessary. If the context could be made to strengthen its attractor to a greater degree, the processing of the prime could be biased to the extent that only one meaning would be significantly activated. Consequently, one might observe facilitation of only one meaning even in the unconscious condition. This leads to the as-yet-untested prediction that as context is "strengthened", the amount of priming of noncontextual meanings should decrease.

Metacontrast Masking and Response Time

Alpern (1953) described a metacontrast masking effect in which a square of light is flashed for 50 ms. After a variable stimulus onset asynchrony (SOA), a pair of *flanker* squares appears on either side of the *center* square, also for 50 ms. As the SOA is increased from zero, subjects first report a solid bar of light at short SOAs (~10 ms), the flankers alone at intermediate SOAs (~75 ms), and then the center square *followed* by the flankers at long SOAs (~120 ms). Thus, subjects were unaware of the center square at intermediate SOAs. Fehrer and Raab (1962) performed a variant of this experiment in which subjects were instructed to respond as soon as *any* stimulus was detected. Subjects took the same amount of time to respond at every SOA (~160 ms). There was a dissociation between awareness and reaction time, because if subjects were responding to their *percept*, then at intermediate SOAs they should be substantially slowed since they do not perceive the center square in that case[1].

Simulation. To model these data, we use a simple two-module architecture that captures the essence of the task. The *perceptual module* takes visual patterns as input and produces objects as output. This module has two input units, one for the center square and one for the flanker stimuli. There are three output units, one for each of three competing percepts—the center square alone, the flankers alone, and the center square and flankers together (a solid bar)—and three corresponding RBF units. The output units of the perceptual module feed into the inputs of a *response module* in a 1-1 manner, and this module makes a presence/absence decision about the input. Its mapping network simply sums the activity of the input units and passes this to a single output unit whose well-formed states are 0 and 1.

We do not model the low-level mechanism that give rise

1. Kolb & Braun (1995) provide a recent example of a similar dissociation between awareness and ability to respond.

to the masking effect[2]. We simply assume that this mechanism operates inside the mapping network of the perceptual module, and has the effect of suppressing activity stemming from the center square.

We used 5 cycles as the short SOA, 35 as the intermediate SOA, and 60 as the long SOA. The two input units were activated and deactivated appropriately to match the experimental conditions. The degree of stability of the output of the perceptual module was used as a measure of awareness, and the response module's settling time represented the RT for the detection response. The results closely matched those of the experiment: at the short SOA, the perceptual module stabilizes on the bar pattern; this occurs because the center and flanker patterns overlap substantially in time, providing enough activity to the combined "bar" pattern to win the competition. At the intermediate SOA, the perceptual module stabilizes on only the flanker pattern; in this case, the input stimuli do not overlap temporally, but the center square is removed before the perceptual module has time to stabilize on it. At long SOAs, the perceptual module stabilizes first on the center pattern, and then on the flanker pattern; in this case, the perceptual module has time to stabilize before the flanker pattern appears. The detection RTs for the three SOAs were not substantially different: 54, 51, and 51 cycles. The detection response is triggered by the initial flow of activity from the presentation of the center square. The dissociation between awareness and detection RT comes about because the two processes are based on different information—detection on activity and awareness on stability of activity.

Discussion. In the detection task, responses can be accurately initiated based on very coarse information. But if the task required finer detail, e.g., discriminating a center square from a center circle, the evidence that accumulated before the flankers masked the center stimulus might not be sufficient to form a response. In earlier simulations, we found that stability was required to initiate discrimination responses (Mathis and Mozer, 1995). Thus, the dissociation between awareness of a stimulus and ability to respond might not be absolute; it might depend on the response task. This leads to the prediction that the present dissociation should weaken or disappear for more complex discrimination tasks.

The Effect of Conscious States on Cognition

Philosophers raise the issue of whether consciousness is just a "read out" process or whether conscious states affect subsequent processing (Flanagan, 1992)[3]. There is no strong experimental test of this hypothesis yet, but an experiment providing some evidence was conducted by Levy and Pashler (1996), who examined the effects of speeded perceptual decisions on subsequent perceptual processing. Subjects viewed a visually degraded word stimulus, and made either a single unspeeded identification response (verbally reporting the word) or two responses: a speeded response in a 600–900 ms window, followed by an unspeeded response. Levy and Pashler found that in both the single- and dual-response conditions, the probability of a correct unspeeded response was .97. But in the dual-response condition this probability depended on the accuracy of the preceding speeded response. The probability of a correct unspeeded response given a *correct* speeded response was .99. But the probability of a correct unspeeded response given an *incorrect* speeded response was .93. One possible explanation for this is that the process of *making the speeded decision* affects subsequent processing, increasing performance when the speeded decision is correct, and hurting it when incorrect.

