To appear in *Advances in Experimental Social Psychology*

Reflection and Reflexion: A Social Cognitive Neuroscience Approach to Attributional Inference[1]

Matthew D. Lieberman
University of California, Los Angeles

Ruth Gaunt
Bar-Ilan University

Daniel T. Gilbert
Harvard University

Yaacov Trope
New York University

> "Knowledge may give weight, but accomplishments give lustre, and many more people see than weigh."
> Lord Chesterfield, *Letters*, May 8, 1750

Lord Chesterfield gave his son, Philip, a great deal of advice—most of it having to do with manipulating other people to one's own ends—and that advice has survived for nearly three centuries because it is at once cynical, distasteful, and generally correct. One of the many things that Lord Chesterfield understood about people is that they form impressions of others based on what they see and what they think, and that under many circumstances, the former tends to outweigh the latter simply because seeing is so much easier than thinking. The first generation of social psychologists recognized this too. Solomon Asch observed that "impressions form with remarkable rapidity and great ease" (1946, p. 258), Gustav Ichheiser suggested that "conscious interpretations operate on the basis of an image of personality which was already performed by the unconscious mechanisms" (1949, p. 19), and Fritz Heider noted that "these conclusions become the recorded reality for us, so much so that most typically they are not experienced as interpretations at all" (1958, p. 82). These observations foretold a central assumption of modern dual-process models of attribution (Trope, 1986; Gilbert, Pelham, & Krull, 1988), namely, that people's inferences about the enduring characteristics of others are produced by the complex interaction of automatic and controlled psychological processes. Whereas the first generation of attribution models described the logic by which such inferences are made (Jones & Davis, 1965; Kelley, 1967), dual-process models describe the sequence and operating characteristics of the mental processes that produce those inferences. These models have proved capable of explaining old findings and predicting new phenomena, and as such, have been the standard bearers of attribution theory for nearly fifteen years.

Dual-process models were part of social psychology's response to the cognitive revolution. But revolutions come and go, and while the dust from the cognitive revolution has long since settled, another revolution appears now to be underway. In the last decade, emerging technologies have allowed us to begin to peer deep into the living brain, thus providing us with a unique opportunity to tie phenomenology and cognitive process to its neural substrates. In this chapter, we will try to make use of this opportunity by taking a "social cognitive neuroscience approach" to attribution theory (Adolphs, 1999; Klein & Kihlstrom, 1998; Lieberman, 2000; Ochsner & Lieberman, 2001). We begin by briefly sketching the major dual-process models of attribution and pointing out some of their points of convergence and some of their limitations. We will then describe a new model that focuses on the phenomenological, cognitive, and neural processes of attribution by defining the structure and functions of two systems, which we call the reflexive system (or X-system) and the reflective system (or C-system).

---

[1]This chapter was supported by grants from the National Science Foundation (BCS-0074562) and the James S. McDonnell Foundation (JSMF 99-25 CN-QUA.05). We gratefully acknowledge Kevin Kim for technical assistance and Naomi Eisenberger for helpful comments on previous drafts. Correspondence concerning this chapter should be addressed to Matthew Lieberman, Department of Psychology, University of California, Los Angeles, CA 90095-1563; email: lieber@ucla.edu.

I. Attribution Theory

A. The Correspondence Bias

In ordinary parlance, "attribution" simply means locating or naming a cause. In social psychology, the word is used more specifically to describe the process by which ordinary people figure out the causes of other people's behaviors. Attribution theories suggest that people think of behavior as the joint product of an actor's enduring predispositions and the temporary situational context in which the action unfolds (Behavior = Disposition + Situation), and thus, if an observer wishes to use an actor's behavior ("The clerk smiled") to determine the actor's disposition ("But is he really a friendly person?"), the observer must use information about the situation to solve the equation for disposition (D = B – S). In other words, people assume that an actor's behavior corresponds to his or her disposition unless it can be accounted for by some aspect of the situational context in which it happens. If the situation somehow provoked, demanded, aided, or abetted the behavior, then the behavior may say little or nothing about the unique and enduring qualities of the person who performed it ("Clerks are paid to smile at customers").

The logic is impeccable, but as early as 1943, Gustav Ichheiser noted that people often do not follow it:

> "Instead of saying, for instance, the individual X acted (or did not act) in a certain way because he was (or was not) in a certain situation, we are prone to believe that he behaved (or did not behave) in a certain way because he possessed (or did not possess) certain specific personal qualities" (p. 152).

Ichheiser (1949, p. 47) argued that people display a "tendency to interpret and evaluate the behavior of other people in terms of specific personality characteristics rather than in terms of the specific social situations in which those people are placed." As Lord Chesterfield knew, people attribute failure to laziness and stupidity, success to persistence and cunning, and generally neglect the fact that these outcomes are often engineered by tricks of fortune and accidents of fate. "The persisting pattern which permeates everyday life of interpreting individual behavior in light of personal factors (traits) rather than in the light of situational factors must be considered one of the fundamental sources of misunderstanding personality in our time" (Ichheiser, 1943, p. 152). Heider (1958) made the same point when he argued that people ignore situational demands because "behavior in particular has such salient properties it tends to engulf the total field" (p. 54).

Jones and Harris (1967) provided the first empirical demonstration of this *correspondence bias* (Gilbert & Malone, 1995) or *fundamental attribution error* (Ross, 1977). In one of their experiments, participants were asked to read a political editorial and estimate the writer's true attitude toward the issue. Some participants were told that the writer had freely chosen to defend a particular position and others were told that the writer had been required to defend that particular position by an authority figure. Not surprisingly, participants concluded that unconstrained writers held attitudes corresponding to the positions they espoused. Surprisingly, however, participants drew the same conclusion (albeit more weakly) about constrained writers. In other words, participants did not give sufficient consideration to the fact that the writer's situation provided a complete explanation for the position the writer espoused and that no dispositional inference was therefore warranted (if B = S, then D = 0).

B. Dual-Process Theories

The correspondence bias proved both important and robust, and over the next few decades social psychologists offered a variety of explanations for it, mostly having to do with the relative salience of behaviors and situations (see Gilbert & Malone, 1995; Gilbert, 1998a, 1998b). The cognitive revolution brought a new class of explanations that capitalized on the developing distinction between automatic and controlled processes. These explanations argued that the interaction of such processes could explain why people err on the side of dispositions so frequentlyas well as why they sometimes err on the side of situations when solving the attributional equation. They specified when each type of error should occur and the circumstances that should exacerbate or ameliorate either.

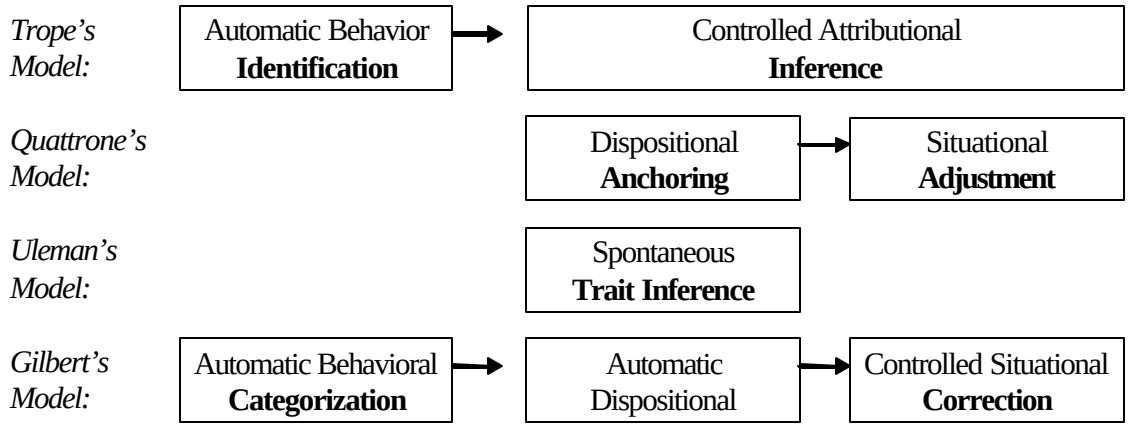
The Identification-Inference model. Trope's (1986) identification-inference model of attribution distinguished between two processing stages. The first, called identification, represents the available information about the person, situation, and behavior in attribution-relevant categories (e.g., anxious person, scary situations, fearful behavior). These representations implicitly influence each other through assimilative processes in producing the final identifications. The influence on any given representation on the process of identification is directly proportional to the ambiguity of the person, behavior, or situation being identified. Personal and situational information influences the identification of ambiguous behavior, and behavioral information influences the identification of personal and

situational information. The second, more controllable process, called inference, evaluates explanations of the identified behavior in terms of dispositional causes ("Bill reacted anxiously because he is an anxious person") or situational causes (e.g., "Bill reacted anxiously scary situations cause people to behave anxiously"). Depending on the availability of cognitive and motivational resources, the evaluation is systematic or heuristic. Systematic (diagnostic) evaluation compares the consistency of the behavior with a favored hypothetical cause to the consistency of the behavior with other possible causes, whereas heuristic (pseudodiagnostic) evaluation is based on the consistency of the behavior with the favored hypothetical cause and disregards alternative causes (Trope & Lieberman, 1993).

Assimilative identifications and heuristic inferences may produce overconfident attributions of behavior to any cause, dispositional or situational, on which the evaluation focuses. At the identification stage, dispositional or situational information produces assimilative influences on the identification of ambiguous behavior. For example, in a sad situation (e.g., funeral), a neutral facial expression is likely to be identified as sad rather than neutral. At the inference stage, the disambiguated behavior is used as evidence for the favored hypothetical cause. Specifically, a pseudodiagnostic inference process is likely to attribute the disambiguated behavior to the favored cause because the consistency of the behavior with alternative causes is disregarded. In our example, the neutral expression is likely to be attributed to dispositional sadness when a dispositional cause is tested for, but the same expression is likely to be attributed to the funeral when a situational explanation is tested for. In general, assimilative identification and heuristic inferences produce overattribution of behavior to a dispositional cause (a correspondence bias) when a dispositional cause is focal and overattribution of behavior to a situational cause when a situational cause is focal.

The Characterization-Correction Model. In the early 1980s, two findings set the stage for a second dual-process model of attributional inference. First, Uleman and his colleagues performed a series of studies that suggested that when people read about a person's behavior ("The plumber slipped $50 into his wife's purse"), they often spontaneously generate the names of the traits ("generous") that those behaviors imply (Winter & Uleman, 1984; Winter, Uleman, & Cuniff, 1985; see Uleman, Newman, & Moskowitz, 1996, for a review). Second, Quattrone (1982) applied Tversky and Kahneman's (1974) notion of anchoring and adjustment to the problem of the correspondence bias by suggesting that people often begin the attributional task by drawing dispositional inferences about the actor ("Let me start by assuming that the plumber is a generous fellow") and then adjust these "working hypotheses" with information about situational constraints ("Of course, he may feel guilty about having an affair, so perhaps he's not so generous after all"). Tversky and Kahneman had shown that in a variety of instances, adjustments of this sort are incomplete. As such, using this method of solving the attributional equation should lead people to display the correspondence bias. Quattrone's studies provided conceptual support for this hypothesis.

As Figure 1 shows, Gilbert et al (1988) incorporated these insights into a single



characterization-correction model, which suggested that (a) the second stage in Trope's model could be decomposed into the two sub-stages that Quattrone had described; and (b) that the first of these sub-stages was more automatic than the second. According to the model, people automatically identify actions, automatically draw dispositional inferences from those actions, and then consciously correct these inferences with information about situational constraints. Gilbert called these stages *categorization*, *characterization*, and *correction*. The key insight of the model was that because correction was the final and most fragile of these three sequential operations, it was the operation most likely to fail when people were unable or unwilling to devote attention to the attributional task. The model predicted that when people were under cognitive load, the correspondence bias would be exacerbated, and subsequent research confirmed this novel prediction (Gilbert, Pelham & Krull, 1988; Gilbert, Krull, & Pelham, 1988).

C. Reflection and Reflexion

Dual-process models make two assumptions about automaticity and control. First, they assume that automatic and controlled processes represent the endpoints on a smooth continuum of psychological processes, and that each can be defined with reference to the other. Fully controlled processes are effortful, intentional, flexible, and conscious, and fully automatic processes are those that lack most or all of these attributes. Second, dual-process models assume that only controlled processes require conscious attention, and thus, when conscious attention is usurped by other mental operations, only controlled processes fail. This suggests that the robustness of a process in the face of cognitive load can define its location on the automatic-controlled continuum. These assumptions are derived from the classic cognitive theories of Kahneman (1973), Posner and Snyder (1975), and Schneider and Shiffrin (1977), and are severely outdated (Bargh, 1989). In the following section we will offer a distinction between reflexive and reflective processes that we hope will replace the shopworn concepts of automaticity and control that are so integral to dual-process models of attribution.

To do so, we will describe the phenomenological features, cognitive operations, and neural substrates of two systems that we call the X-system (for the X in *reflexive*) and the C-system (for the C in *reflective*). These systems are instantiated in different parts of the brain, carry out different kinds of inferential operations, and are associated with different experiences. The X-system is a parallel-processing, sub-symbolic, pattern-matching system that produces the continuous stream of consciousness that each of us experiences as "the world out there." The C-system is a serial system that uses symbolic logic to produce the conscious thoughts that we experience as "reflections on" the stream of consciousness. While the X-system produces our ongoing experience of reality, the C-system reacts to the X-system. When problems arise in the X-system, the C-system attempts a remedy. We will argue that the interaction of these two systems can produce a wide variety of the phenomena that attribution models seek to explain.

II. The X-System

A. Phenomenology of the X-System

The inferences we draw about other people often do not feel like inferences at all. When we see sadness in a face or kindness in an act, we feel as though we are actually seeing these properties in the same way that we see the color of a fire hydrant or the motion of a bird. Inferences about states and traits often require so little deliberation and seem so thoroughly "given" that we are surprised when we find that others see things differently than we do. Our brains take in a steady stream of information through the senses, use our past experience and our current goals to make sense of that information, and provide us with a smooth and uninterrupted flow of experience that we call the stream of consciousness (Tzelgov, 1997). We do not ask for it, we do not control it, and sometimes we do not even notice it, but unless we are deep in a dreamless sleep, it is always there.

Traditionally, psychologists have thought of the processes that produce the stream of consciousness as inferential mechanisms whose products are delivered to consciousness but whose operations are themselves inscrutable. The processes that convert patterns of light into visual experience are excellent examples, and even the father of vision science, Herman von Helmholtz (1910/1925, p. 26-27), suggested that visual experience was the result of unconscious inferences that "are urged on our consciousness, so to speak, as if an external power had constrained us, over which our will has no control." By referring to these processes as inferential, Helmholtz seemed to be suggesting that the unconscious processes that produce visual experiences are structurally identical to the conscious processes that produce higher-order judgments, and that the two kinds of inferences were distinguished only by the availability of the inferential work to conscious inspection (unconscious inferences "never once can be elevated to the plane of conscious

judgments").

This remarkably modern view of automatic processes is parsimonious inasmuch as it allows sensation, perception, and judgment to be similarly construed. Moreover, it captures the phenomenology of *automatization*. For instance, when we learn a new skill, such as how to repair the toaster, our actions are highly controlled and we experience an internal monologue of logical propositions ("If I lift that metal thing, then the latch springs open"). As we repair the toaster more and more frequently, the monologue becomes less and less audible, until one day it is gone altogether and we find ourselves capable of repairing a toaster while thinking about something else entirely. The seamlessness of the phenomenological transition from ineptitude to proficiency suggests that the inferential processes that initially produced our actions have simply "gone underground," and that the internal monologue that initially guided our actions is still being narrated, but now is "out of earshot." When a process requiring propositional logic becomes automatized, we naturally assume that the same process is using the same logic, albeit somewhere down in the basement of our minds.

The idea that automatic processes are merely faster and quieter versions of controlled processes is theoretically parsimonious, intuitively compelling, and wrong. Even before Helmholtz, William James suggested that the "habit-worn paths in the brain" make such inaudible internal monologues "entirely superfluous" (1890, p. 112). Indeed, if the inferential process remained constant during the process of automatization, with the exception of processing efficiency and our awareness of its internal logic, we should expect the neural correlates of the process to remain relatively constant as well. Instead, it appears that there is very little overlap in the parts of the brain used in the automatic and controlled versions of cognitive processes (Cunningham, Johnson, Gatenby, Gore, & Banaji, 2001; Hariri, Bookheimer, & Mazziotta, 2000; Lieberman, Hariri, & Gilbert, 2001; Lieberman, Chang, Chiao, Bookheimer, & Knowlton, 2001; Ochsner; Bunge, Gross, & Gabrieli, 2001; Packard, Hirsh, & White, 1989; Poldrack & Gabrieli, 2001; Rauch et al., 1995). It is easy to see why psychologists since Helmholtz have erred in concluding that the automatic processes responsible for expert toaster repair are merely "silent versions" of the controlled process responsible for amateur toaster repair. From the observer's perspective the changes appear quantitative, rather than qualitative; speed is increased and errors are decreased. Parsimony would seem to demand that quantitative changes in output be explained by quantitative changes in the processing mechanism. Unlike the behavioral output, however, the changes in phenomenology and neural processing are qualitative shifts, and these are the clues that the behavioral output alone masks the underlying diversity of process.

B. Operating Characteristics of the X-System

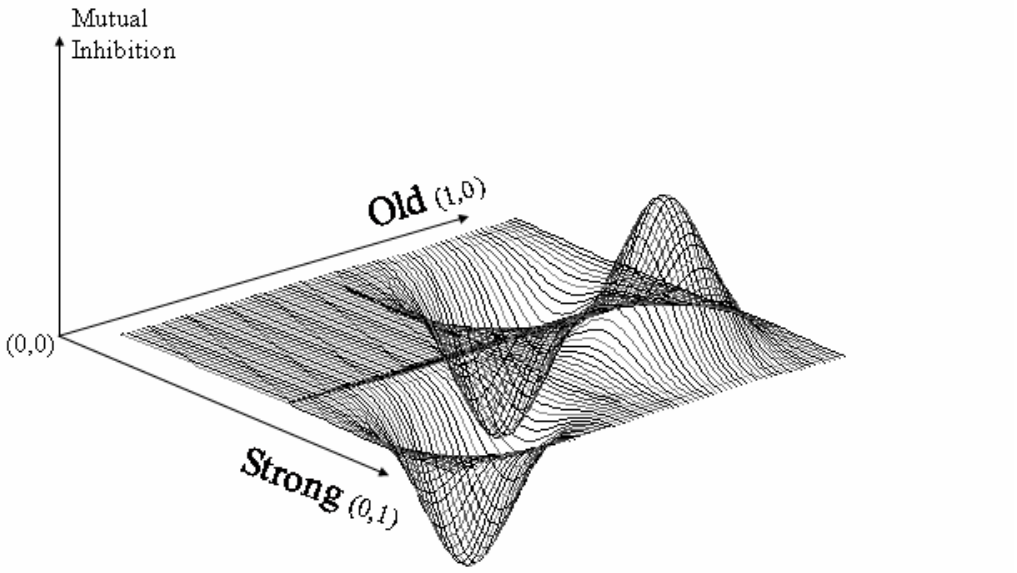
If automatic processes do not have the same structure as controlled processes, then what kind of structure do they have? The X-system is a set of neural mechanisms that are tuned by a person's past experience and current goals to create transformations in the stream of consciousness, and connectionist models (Rumelhart & McClelland, 1986; Smolensky, 1988) provide a powerful and biologically plausible way of thinking about how such systems operate (Smith, 1996; Read, Vanman, & Miller, 1997; Kunda & Thagard, 1996; Spellman & Holyoak, 1992). For our purposes, the key facts about connectionist models are that they are *sub-symbolic* and have *parallel processing* architectures. Parallel processing refers to the fact that many parts of a connectionist network can operate simultaneously rather than in sequence. Sub-symbolic means that no single unit in the processing network is a symbol for anything else—that is, no unit represents a thing or a concept, such as *democracy*, *triangle*, or *red*. Instead, representations are reflected in the pattern of activations across many units in the network, with similarity and category relationships between representations defined by the number of shared units. Being parallel and sub-symbolic, connectionist networks can mimic many aspects of effortful cognition without their processing limitations. These networks have drawbacks of their own, not the least of which is a tendency to produce the correspondence bias.

A Connectionist Primer. The complex computational details of connectionist models are described elsewhere (O'Reilly, Munakata, & McClelland, 2000; Rolls & Treves, 1998), and consequently, we will focus primarily on the emergent properties of connectionist networks and their consequences for attribution. The basic building blocks of connectionist models are *units*, *unit activity*, and *connection weight*. Units are the fundamental elements of which a connectionist network is composed, and a unit is merely any mechanism capable of transmitting a signal to a similar mechanism. Neurons are prototypical units. Unit activity corresponds to the activation level or firing rate of the unit that sends the signal, and connection

weight refers to the strength of the connection between two units. Connection weight determines the extent to which one unit's activity will result in a signal that increases or decreases another unit's activity. If the connection weight between units *a* and *b* is +1, then each unit's full activity will be added to the activity of the other, whereas a connection weight of –1 will lead to each unit's activity to be subtracted from the other unit's activity (within the limits of each unit's minimum and maximum firing rates). Positive and negative connection weights can thus be thought of as facilitating and inhibiting, respectively. Because inter-unit connections are bidirectional, units are simultaneously changing the activity of one another.

When two units have a negative connection weight, the units place competing constraints on the network. *Parallel constraint satisfaction* is the process whereby a connectionist network moves from an initial state of activity (e.g., the ambiguous text of a doctor's handwriting) to a final state that maximizes the number of constraints satisfied in the network and thus creates the most coherent interpretation of the input (e.g., a medical prescription). The process is parallel, because the bidirectional connections allow units to update one another simultaneously. The nonlinear processes of constraint satisfaction can be visualized if all the potential states of the network are graphed in N+1 dimensional space, with N being the number of units in the network and the extra dimension being used to plot the amount of mutual inhibition in the entire network given the set of activations for that coordinate (Hopfield, 1982, 1984).

To provide an oversimplified example (and violate the principle of sub-symbolic units), one can imagine a two unit network in which one unit represents the attribute *old* and the other unit represents the attribute *strong* (see Figure 2). Increasing the activation of one unit increases the strength of that feature in the overall pattern represented in the network. All possible combinations of activation strengths for the two units can be plotted in two dimensions that run from zero to one, representing a unit's minimum and maximum activation levels, respectively. The amount of mutual inhibition present at each coordinate may be plotted on the third dimension. As Figure 2 shows, *old* and *strong* are competing constraints within the network because they are negatively associated. When they are activated simultaneously, each unit inhibits the other according to its own level of activation and the negative connection weight linking them. When both units are activated, there is strong mutual inhibition, which is represented as a hill on the right side of the figure. The least mutual inhibition occurs when either one of the two units is activated alone. In this case, the active unit can fully inhibit the second unit without the second unit being able to reciprocate, because the negative connection weight only helps a unit inhibit another to the degree that it is active. When only a single unit is strongly activated, a valley is formed in the graph since there is no mutual inhibition.



The beauty of this "Hopfield net" illustration is that all of our instincts about gravity, momentum, and potential energy apply when we attempt to understand the way in which initial states will be transformed into final states. Imagine that the units for *old* and *strong* are simultaneously activated, with activations of 0.9 and 0.7, respectively. The network will initially have a great deal of mutual inhibition, but it will quickly minimize the mutual inhibition in the system through parallel constraint satisfaction. Because *old* is slightly more active than *strong*, *old* can inhibit *strong* more than *strong* can inhibit *old*. This will widen the gap in activation strengths between the two units, allowing *old* to have an even larger advantage in inhibiting *strong* after each round of updating their activations, until *old* and *strong* might have activations of .8 and .1, respectively. Following this path on the graph, it appears that the point representing the network's activity started on a hill and then rolled down the hill into the valley associated with *old*. Just as gravity moves objects to points of lower altitude and reduces the potential energy of the object, parallel constraint satisfaction reduces the tension in the network by moving from hills to valleys. Because each valley refers to a state of the network that conceptually "make sense" based on past learning, we refer to them as *valleys of coherence*. These valleys are also referred to as *local minima* or *attractor basins.*

Pattern Matching. The pattern of connection weights between its units may be thought of as its "implicit theory" about the input. Such theories develop as the network "observes" statistical covariations over time between features of the input. As the features of the input co-occur more frequently in the network's experience, the units whose pattern of activation corresponds to those features will have stronger positive connection weights (Hebb, 1949). As the pattern of connection weights strengthens, the network tends to "assume" the presence or absence of features predicted by the implicit theories of the network even when these features are not part of the input. In this sense, the strength of connection weights acts as a schema or a chronically accessibile construct (Higgins, 1987; Neisser, 1967).

For instance, if one end of a bicycle is partly hidden from the network's "sight," it will still be recognized as a bicycle because the network has a theory about what the object is likely to be, based on what it can "see" and what it has seen before. Units associated with the visible part of the bicycle will facilitate all of the units with which they are positively connected, including those typically, but not in this instance, activated by this occluded bicycle. The overall function of connectionist networks can thus be described as one of pattern matching (Smolensky, 1988; Sloman, 1996; Smith & DeCoster, 2000), which means matching imperfect or ambiguous input patterns to representations that are stored as a pattern of connection weights between units. This pattern matching constitutes a form of categorization in which valleys represent categories that are activated based on the degree of feature overlap with the input. In the example of *old* and *strong*, the initial activation (0.9, 0.7) is closer to, and therefore more similar to, the valley for *old* at (1.0, 0.0) than for *strong* at (0.0, 1.0). Thus, when the network sees a person who is objectively both *old* and *strong*, it is likely to categorize the person as old and weak. The network assimilates an instance (a strong, old person) to its general knowledge of the category to which that instance belongs (old people are generally not strong), and thus acts very much like a person who has a strong schema or stereotype.

Overall, the categorization processes of a network are driven by three principles that roughly correspond to *chronic accessibility*, *priming*, and *integrity of input*. Chronically accessible constructs represent categories of information that have been repeatedly activated together in the past. In connectionist terms, this reflects the increasing connection weights that constitute implicit theories about which features are likely to co-occur in a given stimulus. Priming refers to the temporary activation of units associated with a category or feature, and these units may be primed by a feature of the stimulus or by some entirely irrelevant prior event. Finally, the integrity of the input refers to the fact that weak, brief, or ambiguous inputs are more likely to be assimilated than are strong, constant, or unambiguous inputs.

Dispositional and Situational Inference in Connectionist Networks. When politicians are asked questions they cannot answer, they simply answer the questions they can. Connectionist networks do much the same thing. When a connectionist network is confronted with a causal inference problem, for example, it simply estimates the similarity or associative strength between the antecedent and the consequent, which sometimes leads it to make the error of affirming the consequent. Given the arguments "If *p* then *q*" and "*q*" it is illogical to conclude "*p*." Although it is true that "If a man is hostile, he is more likely to be in a fistfight," it is incorrect to infer from the presence of a fistfight that the man involved is hostile. Solving these arguments properly requires the capacity to appreciate unidirectional causality. The bidirectional flow of activity in the units of connectionist networks are prepared to represent associative strength rather than causality and thus are prone to make this inferential

error (Sloman, 1994; Smith, Patalano, & Jonides, 1998). For example, the "Linda problem" (Donovan & Epstein, 1997; Tversky & Kahneman, 1983) describes a woman in a way that is highly consistent with the category *feminist* without actually indicating that she is one. Participants are then asked whether it is more likely that Linda is (a) a bank teller or (b) a bank teller and a feminist. The correct answer is *a*, but the vast majority of participants choose *b*, and feel that their answer is correct even when the logic of conjunction is explained to them. Although one would expect a system that uses symbolic logic to answer *a*, one would expect a connectionist network to answer the question by estimating the feature overlap between each answer and the description of Linda. And in fact, *b* has more feature overlap than *a*.

Why does this matter for problems in attribution? As Trope and Lieberman (1993) suggest, overattribution of behavior to a dispositional or situational cause may be thought of as a case of affirming the consequent. If one wishes to diagnose the dispositional hostility of a participant in a fistfight, one must estimate the likelihood of a fistfight given a hostile disposition, p(fight|disposition), and then subtract the likelihood that even a non-hostile person might be drawn into a fistfight given this particular situation, p(fight|situation). Unfortunately, the X-system is not designed to perform these computations. Instead, the X-system tries to combine all the perceived features of the situation, behavior, and person into a coherent representation. The X-system will activate the network units associated with the dispositional hypothesis (man, hostile), the situation (hostile), and the behavior (fighting). Because these representations have overlapping features, the network will come to rest in a valley of coherence for hostility and conclude that the person looks a lot like a hostile person.

There are three consequences of these processes worth highlighting. First, the process of asking about dispositions in the first place acts as a source of priming—it activates the network's dispositional category for hostile people, which activates those units that are shared by the observed behavior and the dispositional valley of coherence. Simply asking a dispositional question, then, increases the likelihood that a connectionist network will answer it in the affirmative. While this has been the normative question in attribution research and modal question for the typical Westerner, when individuals hold a question about the nature of the situation the X-system is biased towards affirming the situational query (Krull, 1993; Lieberman, Gilbert, & Jarcho, 2001). As in the case of dispositional inference, the overlapping features between the representation of the situation and the behavior would lead the network to conclude that the situation is hostile. Second, if a dispositional question is being evaluated, situations have precisely the opposite effect in the X-system than the logical implications of their causal powers dictate. The same situation that will mitigate a dispositional attribution when its causal powers are considered, will enhance dispositional attributions when its featural associates are activated in the X-system. For example, while ideally the X-system could represent "fighting but provoked", (B-S), it actually represents something closer to "fighting and provoked", (B+S). In a similar manner, if a situational question is being evaluated, information about personal dispositions will produce behavior identifications that enhance rather than attenuate attribution of the behavior to the situations. Third, we have not clearly distinguished between automatic behavior identification and automatic dispositional attribution. This was not accidental. The difference between these two kinds of representations is reflected in the sort of conditional logic that is absent from the X-system. Logically, we can agree that while the target is being hostile at this moment, he may not be a hostile person in general. The X-system learns featural regularities and consequently has no mechanism for distinguishing between "right now" and "in general." This sort of distinction is reserved for the C-system.