Simulation. The simulation consisted of the attractor component of a word-recognition module, as described earlier. Since identifying a visual word requires a fine discrimination, the perceptual module must settle to an attractor before a response can be made. Speeded responses were modeled by forcing the attractor net to settle to the nearest attractor state[4] after a certain number of cycles, chosen such that the probability of a correct speeded response matched that of the experiment (0.61). In the single-response condition, the net was allowed to settle to an attractor, and percent correct was recorded. In the dual-response condition, the net was forced to settle after 20 cycles, and was then allowed to resume processing from that state until settled. The results qualitatively match the data; see Table 2. In the simulation column of the table, all differences were statistically significant except that

Table 2: Simulation of Levy & Pashler (1996)

	human data	simulation
Prob(unspeeded correct | dual response)	.967	.962
Prob(unspeeded correct | single response)	.970	.965
Prob(unspeeded correct | speeded correct)	.992	1.000
Prob(unspeeded correct | speeded incorrect)	.925	.901

2. There is still no agreement as to an adequate model of metacontrast masking (e.g., see di Lollo et. al., 1993).
3. A stronger version of the issue asks whether qualia can affect cognition, but we address the materialist version.
4. Computationally, there are several ways this could be done, for example, setting the β parameter to a value close to zero.

between rows one and two. The basic pattern of results is due to a kind of perseverative effect in the network. Forcing the net to settle early may either help or hinder the network. It helps if the network passes through the neighborhood of the correct attractor, and is forced to settle there early, instead of possibly continuing on to an erroneous attractor (due to the noisy input and brief exposure). It hurts if the state vector passes nearby incorrect states on the way to the correct state, and is forced to settle to an incorrect state.

Discussion. In our model there is both a benefit in accuracy and a cost in accuracy of speeded settling, but in this experiment the benefit and cost balance each other. However, this balance may not hold for other measures, such as reaction time, which our model is well-suited to generate. This experiment provides support for our model, but it is consistent with other models as well. It is possible, for example, that the correlation between speeded and unspeeded response accuracy is due to a third factor that both accuracies depend on, and that there is no causal relationship between the two responses. It would be a challenge to experimentalists to devise a way to test the causality issue.

Conclusion

Our aim in this work is to account qualitatively for broad variety of data, rather than to model the full depth of detail in any given experiment. In doing so we hope to contribute ideas that may help constrain the development of more detailed models.

The particular choice of modules used in our simulations and the type of information processed in them depends on the particular domain being modeled. However, our theory proposes that there is one common set of modules that underlie all cognitive behavior, and that each module takes part in many cognitive tasks. It is the subject of future work to attempt to better identify the set of modules, and their specific location and connectivity in the brain.

References

Alpern, M. (1953). Metacontrast. *Journal of the Optical Society of America*, 43, 648-657

Becker, S., Behrmann, M., & Moscovitch, K. (1993), Word priming in attractor networks. *Proceedings of the cognitive science society*, p.231-236.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the neurosciences*, 2, 263-275

di Lollo, V., Bischof, W., & Dixon, P. (1993). Stimulus-onset asynchrony is not necessary for motion perception or metacontrast masking. *Psychological Science*, 4, 260-263.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1, 1-47

Fehrer, E., & Raab, D. (1962). Reaction time to stimuli masked by metacontrast. *Journal of experimental psychology*, 63, 143-147

Flanagan, O. J. (1992). *Consciousness reconsidered.* Cambridge, Mass. MIT Press.

Greenwald, A. G., Klinger, M. R., & Schuh, E. S. (1995). Activation by marginally perceptible ("subliminal") stimuli: Dissociations of unconscious from conscious cognition. *Journal of experimental psychology: General*, 124, 22-42

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98(1), 74-95

Hintzman, D. L. (1990). Human learning and memory: Connections and dissociations. *Annual Review of Psychology*, 41, 109-139

Kanizsa, G. (1979). *Organization in vision.* New York: Praeger.

Kolb, F. C., & Braun, J. (1995) Blindsight in normal observers. *Nature*, 377, 336-338

Kolers, P. A., & Von Grunau, M. V. (1976). Shape and color in apparent motion. *Vision Research*, 16, 329-335

Levy, J., & Pashler, H. (1996). Does perceptual analysis continue during selection and production of a speeded response? *Acta Psychologia*

Marcel, T. (1980). Conscious and preconscious recognition of polysemous words: Locating the selective effects of prior verbal context. In R.S. Nickerson (Ed.), *Attention and Performance VIII*, Hillsdale, NJ. Erlbaum.

Mathis, D. W., & Mozer, M. C. (1995). On the computational utility of consciousness. In Tesauro, G., Touretzky, D. S., & Leen, T. K. (Eds.) *Advances in neural information processing systems 7*, 11-18

Nisbett, R.E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231-259

Shallice, T. (1988). *From neuropsychology to mental structure.* New York: Cambridge U. Press.

Warren, R. M. (1970). Perceptual restorations of missing speech sounds. Science, 167, 392-393

Young, A.W., & DeHaan, E. H. F. (1990). Impairments of visual awareness. *Mind and language*, 5, 29-48