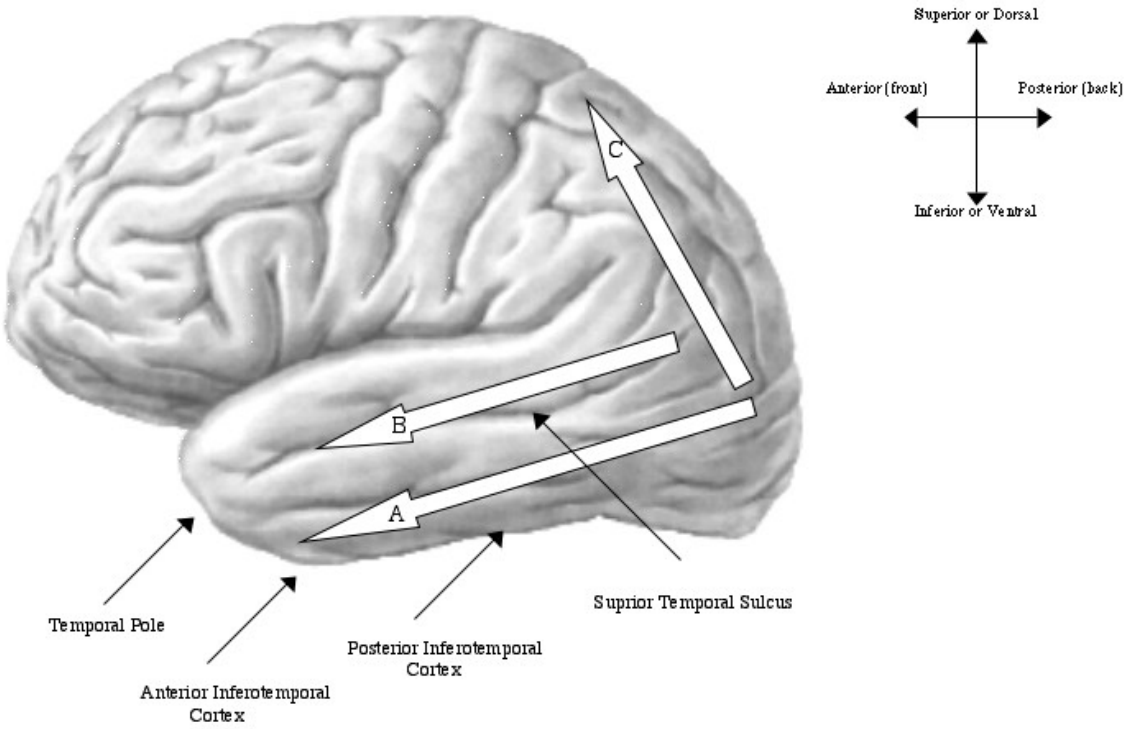
C. Neural Basis of the X-System

The X-system gives rise to the socially and affectively meaningful aspects of the stream of consciousness, allowing people to see hostility in behavior just as they see size, shape, and color in objects. The X-system's operations are automatic inasmuch as they require no conscious attention, but they are not merely fast and quiet versions of the logical operations that do. Rather, the X-system is a pattern-matching system whose connection weights are determined by experience and whose activation levels are determined by current goals and features of the stimulus input.

The most compelling evidence for the existence of such a system is not in phenomenology or design, but in neuroanatomy. The neuroanatomy of the X-system includes *lateral temporal cortex*, *amygdala* and *basal ganglia*. These are not the only regions of the brain involved in automatic processes, of course, but they are the regions most often identifiably involved in automatic social cognition. The amygdala and basal ganglia are responsible for spotting predictors of punishments and rewards, respectively (Adolphs, 1999; Knutson, Adams, Fong, & Hommer, 2001; LeDoux, 1996; Lieberman, 2000; Ochsner &

Schacter, 2000; Rolls, 1999). Although the basal ganglia and amygdala may be involved in automatically linking attributions to the overall valenced evaluation of a target (N. H. Anderson, 1974; Cheng, Saleem, & Tanaka, 1997), the lateral temporal cortex appears to be most directly involved in the construction of attributions. Consequently, this section will focus primarily on lateral (outer) temporal cortex, which is the part of temporal cortex visible to an observer who is viewing the side of a brain (see Figure 3, arrows a and b), in contrast to medial (middle) temporal areas that include the hippocampus and are closer to the center of the brain.

Figure 3: The X-system in Temporal Cortex



Inferotemporal Cortex and Automatic Categorization. Asch (1946), Brunswik (1947), and Heider (1958) suggested that social perception is analogous to object perception. Although this analogy has been occasionally misleading (Gilbert, 1998a), it has much to recommend it even at the neural level. Visual processing may be divided into two "information streams" that are often referred to as the "what pathway" and the "where pathway" (Mishkin, Ungerleider, & Macko, 1983). After passing through the thalamus, incoming visual information is relayed to occipital cortex at the back of the brain, where it undergoes these two kinds of processing. The "where" pathway follows a dorsal (higher) route to the parietal lobe (see Figure 3, arrow c), where the spatial location of an object is determined. The "what" pathway follows a ventral (lower) route through inferotemporal cortex or ITC (see Figure 3, arrow a), where the identity and category membership of the object is determined. This lower route corresponds more closely with the kind of perception relevant to attributional inference.

The ITC performs a pattern matching function. As information moves from the occipital lobe through the ventral pathway towards the temporal pole, a series of different computations are performed, each helping to transform the original input into progressively more abstract and socially meaningful categorizations. In the early stages along this route, neurons in the occipital lobe code for simple features such as line orientation, conjunction, and color. This information is then passed on to posterior ITC, which can represent complete objects in a view-dependent fashion. For instance, various neurons in posterior ITC respond to the presentation of a face, but each responds to a particular view of the face (Wang, Tanaka, & Tanifuji, 1996). Only when these view-dependent representations activate neurons in anterior (forward) ITC is view-invariance achieved. In anterior ITC, clusters of neurons respond equally well to most views of a particular object (Booth & Rolls, 1998) and consequently, this region represents entities abstractly, going beyond the strictly visible.

While visual information is flowing from the back of the brain towards anterior ITC, each area along this path is sending feedback information to each of the earlier processing areas (Suzuki, Saleem, & Tanaka, 2000), making the circuit fully bidirectional. This allows the implicit theories embedded in the more abstract categorizations of anterior ITC to bias the constraint satisfaction processes in earlier visual areas. Moreover, particular categories in anterior ITC may be primed via top-down activations from prefrontal cortex (Rolls, 1999; Tomita et al., 1999). Prefrontal cortex is part of the C-system (to be discussed shortly) that is involved in holding conscious thoughts in working memory. In the case of attribution, prefrontal cortex initially represents the attributional query ('Is he a hostile person?'), which can activate implied categories downstream in anterior ITC (person, hostility). In turn, then, anterior ITC can bias the interpretation of ambiguous visual inputs in posterior ITC and occipital cortex.

Neuroimaging studies in humans provide substantial evidence that anterior ITC is engaged in automatic semantic categorization (Boucart et al., 2000; Gerlach, Law, Gade, & Paulson, 2000; Hoffman & Haxby, 2000). Similar evidence is provided by single-cell recording with monkeys (Rolls, Judge, & Sanghera, 1977; Vogels, 1999) and lesion studies with human beings (Nakamura & Kubota, 1996; Weiskrantz & Saunders, 1984). Neurons in this area begin to process incoming data within 100 ms of a stimulus presentation (Rolls, Judge, & Sanghera, 1977) and can complete their computations within 150ms of stimulus presentation (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001; Thorpe, Fize, & Marlot, 1996). Traditionally, processes occurring in the first 350-500ms after a stimulus presentation are considered to be relatively uncontrollable (Bargh, 1999; Neely, 1991). A recent fMRI study of implicit prototype learning also favors an automaticity interpretation of anterior ITC activations (Aizenstein et al., 2000). Participants were trained to discriminate between patterns that were random deviations from two different prototypes (Posner & Keele, 1968), and though participants showed evidence of implicit category knowledge that correlated with neural activity in ITC, they had no conscious awareness of what they had been learning. In another neuroimaging study, participants were scanned while solving logic problems (Houde et al., 2000). When participants relied on more intuitive pattern-matching strategies, as evidenced by the systematic deviations from formal logic (Evans, 1989), activations were found in ventral and lateral temporal cortex. Finally, patients with semantic dementia, a disorder that damages the temporal poles and anterior ITC (Garrard & Hodges, 1999), show greater deficits in semantic priming tasks than they do in explicit tests of semantic memory (Tyler & Moss, 1998).

Neuroimaging studies also provide evidence that the X-system is a major contributor to the stream of consciousness. Portas, Strange, Friston, Dolan & Frith (2000) scanned participants while they viewed 3D stereograms in which objects suddenly appear to "pop-out" of the image when looked at the right way. ITC was one of the only areas of the brain whose activation was correlated with the moment of "pop-out" that is, the moment when the image emerges into the stream of consciousness. Sheinberg and Logothetis (1997) recorded activity from single cells in monkeys' cortices while different images were simultaneously presented to each of the animal's eyes creating "binocular rivalry". Although both images are processed up to a point in early parts of the visual processing stream, humans report seeing only one stimulus at a time. This suggests that early visual processing areas do not directly shape the stream of consciousness. Unlike early visual areas, ITC's activation tracked subjective experience rather than objective stimulus features. Finally, Bar et al. (2001) provided participants with masked 26ms presentations of several images. With multiple repetitions, participants were eventually able to identify the images, but the best predictor of whether an image would be consciously recognized on a particular trial was the degree to which ITC activations extended forwards towards the temporal pole.

In summary, ITC is responsible for categorical pattern matching. This pattern matching is automatic, relying on parallel processing along bidirectional links, and contributes directly to the stream of consciousness. It is also worth noting that the neurons in ITC appear to be genuinely sub-symbolic, which is necessary for their functions to be appropriately characterized by connectionist models. For instance, Vogels (1999) found that while the combined activations of a group of neurons in ITC accounted for the animals' behavioral categorization of stimuli into "trees" and "non-trees", no single neuron responded to all instances of trees and only to trees. Furthermore, the activity of individual neurons in ITC may not even correspond to smaller features that add up to the larger category. Multiple laboratories have reported being unable to discern any category-relevant features to which individual neurons respond (Vogels, 1999; Desimone et al., 1984; Mikami et al., 1994), suggesting that category activation is an emergent property of the ensemble of neurons. Indeed, it is more accurate to say that the ensembles of X-system neurons act "as if" they are representing a particular category than to say they really are. This

is similar to the way a calculator appears to behave "as if" it were representing mathematical equations (Searle, 1984). It is only to the outside observer that these "as if" representations appear genuine, but there is no evidence that any truly symbolic representations exist in calculators, the X-system, or any other machine. So far as we know, the reflective consciousness associated with the C-system is the only instance of real symbolic representation (Brentano, 1874; Husserl, 1913). It is not surprising that unlike the representations in the X-system, those in the C-system are not distributed over broad ensembles of neurons (O'Reilly, Braver, & Cohen, 1999).

Superior Temporal Sulcus and Behavior Identification. The analysis of the ventral temporal pathway contributes to our understanding of automatic attributional inference up to a point. The "what" pathway in ITC provides a coherent account of automatic category activation and its related semantic sequelae. This pathway performs a "quick and dirty" pattern-matching function that links instances in the world to previously learned categories. The semantic (anterior) and perceptual (posterior) ends of this pathway are bidirectionally linked, allowing activated categories in ITC to assimilate ambiguous perceptual targets. Up to this point, the analogy between social and object perception has been a useful guide, but like all analogies, this one is limited. In object perception, visual data are used to collect enough sensory data to know that a particular object is a shoe, a notebook, or an ice cream cone. Generally, people do not require that a shoe "do something" before they can determine what it is (cf. Dreyfus, 1991). Attribution, however, is generally concerned with how people use (or fail to use) the dynamic information contained in behavior to draw inferences about the *person and the situation*.

All attribution theories suggest that when behavior is unconstrained and intentional, it provides information about the actor's dispositions. For example, knowing whether Ben tripped Jacob "on purpose" or "by accident" is critical to understanding what kind of person Ben is and what can be expected of him in the future. Evolution seems to have picked up on this need to identify intentional behavior long before social psychologists realized its importance. There is mounting evidence that in addition to "what" and "where" pathways in the brain, a sizable strip of lateral temporal cortex constitutes what we might call the "behavior identification" pathway (Allison, Puce & McCarthy, 2000; Haxby, Hoffman, & Gobbin, 2000; Perrett et al., 1989). This pathway lies along the superior temporal suclus or STS (see Figure 3, arrow b) and is just above the "what" pathway. It receives combined inputs from the other two visual pathways, allowing for the conjunction of form and motion, and this conjunction results in an exquisite analysis of behavior.

For example, STS does not respond to random motion (Howard et al., 1996) or to unintentional behaviors, such as a person dropping an object (Perrett, Jellema, Frigerio, & Burt, 2001), but at least some neurons in STS respond to almost any action that could be described as intentional. Different neurons in STS are activated by eye gaze, head movement, facial expressions, lip movement, hand gestures, and directional walking (Decety & Grezes, 1999). Most remarkably, the same neuron tends to respond to entirely different behaviors as long as they are merely different ways of expressing the same intention! For instance, the neurons that respond to an actor facing an observer while staring straight ahead also respond when the actor is standing in profile with his eyes turned towards the observer (Allison, Puce & McCarthy, 2000; Perrett, Hietanen, Oram, & Benson, 1992). Although the visual information is radically different in these two instances, both represent the same intentional action, namely, "He's looking at me." Watching a person reach for an object in several different ways will activate the same neuron (Perrett et al., 1989), even when the only visual data are small points of light attached to the target's joints in an otherwise darkened room (Bonda, Petrides, Ostry, & Evans, 1996; Howard et al., 1996). It is difficult to make sense of these findings without concluding that the STS is identifying action based on the intention expressed.

Interestingly, most of the neurons in STS can be activated by disembodied eyes, hands, lips and bodies that contain no information about who it is that is behaving intentionally. These neurons seem to be activated by the pure intentionality of behavior rather than by the intentionality of particular individuals. This finding bears a striking similarity to earlier work on spontaneous trait inferences (Winter and Uleman, 1984; Moskowitz & Roman, 1992), which showed that trait terms are linked in memory with behaviors rather than individual actors. One study did present monkeys with recognizably different monkey targets to look at the interaction of actors and actions (Hasselmo, Rolls, & Baylis, 1989). In this study, three different monkey targets were each presented making three different facial expressions. Many neurons in ITC were activated whenever any monkey target was presented, and some ITC neurons were active only if a particular monkey was presented, suggesting that these neurons coded for the static identity of particular monkeys as well as the general category of monkeys. These neurons did not respond differentially to the three facial expressions. STS

neurons mostly showed the opposite pattern of activation, differentially responding to the distinct facial expressions but not to the identity of the monkeys. Some neurons in both ITC and STS, however, responded only to a particular monkey making a particular expression, suggesting that these neurons might be coding for the target's disposition as it corresponds to the facial expression. Although these neurons may have been coding a particular monkey's momentary emotional state rather than its disposition, it is important to remember that a connectionist network lacks the symbolic capacity to distinguish between 'now' and 'in general.'

Situations and Lateral Temporal Cortex. Our description of lateral temporal cortex has made explicit contact with attribution theory in terms of the representation of a target person, intentional behavior, and the person X behavior interaction. Conspicuously absent is any mention of situational representations in this guided tour of the X-system's neuroanatomy. While discoveries are changing our understanding of the brain almost daily, currently it is reasonable to say that lateral temporal cortex represents classes of objects and behavior. The representations of objects reflect their dispositional qualities insofar as the X-system is in the business of learning statistical regularities—which are a pretty good proxy for dispositions. Other sentient creatures are clearly the sort of objects the X-system is designed to learn about, but many objects represented by the X-system can be thought of as situations as well. A gun aimed at someone's head is clearly both an object and a situational context for the unfortunate individual at the end of its barrel. Similarly, an amusement park can be characterized in terms of a collection of visible features and is a situational context with general consequences for the behavior and emotional state of its visitors. Given that situations can be objects, there is no reason to think that the X-system cannot represent these situations in the same way that it can represent dispositions. It is even possible for an unseen situation to be activated in the X-system by behavior. Consider advertisements for horror films. All we need is the image of a terrified face and we are spontaneously drawn to thoughts of the terrifying situation that must have caused it.

Thus, the associated features of situations are represented along with the associated features of dispositions and behaviors in the X-system. It should be pointed out, however, that features that are statistically associated with a situation or disposition may be represented in the X-system, but as described earlier, their causal powers are purely the province of the C-system. For instance, funerals are statistically associated with the presence of tombstones, black clothes, and caskets. Funerals are also associated with particular behaviors (crying) and emotional states (sadness). The causal link between funerals and sadness need not be represented in order to learn their association. The activation of "funeral" in the X-system increases the likelihood that an ambigous facial expression will be resolved in favor of sadness, because sadness is primed by funeral and biases the network to resolve in a valley of coherence for sadness. Thus, the impact of situations and dispositions is strictly limited to priming their associates.

III. The C-System

George Miller (1981) observed that "the crowning intellectual accomplishment of the brain is the real world." The X-system is the part of the brain that automatically provides the stream of conscious experience that we take (or mistake!) for reality. The structure of behavior and the structure of the brain suggest that we share this system, and probably its capacities, with many other animals. As Nagel (1974) argued, there is something it is like to be a bat. But if we share with other animals the capacity for an ongoing stream of experience, it is unlikely that most also share our capacity to reflect on the contents of that stream.

Terms such as *reflective awareness* and *stream of consciousness* beg us to be confused, and thus it is worth pausing to consider them. Trying to define the stream of consciousness is a bit like fish trying to define water; it seems to be all encompassing and if it ever disappears no one will be around to say so. The stream of consciousness is the wallpaper of our minds; an ever-present backdrop for hanging the mental pictures that we focus on and it is usually only noticed if there is something very wrong with it. It spans our entire visual field and thus, phenomenologically, the objects in the stream are best thought of in terms of "consciousness *as* _____." That is to say there is no distinction between the stream and the objects in the stream; they are one and the same. There is never an empty part of the stream where there is just consciousness, but no object. *Reflective awareness*, on the other hand, is always "consciousness *of* _____" (Brentano, 1874; Husserl, 1913; Sartre, 1937). Any phenomenon or event in the stream of consciousness (a painting off to the side of one's desk) can be extracted from that stream, attended to ("that is an artprint"), reflected upon ("That's my Magritte. I haven't thought about that in ages."), integrated with other symbols ("Magrittes' don't belong on the same wall with fourteenth century Italian art"), and so on.

The relationship between reflective awareness and the stream of consciousness is roughly analogous to the relationship between figure and ground. Reflective awareness and the stream of consciousness refer to the two kinds of consciousness that give rise to these different kinds of percepts, figure and ground, respectively. The figure is not merely the information at the center of our visual field, rather it is that which emerges as a separate distinct entity from the background (Kohler, 1947). Through this emergence, we become conscious of this entity *as* an entity. The discovery that the object "is what it is" represents at once both the simplest form of reflective awareness and one of the most bewildering achievements of the human mind. Highly evolved mamalian brains are the only organization of matter in the known universe that can intrinsically represent phenomena. Stop signs, gas gages,and cloud formations do not intrinsically represent anything. The representation of an apple that emerges in reflective awareness is truly about something, even when that something is only an illusion (Aristotle, 1941).

When two people argue about whether dogs are conscious, the proponent is usually using that badly bruised term to mean stream of consciousness while the opponent is using it to mean reflective awareness. Both are probably right. Dogs probably do have an experience of yellow and sweet: There is something it is like to be a dog standing before a sweet, yellow thing, even if human beings can never know what that something is. But the experiencing dog is probably not able to reflect on that experience, thinking as it chews, "Damned fine ladyfinger, but what's next?" While the stream of consciousness and reflective awareness are easily confusable when it comes to the metaphysics of canine consciousness, it is worth noting that a wide array of human behaviors belie a sensitivity to the differences between the two. People drink, dance, and binge eat to stem the self-evaluative tide of reflective awareness, but none of these escape activities are aimed at switching off the stream of consciousness (Baumeister, 1990; Csikzentmihalyi, 1974; Heatherton & Baumeister, 1991; Steele & Josephs, 1989). People implicitly know that reflective awareness can be painfully oppressive in a way that the stream of consciousness cannot.

The system that allows us to have the thoughts that dogs cannot is the C-system, which may explain why dogs as a whole seem so much happier than human beings. The C-system is a symbolic processing system that produces reflective awareness, which is typically invoked when the X-system encounters problems it cannot solve—or more correctly, when it encounters inputs that do not allow it to settle into a stable state through parallel constraint satisfaction. When reflective consciousness is invoked, it can either generate solutions to the problems that are vexing the X-system, or it can bias the processing of the X-system in a variety of ways that we will describe. We begin by considering four phenomenological features of the C-system: *authorship, symbolic logic, capacity limits*, and that it is *alarm-driven*.

A. Phenomenology of the C-System

Authorship. By the age of three, most children can appreciate the difference between seeing and thinking, which allows them to distinguish between the products of their imaginations and the products of their senses (Johnson & Raye, 1981). One of the best indicators of a mental representation's origin is how it feels to produce it. Thinking usually feels volitional, controllable, and somewhat effortful, whereas seeing feels thoroughly involuntary, uncontrollable, and easy. We decide to think about something ("I've got to figure out how to get the stain out of my sweater"), and then we go about doing so ("Soap dissolves grease, but hot water dissolves soap, so maybe…"), but we rarely set aside time to do a bit of seeing. And when we do look at an object, we almost never find the task challenging. Seeing is just something that happens when our eyes are open, whether we like it or not.

The fact that we initiate and direct our thinking but not our seeing has two important and interrelated consequences. First, it suggests that our thoughts are more unique than our perceptions, and hence are more closely associated with our selves and our identities. Individuals pride themselves on their intelligence and creativity because they feel personally responsible for the distinctive paths their thinking takes, but they do not generally brag about being "the guy who is great at seeing blue" or instruct their children that "a lady must always do her best to tell horses from brussel sprouts." Second, because we have the sense of having generated our thoughts but not our perceptions, we tend to trust the latter in a way that we do not trust the former. The products of perception have a "given" quality that leads us to feel that we are in direct contact with reality. Thoughts are about things, but perceptions *are* things, which is why we say, "I am thinking about Katie" when Katie is absent, but not "I am having a perception about Katie" when Katie is standing before us. Our perceptions feel immediate and unmediated, our thoughts do not, and that is why it is generally easier to convince someone that they have reached the wrong conclusion ("Just because she's Jewish doesn't mean she's a Democrat") than that they have had the wrong perception ("That was a cow, not a

traffic light").

***Symbolic logic*.** If the crowning achievement of the X-system is the real world, then the crowning achievement of the C-system is symbolic logic. The ability to have a true thought about the world, and then produce a second true thought based on nothing more than its logical consistency with the first, allows every human mind to be its own truth factory. Symbolic logic allows us to escape the limits of empiricism and move beyond the mere representation and association of events in world and into the realms of the possible. The fact that we can execute endless strings of "if-then" statements means that we can consider the future before it happens and learn from mistakes we have never made ("If I keep teasing the dog, then he will bite me. Then I will bleed. Then mom will cry. So this is a really bad idea"). This capacity also ensures that the C-system, unlike the X-system, can represent unidirectional causal relations (Waldmann & Holyoak, 1992) and the causal powers of symbolic entities in general.

It is important to note that the products of the X-system can also be described as the result of executing a series of "if-then" statements, just as the mechanical connection between a typewriter's key and hammer can be described as a representation of the logical rule "If the fifth key in the middle row is depressed, then print the symbol G on the paper." But typewriters do not use symbolic logic anymore than planets use Keppler's equations to chart their courses through the heavens (Dennett, 1984), and so it is with the X-system. It was a flaw of early models of automatic cognition to suggest that symbolic logic was part of the mechanism of X-system processes (Newell, 1990). The C-system, on the other hand, truly *uses* symbolic logic—at least phenomenologically—which is why people who learn logical reasoning skills end up reasoning differently than people who do not (Nisbett, Krantz, Jepson, & Fong, 1982). Because symbolic logic is part of the "insider's" experience of the C-system, symbolic logic must be explained, rather than explained away, by any final accounting of the C-system.

***Capacity Limits*.** The maximum number of bytes of information that we can keep in mind at one time is approximately seven, plus or minus two (Baddeley, 1986; Miller, 1956). But the maximum number of thoughts we can think at once is approximately one, plus or minus zero (James, 1890). Indeed, even when we have the sense that we might be thinking two things at once, careful introspection usually reveals that we are either having a single thought about a category of things ("Phil and Dick sure do get along nicely") or that we are rapidly oscillating between two thoughts ("Phil is so happy…Dick is too…I think Phil is glad to have Dick around.") Indeed, it is difficult to know just what thinking two thoughts at the same time could mean. The fact that reflective thinking is limited to one object or category of objects at any given moment in time means that it must execute its symbolic operations serially rather than in parallel. The effortfulness and sequential nature of reflective thought makes it fragile: A person must be dead or in a dreamless sleep for the stream of consciousness to stop flowing, but even a small, momentary distraction can derail reflective thinking.

Alarm-driven. Wilshire (1982, p. 11) described an unusual play in which the first act consisted of nothing more than a kitchen sink and an apple set upon the stage:

> "The kitchen sink was a kitchen sink but it could not be used by anyone: the faucets were unconnected and its drainpipe terminated in the air. These things were useless. And yet they were meaningful in a much more vivid and complete way than they would be in ordinary use. Our very detachment from their everyday use threw their everyday connections and contexts of use into relief… The things were perceived *as* meaningful… That is, actual things in plain view—not things dressed up or illuminated to be what they are not—are nevertheless seen in an entirely new light."

As silly as this play may seem, it does succeed in transforming the overlooked into the looked over. Kitchen sinks are part of our ordinary stream of conscious experience, and yet, even as we use them, we rarely if ever reflect upon them. Absurdist art is meant to wake us up, to make us reflect on that which we normally take for granted, to become momentarily aware of that which would otherwise slip through the stream of consciousness without reflection. Alas, if there is one clear fact about reflective consciousness, it is that it comes and goes. Like the refrigerator light, reflective consciousness is always on when we check it; but like the refrigerator light, it is probably off more often than on (Schooler, in press). What switches reflective consciousness off and on? Whitehead (1911) argued that acts of reflection "are like cavalry charges in a battle—they are strictly limited in number, they require fresh horses, and must only be made at decisive moments." Normally, a cavalry's decisive moment comes when someone or something is in dire need of rescue, so what might reflective consciousness rescue us from?

Normally, reflective awareness is switched on by problems in the stream of consciousness. As Dewey

noted, reflection is initiated by "a state of perplexity, hesitation and doubt" which is followed by "an act of search or investigation" (1910, p. 10). Heidegger similarly suggested that this moment of doubt is what transforms cognition from "absorbed coping" to "deliberate coping" (Dreyfus, 1991; Heidegger, 1927), or in our terms, from experience to awareness of experience. The X-system's job is to turn information that emanates from the environment into our ongoing experience of that environment, and it does this by matching the incoming patterns of information to the patterns it stores as connection weights. When things match, the system settles into to a stable state and the stream of consciousness flows smoothly. When they do not match, the system keeps trying to find a stable state, until finally the cavalry must be called in. We will have much more to say about this in the next section, and for now we merely wish to note that part of the phenomenology of reflective consciousness is that we often come to it with a sense that something is awry, that an alarm has been sounded to grab our attention, and we use reflective consciousness to figure out what that something is and to fix it. There may be fish in the stream of consciousness, but when an elephant swims by we sit up and take notice.

B. Operating Characteristics of the C-System

If we could describe in detail the operating characteristics of the C-system, we would be collecting the Nobel Prize rather than sitting here typing. That description would be a conceptual blueprint for a machine that is capable of reflective awareness, and such a blueprint is at least a quantum leap beyond the grasp of today's science. In our discussion of the X-system, we suggested that its operations may be described in terms of symbolic logic, but that it actually functions as a connectionist network. The C-system, on the other hand, uses symbolic logic, and no one yet knows what kind of system can do that. A good deal is known about the necessary conditions for reflective awareness; it is probably necessary that the critter in question be a living primate with a functioning prefrontal cortex. As for the sufficient conditions, it is arguable that not a single positive fact has been generated going back to pre-socratics (Schrodinger, 1992); that is to say, we don't know *why* any of the necessary conditions are necessary. One thing we can say is that the C-system is probably not a connectionist network, and we know this because connectionist networks cannot be made to do the things that the C-system does (Fodor, 2000). Until science can provide a full account of the C-system's operating characteristics, including our experience of using symbolic logic, we must be satisfied to note the functional aspects of the system.

The Alarm System. As early as the 1950's, scientists showed that cybernetic systems—that is, systems that use their output (past behavior) as input (information) for more output (future behavior)—could be made to perform a variety of interesting tricks. Miller, Galanter, and Pribram (1960) showed that many complex, purposive behaviors could be produced by a system that simply computes the difference between its current state and its desired state, and then acts to reduce that difference. For example, a thermostat's "goal" is that the temperature in a room should be 72 degrees Fahrenheit, and when the temperature falls below that standard, an alarm signal triggers the thermostat to "wake up" and run the furnace. The thermostat keeps checking to see if it has reached its desired state, and when it does, it goes back to sleep. This apparently complex behavior requires only that the thermostat execute what Miller et al called a TOTE loop—that is, a series of operations that can be described as Testing ("Is it 72? No"), Operating ("Turn on the furnace"), Testing ("Is it 72? Yes"), and Exiting ("Goodnight!").

The thermostat is a good model of some aspects of the relationship between the X- and C-systems. In an old-fashioned (non-electronic) thermostat, temperature deviations are represented as changes in the height of the mercury in a thermometer, and if the mercury falls below the standard, a circuit is completed that activates the thermostat. The thermostat then operates until the mercury level rises to the standard and disconnects the circuit. The height of the mercury, then, constitutes a kind of alarm system. Similarly, the amount of sustained mutual inhibition in the X-system—which represents the degree to which constraint satisfaction processes have failed to match the information emanating from the environment to an existing pattern—can serve to switch on the C-system. To visualize this, simply recall the hills and valleys in Figure 2 and imagine a set of low-lying clouds hovering over this terrain. When the level of mutual inhibition in the X-system reaches the cloud layer, a circuit is completed and the C-system is brought on line. The beauty of this arrangement is that the C-system does not need to continuously monitor the X-system for problems; rather, the X-system automatically wakes the C-system up when problems arise. The C-system does not need to go looking for trouble. Trouble finds it.

People are, of course, more complex than thermostats, and the analogy breaks down at some points. Whereas a thermostat's standard is determined by the human being in whose home it is

installed, the X-system's standard is in part determined by the C-system. The goals and concerns that are represented in reflective consciousness serve to bias the alarm system's sensitivity, causing it to sound at higher or lower levels of mutual inhibition. In a sense, the C-system is like a person who sets an alarm clock: It does not need to be continuously or intermittently awake throughout the night in order to test the current time against the desired time of awakening. Rather, it simply sets the alarm and goes to sleep, thereby determining the conditions under which it will be woken without having to watch for them.

The C-system, then, is automatically brought on line by sufficiently vexing problems in the X-system, and the C-system helps determine what "sufficiently vexing" means. These facts have implications for attribution. For example, the characterization-correction model (Gilbert, 1989) suggests that cognitive load prevents people from using their knowledge of situational constraints to adjust their automatic dispositional inferences. The current analysis suggests that situational information may not be used under conditions of cognitive load because (a) cognitive load prevents reflective consciousness from carrying out the logical process of integrating *p*(B/S) with *p*(B/D), or (b) cognitive load resets the sensitivity of the alarm system, and thus the C-system is insensitive to the incoherence in the X-system. It is not clear whether load causes reflective consciousness to stumble in its attempts to correct the dispositional inference, or whether the person simply "sleeps right through" the problem.

The alarm system cannot easily be labeled automatic or controlled because it shares characteristics with both kinds of processes. On the one hand, the alarm is spontaneously triggered when a preset amount of mutual inhibition is present in the X-system. On the other hand, cognitive load may impair the sensitivity of the alarm system, a telltale sign of a controlled process. Rather than trying to resolve whether the alarm system is automatic or controlled, we suggest that the alarm system is an ideal example of why current views of automaticity and control have outlived their usefulness. We have, however, made the claim that the alarm system is part of the G-system, rather than the X-system. This decision reflects the fact that the activity system activity is almost perfectly correlated with the activity of the rest of the C-system.*(?)* This is as it should be. It would be very strange indeed if the detection of the need for control were not highly correlated with the actual exertion of control.

Correction. Our discussion of the mechanisms that activate reflective consciousness may seem to suggest that when the G-system wakes up, the X-system goes to sleep. This is not the case, of course, as it would leave us with a person who has reflective awareness without any experience of which to be aware. If the C-system is like a refrigerator light that goes on and off, the X-system is like the contents of the refrigerator: They are always there, and if they were gone, there would be nothing for the refrigerator light to illuminate.

The fact that the X-system is always on means that even when the C-system wakes up and attempts to use symbolic logic to solve problems that the pattern matching X-system has failed to solve, the X-system continues to match patterns and produce experience. The fact that both systems can be operating at the same time gives rise to some familiar dissociations between what we think and what we see. Consider the case of optical illusions. When we peer into an Ames room, we see a huge person in one corner and a tiny person in another, and when the two people change places, they seem to shrink and grow, respectively. A quick trip around the room with a measuring tape is enough to convince us that the people are actually the same size and the walls are trapezoidal, but even after we are so enlightened, when we look into the room again we see a giant and a midget. The problem is that while the C-system has used symbolic logic to understand the visual effects of trapezoidal walls, the X-system continues to compute height the old-fashioned way, and its products continue to shape the stream of conscious experience. Indeed, the only way to resolve the dissociative dilemma is to shut one's eyes, thereby depriving the X-system of the input that it cannot process correctly.

The same sort of dissociation can occur when we attempt to diagnose the states and traits of others. When we see a person fidgeting nervously in a chair, the X-system matches that pattern of behavior and leads us to "see" dispositional anxiety. The C-system may use symbolic logic to consider the causal implications of the situation—for example, that the person is waiting for the results of a medical test. But because both systems are on, and because the X-system does not stop producing experience even when the G-system knows that that experience is wrong, the observer can be left with the unusual feeling that the actor is dispositionally anxious even though the observer knows that the actor ought to be forgiven his twitching. When the X- and C-systems collide, the resolution is often a compromise. The X-system votes for dispositional anxiety, the C-system votes against it, and when asked, the observer says, "Well, he's probably just anxious about the test results, but still, he's fidgeting terribly, so…I don't know, I guess perhaps he's a slightly anxious

person." And indeed, this is precisely what real people tend to say when they observe others behaving in line with strong situational constraints.

Diagnosticity and Pseudodiagnosticity. The foregoing example assumes that the C-system is not only alerted that its vote is required, but also that the C-system is a conscientious citizen that carefully considers all the candidate theories before voting. A diagnostic evaluation of the target's disposition requires computing the likelihood that a behavior would occur if an actor has the disposition, p(B/D), and subtracting out the likelihood of the alternative theory that the situational constraints might cause anyone to behave that way, p(B/S). Frequently, the C-system is either too busy with other activities to vote or isn't concerned enough to educate itself thoroughly before casting a ballot. In both of these cases, when person is under cognitive load or is not motivated to be accurate, the original dispositional hypothesis will often be affirmed. This occurs even though the alarm system has just woken up the C-system to alert it to the conflict in the X-system. In these cases, the system engages in pseudodiagnostic processing of the evidence such that only the probability of the behavior occurring given the hypothesized disposition, p(B/D), is calculated in order to assess the actor's disposition (Trope & Liberman, 1993, 1996; Trope & Gaunt, 1999). Although this is not strictly a pattern-matching function, pseudodiagnostic processing will produce outcomes similar to those produced by the X-system, namely, affirming the consequent and generating a correspondence bias (Corneille, Leyens, Yzerbyt, & Walther, 1999, Tetlock, 1985; Webster, 1993).

Therefore, unwarranted attributions do not necessarily reflect the biased output of the X-system alone. There is plenty of blame to go around, and when lack of motivation or cognitive resources lead the C-system to perform a simple pseudodiagnostic evaluation, the C-system may produce faulty conclusions. In this case, the C-system may fail to use information about the alternative causes of behavior ("the person is waiting for the results of a medical test") and thus it may vote for dispositional anxiety even when it "knows" about an anxiety provoking situation.

Prior Beliefs. Formal logic may define the constellation of operations that can be used to compare, contrast, and construct new knowledge from existing symbols, but our knowledge of the world and our culturally-driven prior beliefs shape the rules we actually use. For behaviors, these beliefs specify how dispositions and situations interact to produce various behaviors (see e.g. Dweck, Hong, & Chiu, 1993; Morris & Larrick, 1995; Shoda & Mischel, 1993; Trope & Liberman, 1993).

For example, observers' prior beliefs vary systematically as a function of the behavioral domain they are contemplating. Some behaviors are conceived as primarily dependent on personal dispositions, whereas others are assumed to result mainly from situational inducements. Reeder and his colleagues found that people believe that excellent performances necessarily imply excellent ability, whereas poor performance may result either from poor ability or situational constraints (Reeder, 1985, 1993; Reeder & Brewer, 1979; Reeder & Fulks, 1980). Similarly, observers believe that moral actors are unlikely to commit immoral acts regardless of situational influences, whereas immoral actors may sometimes act morally in the presence of situational incentives (Reeder & Spores, 1983). Thus, observers are likely to draw confident dispositional attributions when they believe that the corresponding disposition is necessary for the occurrence of behavior and to attenuate their attributions when they believe that the behavior could also be produced by other factors (see Kelley, 1972; Trope, 1986; Trope & Liberman, 1993).

A demonstration of the important role played by prior beliefs was provided by a study that used the attitude attribution paradigm (Morris & Larrick, 1995) In this study, participants' beliefs about the relationships between the actor's behavior, the actor's personal attitude, and the situational incentives were measured before the inferential task. Participants who did not believe that the situational incentive was sufficient to generate counterattitudinal behavior inferred that the essay reflected the writer's true attitude, even when they were aware of the situational incentive. Only participants who believed the situational incentive was sufficient to cause counterattitudinal behavior used the situation to discount their attitude attribution. Thus, a prior belief that defines the alternative cause as sufficient for the occurrence of behavior is crucial for the discounting of an attribution (see also Bierbrauer, 1979; Sherman, 1980).

C. Neural Basis of the C-System

In our characterization of the X-system, we reviewed several neuroimaging studies that localized automatic pattern-matching and the emergence of these patterns into the stream of consciousness. Several of those studies also included other processing conditions that targeted the activity for which we believe the C-system is responsible. For instance, whereas implicit pattern learning was associated with anterior ITC activations, instructions to explicitly search for meaningful patterns did not activate anterior ITC.

Rather, they activated prefrontal cortex and hippocampus (Aizenstein et al., 2000). Houde et al., (2000) found that while similarity-based pattern-matching activated lateral temporal cortex, rule-based processing of the identical problems led to prefrontal, anterior cingulate and hippocampus activations. In the 3D stereogram pop-out study (Portas et al., 2000), the initial recognition of the hidden image was associated with ITC activation, whereas the ability to maintain focus on the percept was associated with prefrontal and hippocampal activations. Another fMRI study (Mummery, Shallice, & Price, 1999) examined the strategic and automatic components of semantic priming in a lexical decision task. Whereas the automatic components of semantic priming tended to correspond to anterior ITC activations, more strategic components were associated with prefrontal cortex and anterior cingulate. Additionally, most fMRI studies of symbolic logic have implicated prefrontal cortex (Goel & Dolan, 2000; Goel, Gold, Kapur, & Houle, 1997; Just, Carpenter, & Varma, 1999; Smith, Patalano, & Jonides, 1998; Waltz et al., 1999; Wharton & Grafman, 1998)

We propose that the C-system performs three inter-related operations: Identifying when problems arise in the X-system, taking control away from the X-system, and remembering situations in which such control was previously required. Based on a review of cognitive neuroscience research, we believe that these functions are served by the *anterior cingulate*, *prefrontal cortex*, and *hippocampus*, respectively. In the following sections, we will review evidence tying each of these neural structures to its function.

Anterior Cingulate, Pain, and Affect. The anterior cingulate is an area of the cortex that sits just above the corpus collosum on the medial (middle) wall of each hemisphere. Different parts of the anterior cingulate receive input from various neural structures including the amygdala, basal ganglia, lateral temporal cortex, hippocampus, prefrontal cortex, and regions associated with somatic and visceral sensations (Barbas, 2000; Rolls, 1999). A growing body of research is converging on the notion that the anterior cingulated is an alarm system (Bush, Luu, & Posner, 2000; Botvinick, Braver, Barch, Carter & Cohen, in press; Ochsner & Feldman-Barrett, in press).

For example, the most basic alarm signal is pain, which lets us know that we had best stop what we are doing and pay attention to the source of the pain before we get into serious trouble. The anterior cingulate is one of only two areas of the brain whose activation covaries with subjective reports of unpleasantness in response to both somatic and visceral pain (Baciu et al, 1999; Ladabaum, Minoshima, & Owyang, 2000; Rainville, Duncan, Price, Carrier, & Bushnell, 1997). Angina attacks are associated with activation of the anterior cingulate, but "silent" myocardial ischemia (which lacks the subjective component of angina) is not (Rosen et al, 1996). People who have had their anterior cingulate lesioned report that they feel the physical intensity of their pain, but that it no longer seems unpleasant and no longer concerns them (Foltz & White, 1968; Hurt & Ballantine, 1974). Interestingly, this separation of sensation from the emotional meaning of the sensation is analogous to depersonalization symptoms associated with schizophrenia and the use of certain hallucinogens, both of which are thought to involve alterations of anterior cingulate activity (Mathew et al, 1999; Sierra & Berrios, 1998).

Appraisal theories of emotion suggest that emotions are a good indicator of how a person is currently appraising the match between his or her goals and the current state of the world (Fridja, 1986; Lazarus, 1991; Ortony, Clore & Collins, 1988). The anterior cingulate is commonly activated by emotional stimuli, scripts, and internally generated memories (Dougherty et al., 1999; George et al., 1995; Mayberg et al., 1999; Shin et al., 2000). The anterior cingulates of women with young children are more reactive to infant cries than to a variety of other auditory stimuli (Lorberbaum et al., 1999). Both clinical and transient anxiety are correlated with anterior cingulate reactivity (Kimbrell et al., 1999; Osuch et al., 2000). Consistent with two-factor theories of emotion (James, 1894; Schacter & Singer, 1962; cf. Ellsworth, 1994), the anterior cingulate, along with the insula, is one of the major sites of autonomic feedback (Buchanan, Valentine & Powell, 1985; Critchley, Corfield, Chandler, Mathias, & Dolan, 2000; Soufer et al., 1998). Finally, awareness of one's emotional state correlates with anterior cingulate activity (Lane et al, 1998), and alexithymia (a disorder characterized by impaired identification of one's own emotional states) is associated with reduced cingulate activity in response to emotionally evocative stimuli (Berthoz et al., 2000).

Anterior Cingulate and Cognitive Errors. In addition to responding to painful and affectively significant stimuli, different components of the anterior cingulate respond to cognitive and perceptual tasks that evoke increased controlled processing (Bush, Luu, & Posner, 2000; Derbyshire, Vogt, & Jones, 1998). In the Stroop task, for instance, individuals are required to name the color in which a word is written (e.g., for the word "r-e-d" written in blue ink, the correct answer is "blue"). This task is difficult because people automatically read the word

and thus have a prepotent linguistic response that they find nearly impossible to ignore. Controlled processes identify this prepotent response as inaccurate, inhibit it, and generate the correct response. This process requires more time when the ink color and the word are incongruent, and indeed, there is more anterior cingulated activation on the trials with longer reaction times (MacDonald, Cohen, Stenger, & Carter, 2000). In addition, anterior cingulate lesions exacerbate the classic Stroop interference effect (Ochsner et al., 2001).

Though the activity of the anterior cingulate most often correlates with conscious concern or hesitation, anterior cingulate activation has also been observed in conflict situations for which there is no conscious awareness of the conflict. Berns, Cohen, and Mintun (1997) found anterior cingulate activations when a sequence that had been learned implicitly was altered so that its pattern no longer matched the previous presentations. The anterior cingulate is also activated when multistable visual illusions (such as the Necker cube) switch from one view to another (Lumer, Friston, & Rees, 1998). Although the switch itself is conscious, there is no sense of conscious effort associated with the resolution of the stimulus. These two studies suggest that the anterior cingulate's activation is responsive to the tension in the various networks that constitute the X-system and its perceptual precursors.

Anterior Cingulate as Alarm System. In most controlled processing tasks, detecting the need for control and exercising control are confounded because exercising control typically follows immediately on the heels of detection. Recent studies have shown that the anterior cingulate is responsible for the detection of conflict rather than the exercise of control (Carter et al., 1998, 2000) and thus is a good candidate for our previously described alarm system. In one study, participants were given two different blocks of Stroop trials. In one block, participants were given the accurate expectation that the trials would be mostly congruent ("R-E-D" in red ink), whereas in the other block they were given the accurate expectation that the trials would be mostly incongruent ("R-E-D" in blue ink). In each block, 80% of the trials matched the participant's expectation and 20% did not. Regardless of the participant's expectations for the block, trials that were incongruent required the same amount of control to be exercised to override the prepotent response. They differed only in whether the incongruency was expected beforehand or needed to be detected in order to initiate the exertion of control. In line with the alarm system hypothesis, Carter et al found a larger anterior cingulate response during unexpected incongruent trials than expected incongruent trials. In other words, when the detection of need for control was decoupled from the exercise of control, the activity of the anterior cingulate was associated with the former and not the latter.

These findings also demonstrate that the anterior cingulate's response to conflict varies as a function of the explicit expectations of the C-system. This flexibility has paradoxical consequences for the phenomenology associated with the alarm system. In an fMRI study of emotional responses to pain (Sawamoto et al., 2000; also see Kropotov, Crawford & Polyakov, 1997; Rainville et al., 1997), participants underwent three blocks of trials during which they were exposed to painful or non-painful stimulation. In two of the blocks, participants were given the accurate expectation that all the trials would be painful or painless. In a third block, participants expected mostly painless trials and a few painful trials. Not surprisingly, given the strong relationship between subjectively experienced pain and anterior cingulate activations, the anterior cingulate was more active during the predictable pain trials than the predictable painless trials. What was surprising was that unpredictable painless trials evoked an anterior cingulate response that looked more like the predictable painful trials than the predictable painless trials. Remarkably, the unpredictable painless trials were also reported to be significantly more painful than the predictable painless trials, and this bias in the subjective experience of pain correlated with anterior cingulate activation and with no other region in the brain. These results suggest that the subjective experience of pain and other error signals correlate with anterior cingulate activity, but that anterior cingulate activity is not purely a function of the objective level of conflict detected by the X-system. Rather, the anterior cingulate seems to create what it is looking for, at least to some extent. How the anterior cingulate manufactures some of the pain it is monitoring for is still a mystery, though it may turn out to be analogous to the Heisenbergian measurement problem in quantum physics. That is, whatever "spotlight" is used by the anterior cingulate to measure activity in the X-system may in fact alter this same activity. This process in the anterior cingulate is also analogous to, and may be the neural instantiation of, Wegner's (1994) ironic monitoring in which the attempt to monitor whether one is successfully avoiding thoughts about a particular object (a white bear) actually produces that unwanted thought.

In an earlier section, we suggested that asking a question about an actor's dispositions might prime

the observer's X-system by way of the connections between prefrontal cortex and anterior ITC. The activation of the dispositional category would make the network more likely to settle in a dispositional valley of coherence. The "pain X expectation" study of Sawamoto et al (2000) suggests that the anterior cingulate may play a different functional role in inference processes. Both are types of neural self-fulfilling prophecies, but the prefrontal to anterior ITC circuit leads to the assimilation of ambiguous stimuli to hypothesis-relevant categories, whereas anterior cingulate conflict monitoring leads to the creation of expected errors and incoherencies where none exist. This suggests that the personality and contextual factors that create pre-existing doubt regarding the accuracy of one's own inference processes may dramatically affect the impact of the C-system in attribution, especially when the input would otherwise be coherently assimilated by the X-system.

Finally, the sensitivity of the anterior cingulate to conflict signals in the X-system appears to be dulled by cognitive load. Under cognitive load, the anterior cingulate showed a smaller increase in activity in response to pain than did other areas of the pain network (Petrovic, Petersson, Ghatan, Stone-Elander & Ingvar, 2000). When the outputs associated with a particular process are absent under conditions of cognitive load it is typically assumed that the process itself requires controlled processing resources to operate. The anterior cingulate, however, is in the business of detecting when control is needed, not exerting control itself. Given that the anterior cingulate's sensitivity to the incoherence of X-system processes can be disrupted by cognitive load, a new class of explanations for cognitive load effects is suggested. Controlled attribution processes (correction, integration) may sometimes be absent under conditions of cognitive load because the anterior cingulate is less likely to detect the need for controlled process intervention, and not just because controlled attribution processes lack the cognitive resources to operate. If the anterior cingulate sensitivity is dulled by cognitive load, relatively large X-system coherences may go unnoticed and the rest of the C-system will never be called upon to intervene. In the end, the two ways that cognitive load can derail performance are complementary, not competing, views. Both may occur under most forms of cognitive load, but an understanding of the role of cognitive load in altering the perception of need for control should contribute to an evolving view of the human mind and how we view its accountability for judgments and behaviors.

Prefrontal Cortex, Propositional Thought, and Implementing Control. If the anterior cingulate is the alarm system at the gate, then the prefrontal cortex is the lord of the manor. Differences across species in the ability to exert control are correlated with the ratio of prefrontal cortex to the rest of the brain (Fuster, 1997). The C-system is turned on when the X-system's output is incoherent, and thus the prefrontal cortex is largely in the business of challenging or correcting the X-system's output. The C-system allows human beings to modify their judgments and behaviors in light of information that either eludes the X-system or that the X-system misinterprets (for reviews see Fuster, 1997; Miller & Cohen, 2001; Rolls, 1999). Individually, these operations consist of generating and maintaining symbols in working memory, combining these symbols with rule-based logical schemes, and biasing the X-system and motor systems to behave in line with the goals and outputs of working memory.

One of the oldest findings in neuropsychology is that patients with damage to the prefrontal cortex are impaired in the generation of new goals and hypotheses (Milner, 1963). For instance, in the Wisconsin Card Sorting Task (WCST; Grant & Berg, 1948), cards depicting different numbers of colored shapes must be sorted on the basis of their color, shape, or number. Once participants successfully learn the sorting rule, a new rule is chosen unbeknownst to the participant. People with prefrontal damage continue to use the old rule long after a new rule has been put in place. These people are unable to generate new hypotheses in the face of prepotent hypotheses. A recent fMRI study suggests that internally generated inferences are associated with activations near the very front of prefrontal cortex (Christoff & Gabrieli, 2000).

The ability to hold information in active memory has long been associated with prefrontal cortex. (Braver et al., 1997; Smith & Jonides, 1999). Typically, experiments have examined the number of items that people can simultaneously keep in working memory (Miller, 1956), and have produced two findings. First, although working memory can maintain almost any variety of information and is thus enormously flexible, it is only able to juggle a few items at once. Second, because the experimental tasks are normally so decontextualized (e.g., remembering a string of digits), they may seem to suggest that the main purpose of working memory is to allow people to remember telephone numbers when pens and paper are out of reach. Indeed, only in the last decade has it become clear that the contents of working memory are often our goal representations, or the new rules we are currently trying to use in a novel situation. According to

Cohen's model of prefrontal control, the goal representations in prefrontal cortex serve to prime weaker, but relevant, neural representations in the X-system (Miller & Cohen, 2001; O'Reilly, Braver & Cohen, 1999; Tomita, Ohbayashi et al., 1999-Nature). Thus, by consciously thinking "pay attention to the color of the word" in the Stroop task, downstream areas associated with color identification are given a boost so that they can compete more effectively with word identification units (Banich et al., 2000; MacDonald et al., 1998).

Hippocampus and Memory for Exceptions. The hippocampus and the surrounding structures in the medial temporal lobes, are crucial for episodic memory, which is the memory for specific events and the context in which they happened (Squire & Knowlton, 2000). Semantic memory in lateral temporal cortex and episodic memory in the medial temporal lobe work together to capture the two types of variance in the world that need to be learned. Semantic memory represents what is common across situations, whereas episodic memory represent the exceptions when something important or unexpected happens that thwarts the X-system. McClelland, McNaughton and O'Reilly (1995) demonstrated computationally that it is virtually impossible to build a mechanism that can perform both memory operations simultaneously without catastrophic memory losses under certain common processing conditions.

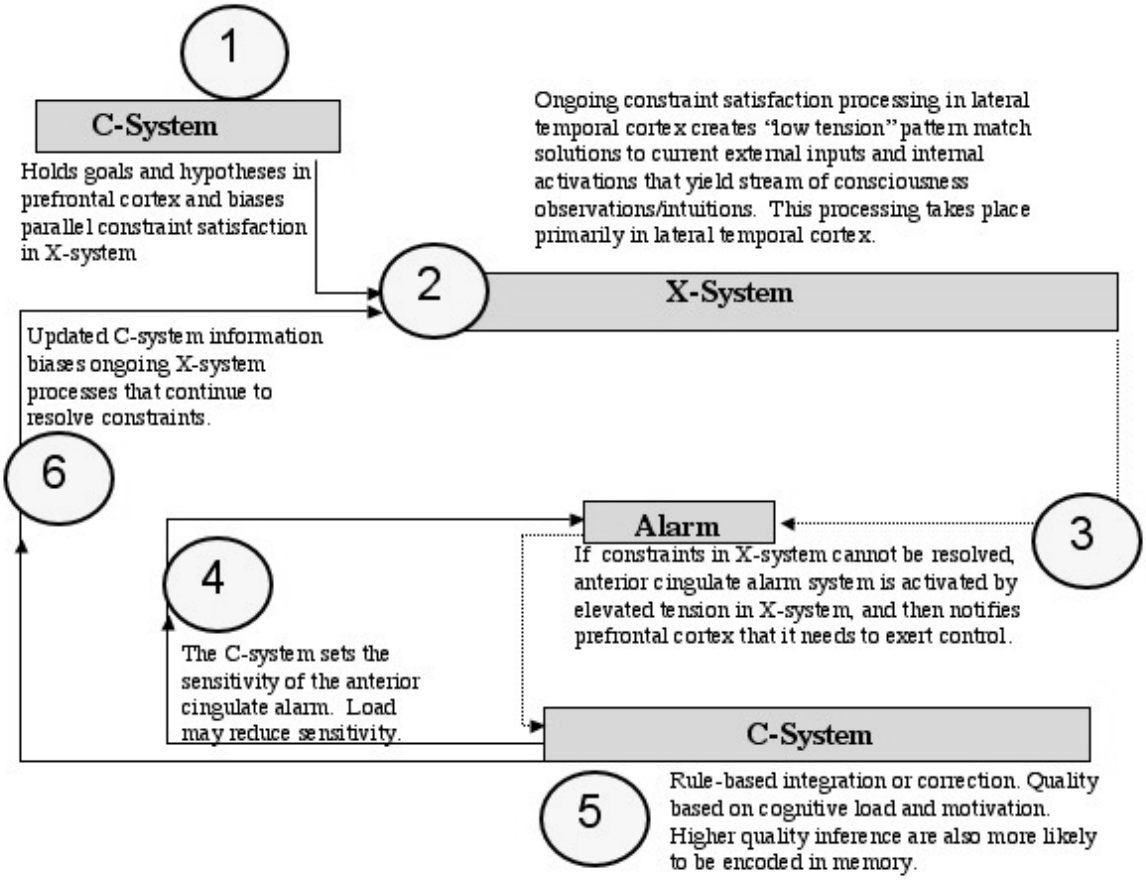
Enduring episodic memories for specific events are formed when the X-system hits a road block and the C-system comes to the rescue. Consequently, situations that require more control tend to lead to stronger episodic memories in the hippocampus. Researchers have long known that depth of explicit processing is correlated with the strength of episodic memory (Craik & Tulving, 1975) but not with the strength of implicit memory (Graf & Mandler, 1984). More recently, fMRI studies have shown that recall is correlated with the amount of prefrontal and hippocampal activity at encoding (Brewer, Zhao, Desmond, Glover, & Gabrieli, 1998; Otten, Henson, & Rugg, 2001; Wagner et al., 1998). Less attention has been given to the kinds of events that give rise to greater prefrontal activity at encoding. Our analysis suggests that this occurs when the X-system cannot settle into a coherent state.

In describing the representations of lateral temporal cortex, we suggested that the causal powers of situations may not be as well represented as their featural associates. In contrast, work with rats has shown that the hippocampus may be necessary to learn about the way a situation conditionalizes the implications of entities in the situation. Rats with hippocampal lesions are quite capable of learning that a tone predicts a subsequent shock (Kim & Fanselow, 1992; Philips & LeDoux, 1992). The rats cannot, however, learn that the tone predicts shock in a blue but not a yellow box. Essentially, the rat cannot take into account the impact of the different situations on the value of the tone when its hippocampus is removed. A recent experiment with rats has shown the hippocampus is necessary for discrimination between situations in general, without regard for what those situations predict (Frankland, Cestari, Filipkowski, McDonald, & Silva, 1998). Given that episodic retrieval generally involves prefrontal cortex (Henson, Shallice, & Dolan, 1999), the use of situational or personal information in terms of their causal powers to shape behavior would seem to occur only when the C-system is activated.

This creates something of a Catch-22 in the system. Protecting against the correspondence bias requires awareness and use of the causal powers of situational constraints in the inference process. Although situations per se may be represented in the X-system, the knowledge of their causal powers is stored in the C-system (Brewer et al., 1998; Wagner et al., 1998; Eldridge et al, 2000; LaPage, Ghaffar, Nyberg & Tulving, 2000). The problem is that the C-system is only activated when the X-system fails. In this sense, the C-system is like a safe that contains the key to the safe. The causal meaning of situational information represented by the C-system can only be used if the X-system has been prevented from settling into a valley of coherence. Unfortunately, the best way to keep the X-system from settling is to invoke the causal implications of situational constraints in order to sensitize the anterior cingulate to smaller incoherences in the X-system. In short, it may be necessary for the observer to have a pre-existing doubt about the veracity of his or her own attributional inferences to activate the C-system—a tendency that we suspect is not all that prevalent. This represents yet another way in which neural architecture may encourage the correspondence bias.

IV. Contributions to Attribution Theory

The X- and C-systems are new labels, but they are not new discoveries. Indeed, thinkers since Plato have tried to explain why people think, feel, and act as they do by dividing the mind into interacting (and occasionally warring) parts (see Gilbert, 1999; Hundert, 1995). This explanatory strategy has endured for two millennia because the notion of interacting parts is more than a metaphor: It is precisely how the brain operates. With that said, it is very difficult to specify with any precision the boundaries of those parts and the nature of their



interaction. Our analysis is broadly consistent with many dual-process models that pit conscious, cognitive deliberation against unconscious, perceptual inference (Ashby-Psych Rev; James, 1890; Petty & Cacioppo, 1986; Chaiken, Liberman, & Eagly, 1989; Epstein, 1990; Fiske & Neuberg, 1991; Sloman, 1996; Smith & DeCoster, 1999). Nonetheless, we believe that our analysis paints a somewhat fuller picture by locating these systems in the brain, by suggesting some of the ways in which the systems may operate and interact, by specifying the circumstances under which the C-system will be activated, and by specifying the consequences of the fact that the X-system is always activated. Before describing the relevance of this analysis for attribution theory, it may behoove us to summarize its key points.

A. Reflexion-Reflection Vs. Dual-Process Models

The X-system is responsible for what psychologists generally refer to as automatic processes, and what ordinary people call perception. It is instantiated in the lateral temporal cortex, basal ganglia, and amygdala, and its main function is to produce the stream of consciousness that we experience as the real world—not just the objects of the real world, but also the semantic and affective associations of those objects, which are also experienced as the real world. Although the X-system may appear to be using the symbolic logic that characterizes the operations of the C-system, it actually performs similarity-based pattern-matching operations on incoming data, which are continuously assimilated to valleys of coherence. The actions of the X-system are described by the parallel constraint satisfaction processes of connectionist theory.

The C-system is responsible for what psychologists generally refer to as controlled processes and reflective awareness, and what ordinary people call thought. It is instantiated in the anterior cingulate, prefrontal cortex, and hippocampus. The anterior cingulate is activated by problems in the pattern-matching operations of the X-

system, and it in turn activates the prefrontal cortex. The prefrontal cortex can use symbolic logic to solve problems that the X-system cannot, and it uses this ability to influence or override the X-system. The hippocampus remembers the situations in which the C-system was activated, presumably to facilitate problem-solving the next time a similar situation arises.

The flow of information through the X- and C-systems during the process of dispositional inference is shown in Figure 4. Four differences between this model and standard dual-process models are worth noting. First, both the X- and C-systems can be involved in the process from beginning to end. The C-system may be involved early in the process when it represents the attributional question or hypothesis and biases the operations of the X-system, and it may be involved later on when it generates alternative solutions that supplant those generated by the X-system. The X-system, on the other hand, is continuously engaged in constraint satisfaction processes to categorize and identify events and actors in the world, and these operations produce the stream of consciousness.

The second point of divergence with dual-process models is that our model is fully recurrent, with each system sending information to and receiving information from the other. Standard dual-process models posit a sequential path from the observer's initial goal to the automatic categorization processes and to controlled causal reasoning processes. In the current model, however, the goals and causal reasoning processes of the C-system bias the ongoing attempts of the X-system to make sense of the world. Furthermore, the initiation of the causal reasoning process depends not only on cognitive load and motivation, as suggested by standard dual-process models, but also on the coherence of X-system's solutions (Thagard, 2000).

Third, our model describes an alarm system that wakes the C-system when sustained mutual inhibition in the X-system passes some threshold. Although the alarm is triggered without active monitoring by the C-system, the C-system can set the threshold on the alarm system and thereby determine when it will be activated. Fourth and finally, our model suggests that the sensitivity of the alarm system may be affected by cognitive load. Dual-process models generally assume that cognitive load impairs the C-system's ability to implement *causal reasoning*. Our model suggests that in addition, cognitive load may prevent the C-system from being notified that its services are needed.

B. How are Situational and Dispositional Information Used?

Our model sheds light on the different ways that the information contained in a behavioral episode may be represented in the brain and used during attributional inference. The X-system represents associations among the features of dispositions, situations, and behaviors. The X-system is biased towards those features that are present in the question or hypothesis posed by the C-system. Parallel constraint satisfaction processes change the initial representations of these different sources of information. It is through such processes that the X-system disambiguates behavioral information and identifies it as consistent with the representations of the situation and actor's dispositions. If the X-system fails to settle into a valley of coherence, the C-system may be engaged to help generate a conclusion, though the C-system's involvement depends on the sensitivity of the alarm system to incoherences of the X-system. When the C-system is brought on line, the disambiguated behavior identifications are treated as facts, and dispositions and situations are treated as potential causal explanations. Under optimal conditions, people may engage in diagnostic evaluation of these causes. Such evaluations may show that the behavior is consistent with both dispositional and situational causes and thus lead to moderate attributional inferences. However, under suboptimal conditions, when people lack the motivation or attentional resources needed, pseudodiagnostic processing may prevail and people may focus on a single hypothetical cause, assess the consistency of the identified behavior with that cause, and disregard other potential causes. When the favored hypothetical cause is dispositional, pseudodiagnostic evaluation is likely to lead to overconfident dispositional attributions (a correspondence bias). When the favored hypothetical cause is situational, pseudodiagnostic evaluation is likely to yield overconfident situational attributions.

A final cautionary note regarding attribution methodologies is in order. Researchers often treat paper-and-pencil scenarios and visually observed behavior as equivalent delivery systems for behavioral information. Our model suggests that they are not equivalent and may have very different consequences for the inferential path taken through the X- and C-systems. Observed behavior must necessarily pass through the X-system before reaching the C-system, if the C-system is ever reached at all. Linguistic information in the form of vignettes, sentences, or overheard conversations may have a direct pipeline to the C-system and avoid the X-system altogether. Recall that the C-system uses

symbolic logic and that the X-system does not. Language is fundamentally symbolic and may well be the basis of symbolic logic in the C-system (Fodor, 1975). As such, presenting subjects with the "symbolic equivalent" of behavior is fundamentally different than presenting subjects with actual behavior, because the former requires the involvement of the C-system while the latter does not. Presenting people with the phrase "an attractive person" probably activates at best a faint image in the stream of consciousness for most people, while fully activating the symbol in reflective awareness. In fact, in order to make the weak image stronger, the C-system will probably become increasingly activated as it helps generate a richer image. Presenting people with an attractive person, on the other hand, always generates a strong percept in the stream of consciousness and only alerts the C-system if the percept is incoherent.

VI. Coda

Human beings are the most complex and significant stimuli that human beings ever encounter, and understanding what makes other people behave as they do is a critical determinant of our health, wealth, happiness, and survival. Social psychologists have made dramatic progress over the last half-century in understanding the psychological processes by which attributions are made, but as we hope this chapter has made clear, there is much more left to know. As revolutions come and go, psychology's subdisciplines wax and wane, radically transforming themselves as interests shift with each new generation. Social psychology has been unusual in its ability to maintain a steady focus on a core set of intellectual problems and to use the fruits of each new scientific revolution to solve them. Psychology's newest revolution is just beginning to unfold, and with it comes all the usual naïve talk about *this* revolution being the last one. We do not believe that we have come to the end of history, and we do not believe that brain science will supplant social psychology, but we do believe that something very important is happening next door. So we have done what social psychologists always do under such circumstances, sneaking into the neighbor's yard, taking what is best, and bringing it home to help illuminate the enduring questions that motivate our discipline. We hope that as brain science matures, it will become wise enough to rob us in return.

VII. References

Adolphs, R. (1999). Social cognition and the human brain. *Trends in Cognitive Sciences, 3,* 469-479.

Aizenstein, H. J., MacDonald, A. W., Stenger, V. A., Nebes, R. D., Larson, J. K., Ursu, S., & Carter, C. S. (2000). Complementary category learning systems identified using event-related functional MRI. *Journal of Cognitive Neuroscience, 12*, 977-987.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences, 4*, 267-278.

Anderson, N. H. (1974). Information integration: A brief survey. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (pp. 236-305). San Francisco: Freeman.

Aristotle (1941) *The basic works of Aristotle*. New York: Random House.3-

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258-290.

Ashby, F. G., Alfonso-Reese, L., Turken, A. U., & Waldron, E. M., (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442-481.

Baciu, M. V., Bonaz, B. L., Papillon, E., Bost, R. A., Le Bas, J. F., Fournet, J., & Segebarth, C. M. (1999). Central processing of rectal pain: a functional MR imaging study. *American Journal of Neuroradiology, 20*, 1920-1924.

Baddeley, A. (1986). *Working memory.* Oxford, England: Clarendon Press.

Banich, M. T., Milham, M. P., Atchley, R. A., Cohen, N. J., Webb, A., Wszalek, T., Kramer, A. F., Liang, Z., Barad, V., Gullett, D., Shah, C., & Brown, C. (2000). Prefrontal regions play a predominant role in imposing an attentional 'set': Evidence from fMRI. *Cognitive Brain Research, 10*, 1-9.

Bar, M., Tootell, R. B. H., Schacter, D. L., Greve, D. N., Fischl, B., Mendola, J. D., Rosen, B. R., & Dale, A. M. (2001). Cortical mechanisms specific to explicit visual object recognition. *Neuroon, 29*, 529-535.

Barbas, H. (2000). Connections underlying the synthesis of cognition, memory, and emotion in primate prefrontal cortices. *Brain Research Bulletin, 52*, 319-330.

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. *Unintended thought.* J. S. Uleman, & J. A. Bargh. New York, Guilford**:** 3-51.

Baumeister, R. F. (1990). Anxiety and deconstruction. On escaping the self. In J. M. Olson & M. P. Zanna (Eds.), *Self-inference processes: The Ontario symposium* (Vol. 6, pp. 259-291). Hillsdale, NJ: Erlbaum.

Baumeister, R. F. (1986). *Identity*. New York: Oxford University Press.

Berns, G. S., Cohen, J. D., & Mintun, M. A. (1997). Brain regions responsive to novelty in the absence of awareness. *Science, 276*, 1272-1275.

Berthoz, S., Artiges, E., Van de Moortele, P. F., Poline, J. B., Rouquette, S., & Martinot, J. L. (2000). Emotion-induced stimuli processing in alexithymia: An fMRI study. *Biological Psychiatry, 47(suppl.)*, 110s.

Bierbrauer, G. (1979). Why did he do it? Attribution of obedience and the phenomenon of dispositional bias. *European Journal of Social Psychology, 9*, 67—84.

Bonda, E., Petrides, M., Ostry, D., & Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal of Neuroscience, 16*, 3737-3744.

Booth, M. C. A., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal cortex. *Cerebral Cortex, 8*, 510-523.

Botvinick, M. M., Braver, T. D., Barch, D. M., Carter, C. S., & Cohen, J. D. (in press). Conflict monitoring and cognitive control. *Psychological Review*.

Boucart, M., Meyer, M. E., Pins, D., Humphreys, G. W., Scheiber, C., Gounod, D., & Foucher, J. (2000). Automatic object identification: an fMRI study. *NeuroReport, 11*, 2379-2383.

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage, 5*, 49-62.

Brentano, F. (1874/1995). *Psychology from an empirical standpoint*. London: Routledge

Brewer, J. B., Zhao, Z., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. E. (1998). Making memories: Brain activity that predicts how well visual experience will be remembered. *Science, 281*, 1185-1187.

Brunswik, E. (1947). *Systematic and representative design of psychological experiments*. Berkeley: University of California Press.

Buchanan, S. L., Valentine, J., & Powell, D. A. (1985). Autonomic responses are elicited by electrical stimulation of medial but not lateral frontal cortex in rabbits. *Behavioral Brain Research, 18*, 51-62.

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences, 4*, 215-222.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science, 280*, 747-749.

Carter, C. S, MacDonald, A. M., Botvinick, M., Ross, L. L., Stenger, V. A., Noll, D., & Cohen, J. D. (2000). Parsing executive processes: Strategic vs. evaluative functions of the anterior cingulate cortex. *Proceedings of the National Academy of Science, 97*, 1944-1948.

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic processing within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 212-252). New York: Guilford Press.

Cheng, K., Saleem, K. S., & Tanaka, K. (1997). Organization of corticostriatal and corticoamygdala projections arising from the anterior inferotemporal area TE of the Macaque monkey: A Phaseolus vulgaris leucoagglutinin study. *Journal of Neuroscience, 17*, 7902-7925.

Christoff, K., & Gabrieli, J. D. E. (2000). The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within human prefrontal cortex. *Psychobiology, 28*, 168-186.

Corneille, O., Leyens, J. P., Yzerbyt, V. Y., & Walther, E. (1999). Judgeability concerns: The interplay of information, applicability, and accountability in the overattribution bias. *Journal of Personality and Social Psychology, 76*, 377-387.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268-294.

Critchley, H. D., Corfield, D. R., Chandler, M. P., Mathias, C. J., & Dolan, R. J. (2000). Cerebral correlates of autonomic cardiovascular arousal: a functional neuroimaging investigation in humans. *Journal of Physiology, 523*, 259-270.

Csikszentmihalyi, M. (2000). Beyond boredom and anxiety: Experiencing flow in work and play. New York: Jossey-Bass.

Cunningham, W. A., Johnson, M. K., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2001, April). *An fMRI study on the conscious and unconscious evaluations of social groups*. Paper presented at the UCLA Conference on Social Cognitive Neuroscience, Los Angeles, CA

Decety, J., & Grezes, J. (1999). Neural mechanisms subserving the perception of human actions. *Trends in Cognitive Sciences, 3*, 172-178.

Dennett, D. (1984). *Elbow room: The varieties of free will worth wanting*.

Cambridge, MA, MIT Press.

Derbyshire, S. W. G., Vogt, B. A., & Jones, A. K. P. (1998). Pain and Stroop interference tasks activate separate processing modules in anterior cingulate cortex. *Experimental Brain Research, 118*, 52-60.

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience, 4*, 2051-2062.

Dewey, J. (1910). *How we think*. Boston: D. C. Heath & Co.

Donovan, S., & Epstein, S. (1997). Conjunction problem can be attributed to its simultaneous concrete and unnatural representation, and not to conversational implicature. *Journal of Experimental Social Psychology, 33*, 1-20.

Dougherty, D. D., Shin, L. M., Alpert, N. M., Pitman, R. K., Orr, S. P., Lasko, M., Macklin, M. L., Fischman, A. J., & Rauch, S. L. (1999). Anger in healthy men: A PET study using script-driven imagery. *Biological Psychiatry, 46*, 466-472.

Dreyfus, H. L. (1991). *Being-in-the-world: A commentary on Heidegger's being and time, division I*. Cambridge, MA: MIT Press.

Dweck, C. S., Hong, Y., & Chiu, C. (1993). Implicit theories: Individual differences in the likelihood and meaning of dispositional inference. *Personality and Social Psychology Bulletin, 19*, 633-643.

Eldridge, L. L., Knowlton, B. J., Furmanski, C. S., Bookheimer, S. Y., & Engel, S. A. (2000). Remembering episodes: A selective role for the hippocampus during retrieval. *Nature Neuroscience, 3*, 1149-1152.

Ellsworth, P. C. (1994). William James and emotion: Is a century of fame worth a century of misunderstanding? *Psychological Review, 101*, 222-229.

Epstein, S. (1990). Cognitive experiential self-theory. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 165-192). New York: Guilford Press.

Evans, J. B. T. (1989). *Biases in human reasoning*. Hove, UK: Erlbaum.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*, 171-180.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1-74). San Diego, CA: Academic Press.

Fodor, J. A. (1975). *The language of thought*. New York: Crowell.

Fodor, J. A. (2000). *The mind doesn't work that way: The scope and limits of computational psychology.* Cambridge, MA: MIT Press.

Foltz, E. L., & White, L. E. (1968). The role of rostral cingulumotomy in "pain" relief. *International Journal of Neurology, 6*, 353-373.

Frankland, P. W., Cestari, V., Filipowski, R. K., McDonald, R. J., & Silva, A. J. (1998). The dorsal hippocampus is essential for context discrimination but not for contextual conditioning. *Behavioral Neuroscience, 112*, 863-874.

Fridja, N. H. (1986). *The emotions*. New York: Cambridge University Press.

Fuster, J. M. (1997). *The prefrontal cortex: Anatomy, physiology, and neuropsychology of the frontal lobe*. New York: Lippincott-Raven.

Garrard, P., & Hodges, J. R. (1999). Semantic dementia: Impliations for the neural basis of language and meaning. *Aphasiology, 13*, 609-623.

George, M. S., Ketter, T. A., Parekh, P. I., & Horwitz, B. (1995). Brain activity during transient sadness and happiness in healthy women. *American Journal of Psychiatry, 152*, 341-351.

Gerlach, C., Law, I., Gade, A., & Paulson, O. B. (2000). Categorization and category effects in normal object recognition: A PET study. *Neuropsychologica, 38*, 1693-1703.

Gilbert, D. T. (1989). Thinking lightly about others. Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 189-211). New York: Guilford.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*, 21-30.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46*, 107-119.

Gilbert, D. T. (1998a). Ordinary personology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., pp. 89-150). New York: Oxford University Press.

Gilbert, D. T. (1998b). Speeding with Ned: A personal view of the correspondence bias. In J. M. Darley & J. Cooper (Eds.), *Attribution and social interaction: The legacy of Edward E. Jones* (pp. 5-36).. Washington, DC: American Psychological Association.

Gilbert, D. T., Krull, D. S., & Pelham, B. W. (1988). Of thoughts unspoken: Social inference and the self-regulation of behavior. *Journal of Personality and Social Psychology, 55*, 685-694.

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology, 54*, 733-740.

Goel, V., & Dolan, R. J. (2000). Anatomical segregation of component processes in an inductive inference task. *Journal of Cognitive Neuroscience, 12*, 110-119.

Goel, V., Gold, B., Kapur, S., & Houle, S. (1997). The seats of reason? An imaging study of deductive and inductive reasoning. *NeuroReport, 8*, 1305-1310.

Graf, P., & Mandler, G. (1984). Activation makes words more accessible, but not necessarily more retrievable. *Journal of Verbal Learning & Verbal Behavior, 23*, 553-568.

Grant, D. A., & Berg, E. A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card sorting problem. *Journal of Experimental Psychology, 38*, 404-411.

Hasselmo, M. E., Rolls, E. T., & Baylis, G. C. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioral Brain Research, 32*, 203-218.

Hariri, A. R., Bookheimer, S. Y., & Mazziotta, J. C. (2000). Modulating emotional responses: Effects of a neocortical network on the limbic system. *Neuroreport, 11*, 43-48.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*, 223-233.

Heatherton, T. F., & Baumeister, R. F. (1991). Binge eating as escape from self-awareness. *Psychological Bulletin, 110*, 86-108.

Hebb, D. O. (1949). *The organization of behavior*. New York: John Wiley and Sons.

Heidegger, M. (1927/1962). *Being and time.* New York: Harper & Row.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Helmholtz, H. V. (1925). *Treatise on psychological optics*. Vol. 3. Menasha,

Wisconsin, Banta. (Original work published in 1910).

Henson, R. N. A., Shallice, T., & Dolan, R. J. (1999). Right prefrontal cortex and episodic memory retrieval: a functional MRI test of the monitoring hypothesis. *Brain, 122*, 1367-1381

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review, 94*, 319-340.

Hoffman, E. A. & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience, 3*, 80-84.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79*, 2554-2558.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences, 81*, 3088-3092.

Houde, O., Zago, L., Mellet, E., Moutier, S., Pineau, A., Mazoyer, B., & Tzourio-Mazoyer, N. (2000). Shifting from the perceptual brain to the logical brain: The neural impact of cognitive inhibition training. *Journal of Cognitive Neuroscience, 12*, 721-728.

Howard, R. J., Brammer, M., Wright, I., Woodruff., P. W., Bullmore, E. T., & Zeki, S. (1996). A direct demonstration of functional specialization within motion-related visual and auditory cortex of the human brain. *Current Biology, 6*, 1015-1019.

Hundert, E. M. (1995). *Lessons from an optical illusion: On nature and*

nurture, knowledge and values. Cambridge, MA, Harvard University Press.

Hurt, R. W., & Ballantine, H. T. (1974). Stereotactic anterior cingulate lesions for persistent pain: a report on 68 cases. *Clinical Neurosurgery, 21*, 334-351.

Husserl, E. (1913/1962). *Ideas*. New York: Collier Press.

Ichheiser, G. (1943). Misinterpretations of personality in everyday life and the psychologist's frame of reference. *Character and Personality, 12*, 145-160.

Ischeiser, G. (1949). Misunderstandings in human relations: A study in false social perception. *American Journal of Sociology, 55*, part 2

James, W. (1890/1950). *The principles of psychology.* New York: Dover.

James, W. (1894). The physical basis of emotion. *Psychological Review, 1*, 516—529.

Johnson, M. K., & Raye, C. L. (1981). "Reality monitoring." *Psychological*

Review, 88, 67-85.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 220-266). New York: Academic Press.

Just, M. A., Carpenter, P. A., & Varma, S. (1999). Computational modeling of high-level cognition and brain function. *Human Brain Mapping, 8*, 128-136.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, pp. 192-240). Lincoln: University of Nebraska Press.

Kelley, H.H. (1972). Causal schemata and the attribution process. In E.E. Jones, D.E. Kanouse, H.H. Kelley, R.E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 151-174). Morristown, NJ: General Learning Press.

Kim, J. J., & Fanselow, M. S. (1992). Modality-specific retrograde amnesia of fear. *Science, 256*, 675-577.

Kimbrell, T. A., George, M. S., Parekh, P. I., Ketter, T. A., Podell, D., M., Danielson, A. L., Repella, J. D., Benson, B. E., Willis, M. W., Herscovitch, P., & Post, R. M. (1999). Regional brain activity during transient self-induced anxiety and anger in healthy adults. *Biological Psychiatry, 46*, 454-465.

Klein, S. B., & Kihlstrom, J. F. (1998). On bridging the gap between social-personality psychology and neuropsychology. *Personality and Social Psychology Review, 2*, 228-242.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience, 21*, RC159:105.

Kohler, W. (1947). *Gestalt psychology*. New York: Liveright.

Kropotov, J. D., Crawford, H. J., & Polyakov, Y. I. (1997). Somatosensory event-related potential changes to painful stimuli during hypnotic analgesia: anterior cingulate cortex and anterior temporal cortex intracranial recordings. *International Journal of Psychophysiology, 27*, 1-8.

Krull, D. S. (1993). Does the grist change the mill? The effect of the perceiver's inferential goal on the process of social inference. *Personality and Social Psychology Bulletin, 19*, 340-348.

Kunda, Z. & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel constraint satisfaction theory. *Psychological Review, 103*, 284-308.

Ladabaum, U., Minoshima, S., & Owyang, C. (2000). Pathobiology of visceral pain: molecular mechanisms and therapeutic implications V. Central nervous system processing of somatic and visceral sensory signals. *American Journal of Physiology: Gastrointestinal and Liver Physiology, 279*, G1-6.

Lane, R. D., Reiman, E. M., Axelrod, B., Yun, L.-S., Holmes, A., & Schwartz, G. E. (1998). Neural correlates of levels of emotional awareness: Evidence of an interaction between emotion and attention in the anterior cingulate cortex. *Journal of Cognitive Neuroscience, 104*, 525-535.

LaPage, M., Ghaffar, O., Nyberg, L., & Tulving, E. (2000). Prefrontal cortex and episodic memory retrieval mode. *Proceedings of the National Academy of Science, 97*, 506-511.

Lazarus, R. S. (1991). *Emotion and adaptation*. New York NY: Oxford University Press.

LeDoux, J. E. (1996). The emotional brain: The mysterious underpinnings of emotional life. New York: Simon & Schuster.

Lieberman, M. D. (2000). Intuition: A social cognitive neuroscience approach. *Psychological Bulletin, 126*, 109-137.

Lieberman, M. D., Chang, G. Y., Chiao, J., Bookheimer, S. Y., & Knowlton, B. J. (2001). *An event-related fMRI study of artificial grammar learning*. Unpublished manuscript. University of California, Los Angeles.

Lieberman, M. D., Gilbert, D. T., & Jarcho, J. M. (2001). Culture's impact on the controlled, but not the automatic, processes in attributional inference. Unpublished manuscript. *University of California, Los Angeles.*

Lieberman, M. D., Hariri, A. R., & Bookheimer, S. Y. (2001, April). *Controlling automatic stereotyping: An fMRI study*. A paper presented at the UCLA Conference on Social Cognitive Neuroscience, Los Angeles, CA.

Lorberbaum, J. P, Newman, J. D., Dubno, J. R., Horwitz, A. R., Nahas, Z., Teneback, C. C., Bloomer, C. W., Bohning, D. E., Vincent, D., Johnson, M. R., Emmanuel, N.,Brawman-Mintzer, O., Book, S. W., Lydiard, R. B., Ballenger, J. C., & George, M. S. (1999). Feasibility of using fMRI to study mothers responding to infant cries. *Depression and Anxiety, 10*, 99-104.

Lumer, E. D. Friston, K. J., & Rees, G. (1998). Neural correlates of perceptual rivalry in the human brain. *Science, 280*, 1930-1934.

MacDonald, A. W., Cohen, J. D., Stenger, V. A. & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science, 288, 1835-1838.

Mathew, R. J., Wilson, W. H., Chiu, N. Y., Turkington, T. G., Degrado, T. R., & Coleman, R. E. (1999). Regional cerebral blood flow and depersonalization after tetrahydrocannbinol administration. *Acta Psychiatrica Scandinavica, 100*, 67-75.

Mayberg, H. S., Liotti, M., Brannan, S. K., McGinnis, S., Mahurin, R. K., Jerabek, P. A., Silva, J. A., Tekell, J. L., Martin, C. C., Lancaster, J. L. & Fox, P. T (1999). Reciprocal limbic-cortical function and negative mood: Converging PET findings in depression and normal sadness. *American Journal of Psychiatry, 156*, 675-682.

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

Mikami, A., Nakamura, K., & Kubota, K. (1994). Neuronal responses to photographs in the superior temporal sulcus of the rhesus monkey. *Behavioral Brain Research, 60*, 1-13.

Miller, E. K., & Cohen, J. D. (in press). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*.

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

Miller, G. A. (1981). Trends and debates in cognitive psychology. *Cognition, 10*, 215-225.

Miller, G. A., Galanter, E., & Pribram, K. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart, & Winston.

Milner, B. (1963). Effects of different brain lesions on card sorting. *Archives of Neurology, 9*, 90-100.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neuroscience, 6*, 414-417.

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review, 102*, 331-355.

Moscowitz, G. B., & Roman, R. J. (1992). Spontaneous trait inferences as self-generated primes: Implications for conscious social judgment. *Journal of Personality and Social Psychology, 62*, 728-738.

Mummery, C. J., Shallice, T., & Price, C. J. (1999). Dual-process model in semantic priming: A functional imaging perspective. *NeuroImage, 9*, 516-525.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*, 435-450.

Nakamura, K., & Kubota, K. (1996). The primate temporal pole: its putative role in object recognition and memory. *Behavioral Brain Research, 77*, 53-77.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Ed.), *Basic processes in reading: Visual word recognition* (pp. 264-336). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Nisbett, R. E., Krantz, D. H., Jepson, C., & Fong, G. T. (1982). Improvinginductive inference. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Ochsner, K. N., Bunge, S. A., Gross, J., & Gabrieli, J. D. E. (2001, April). *Rethinking feelings: Exploring the neurocognitive mechanisms of emotion control*. Paper presented at the UCLA Conference on Social Cognitive Neuroscience, Los Angeles, CA.

Ochsner, K. N. & Feldman-Barrett, L. (2001). A multiprocess perspective on the neuroscience of emotion. To appear in T. J. Mayne, G. Bonnano (Eds.). *Emotion: Current issues and future directions* (pp. 39-81). New York: Guilford Press.

Ochsner, K. N., Kosslyn, S. M., Cosgrove, G. R., Cassem, E. H., Price, B. H., Nierenberg, A. A., & Rauch, S. L. (2001). Deficits in visual cognition and attention following bilateral anterior cingulotomy. *Neuropsychologia, 39*, 219-230.

Ochsner, K. N., & Lieberman, M. D. (2001). The emergence of social cognitive neuroscience. *American Psychologist, 56*, 717-734.

Ochsner, K. N. & Schacter, D. L. (2000). A social-cognitive neuroscience approach to emotion and memory. In J. C. Borod (Ed.), *The Neuropsychology of Emotion*. Oxford University Press: New York.

O'Reilly, R., C., Braver, T. S., & Cohen, J. D. (1999). A biologically based computational model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 375-411). New York, NY: Cambridge University Press.

O'Reilly, R. C., Munakata, Y. & McClelland, J. L. (2000). *Cognitive neuroscience: A computational exploration*. Cambridge, MA: MIT Press.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotion*. Cambridge, England: Cambridge University Press.

Osuch, E. A., Ketter, T. A., Kimbrell, T. A., George, M. S, Benson, B. E., Willis, M. W., Herscovitch, P., & Post, R. M. (2000). Regional cerebral metabolism associated with anxiety symptoms in affective disorder patients. *Biological Psychiatry, 48*, 1020-1023.

Otten, L. J., Henson, R. N. A., & Rugg, M. D. (2001). Depth of processing effects on neural correlates of memory encoding: Relationship between findings from across- and within-task comparisons. *Brain, 124*, 399-412.

Packard, M. G., Hirsh, R., & White, N. M. (1989). Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: Evidence for multiple memory systems. *Journal of Neuroscience, 9*, 1465-1472.

Perrett, D. I., Harries, M. H., Bevan, R., Thomas, S., Benson, P. J., Mistlin, A. J., Chitty, A. J., Hietanen, J. K., & Ortega, J. E. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology, 146*, 87-113.

Perrett, D. I., Hietanen, J. K., Oram, M. W., & Benson, P. J.(1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London (Series B), 335*, 23-30.

Perrett, D.I., Jellema, T., Frigerio, E., & Burt, M. (2001, April). *Using 'social attention' cues (where others are attending) to interpret actions, intentions and emotions of others*. Paper presented at the UCLA Conference on Social Cognitive Neuroscience, Los Angeles, CA.

Petty, R. E. & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In. L. Berkowitz (Ed.), *Advances in experimental social psychology.* (Vol. 19, pp. 123-205). New York: Academic Press.

Petrovic, P., Petersson, K. M., Ghatan, P. H., Stone-Elander, S., & Ingvar, M. (2000). Pain-related cerebral activation is altered by a distracting cognitive task. *Pain, 85*, 19-30.

Philips, R. G., & LeDoux, J. E. (1992). Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. *Behavioral Neuroscience, 106*, 274-285.

Poldrack, R. A., & Gabrieli, J. D. E. (2001). Characterizing the neural mechanisms of skill learning and repetition priming. Evidence from mirror-reading. *Brain, 124*, 67-82.

Portas, C. M., Strange, B. A., Friston, K. J., Dolan, R. J., & Frith, C. D. (2000). How does the brain sustain a visual percept? *Proceedings of the Royal Society of London (Series B), 267*, 845-850.

Posner, M. I. & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.

Posner, M. I. & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information Processing and Cognition: The Loyola Symposium* (pp. 550-585). Hillsdale, NJ: Erlbaum.

Quattrone, G. A. (1982). Overattribution and unit formation: When behavior engulfs the person. *Journal of Personality and Social Psychology, 42*, 593-607.

Rainville, P., Duncan, G. H., Price, D. D., Carrier, B., & Bushnell, M. D. (1997). Pain affect encoded in human anterior cingulate but not somatosensory cortex. *Science, 277*, 968-971.

Rauch, S. L., Savage, C. R., Brown, H. D., Curran, T., Alpert, N. M., Kendrick, A., Fischman, A. J. & Kosslyn, S. M. (1995). A PET investigation of implicit and explicit sequence learning. *Human Brain Mapping, 3*, 271-286.

Read, S. J. & Marcus-Newhall, A. R. (1993). Explanatory coherence in social explanations: A parallel dstributed processing account. *Journal of Personality and Social Psychology, 65*, 429-447.

Read, S. J., Vanman, E. J., & Miller, L. C. (1997). Connectionism, parallel constraint satisfaction processes, and gestalt principles: (Re)Introducing cognitive dynamics to social psychology. *Personality and Social Psychology Review, 1*, 26-53.

Reeder, G. D. (1985). Implicit relations between dispositions and behaviors: Effects on dispositional attribution. In J. H. Harvey & G. Weary (Eds.), *Attribution: Basic issues and application* (pp. 87-116). New York: Academic Press.

Reeder, G. D. (1993). Trait-behavior relations and dispositional inference. *Personality and Social Psychology Bulletin, 19*, 586-593.

Reeder, G.D., & Brewer, M.B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86*, 61-79.

Reeder, G. D., & Fulks, J. L. (1980). When actions speak louder than words: Implicational schemata and the attribution of ability. *Journal of Experimental Social Psychology, 16*, 33-46.

Reeder, G. D. & Spores, J. M. (1983). The attribution of morality. *Journal of Personality and Social psychology, 44*, 736-745.

Rolls, E. T. (1999). *The brain and emotion*. New York: Oxford University Press.

Rolls, E. T., & Treves, A. (1998). *Neural networks and brain function.* New York: Oxford University Press.

Rolls, E. T., Judge, S. J., & Sanghera, M. K. (1977). Activity of neurones in the inferotemporal cortex of the alert monkey. *Brain Research, 130*, 229-238.

Rosen, S. D., Paulesu, E., Nihoyannopoulos, P., Tousoulis, D., Frackowiak, R. S., Frith, C. D., Jones, T., and Camici, P. G. (1996). Silent ischemia as a central problem: regional brain activation compared in silent and painful myocardial ischemia. *Annals of Internal Medicine, 124*, 939-949.

Ross, L. (1977). The intuitive psychologist and his shortcomings. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 173-220). New York: Academic Press.

Rumelhart, D. E., & McClelland, J. (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.

Sartre, J. P. (1937). *Transcendence of the ego*. New York: Hill & Wang.

Sawamoto, N., Honda, M., Okada, T., Hanakawa, T., Kanda, M., Fukuyama, H., Konishi, J., & Shibasaki, H. (2000). Expectation of pain enhances responses to nonpainful somatosensory stimulation in the anterior cingulate cortex and parietal operculum/posterior insula: an event-related functional magnetic resonance imaging study. *Journal of Neuroscience, 20*, 7438-7445.

Schachter, S., & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*, 379-399.

Schacter, D. L. (1992). Understanding implicit memory: A cognitive neuroscience approach. *American Psychologist, 47*, 559-569.

Schneider, W., & Shiffrin, R. M. (1977). Contorlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1-66.

Schooler, J. W. (in press). Discovering memories in the light of meta-consciousness. *The Journal of Aggression, Maltreatment, and Trauma*.

Schrodinger, E. (1967/1992). *What is life? with Mind and matter*. New York: Cambridge University Press.

Sheinberg, D. L., & Logothetis, N. K. (1997). Noticing familiar objects in real world scenes: The role of temporal cortical neurons in natural vision. *Journal of Neuroscience, 21*, 1340-1350.

Sherman, S. J. (1980). On the self-erasing nature of errors of prediction. *Journal of Personality and Social Psychology, 39*, 211—221.

Shin, L. M., Dougherty, D. D., Orr, S. P., Pitman, R. K., Lasko, M., Macklin, M. L., Alpert, N. M., Fischman, A. J., & Rauch, S. L. (2000). Activation of anterior paralimbic structures during guilt-related script-driven imagery. *Biological Psychiatry, 48*, 43-50.

Shoda, Y., & Mischel, W. (1993). Cognitive social approach to dispositional inferences: What if the perceiver is a cognitive social theorist? *Personality and Social Psychology Bulletin, 19*, 574-595.

Shultz, T. R. & Lepper, M. R. (1995). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review, 103*, 219-240.

Sierra, M., & Berrios, G. E. (1998). Depersonalization: neurobiological perspectives. *Biological Psychiatry, 44*, 898-908.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition, 52*, 1-21.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3-22.

Smith, E. E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science, 283*, 1657-1661.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998) Alternative strategies of categorization. *Cognition, 65*, 167-196.

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology, 70*, 893-912.

Smith, E. R. & DeCoster, J. (1999). Associative and rule-based processing: A connectionist interpretation of dual-process models. In S. Chaiken & Y. Tropez (Eds.), *Dual-process theories in social psychology* (pp. 323-336). New York, NY: The Guilford Press.

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

Soufer, R., Bremner, J. D., Arrighi, J. A. Cohen, I., Zaret, B. L., Burg, M. M., Goldman-Rakic, P. (1998). Cerebral cortical hyperactivation in response to mental stress in patients with coronary artery disease. *Proceedings of the National Academy of Sciences, 95*, 6454-6459.

Spellman, B. A., & Holyoak, K. J. (1992). If Saddam is Hitler then who is George Bush? Analogical mapping between systems of social roles. *Journal of Personality and Social Psychology, 62*, 913-933.

Squire, L. R., & Knowlton, B. J. (2000). The medial temporal lobe, the hippocampus, and the memory systems of the brain. In M. Gazzaniga (Ed.), *The new cognitive neurosciencs* (2nd ed., pp. 765-779). Cambridge, MA: MIT Press.

Steele, C. M., & Josephs, R. A. (1990). Alcohol myopia: Its prized and dangerous effects. *American Psychologist, 45*, 921-933.

Suzuki, W., Saleem, K. S., & Tanaka, K. (2000). Divergent backward projections from the anterior part of the inferotemporal cortex (area TE) in the macaque. *Journal of Comparative Neurology, 422*, 206-228.

Taylor, S. E., & Fiske, S. T., (1978). Salience, attention, and attribution: Top of the head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 249-288). New York: Academic Press.

Tetlock, P. E. (1985). Accountability: A social check on the fundamental attribution error. *Social Psychology Quarterly, 48*, 227-236.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520-522.

Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I., & Miyashita, Y. (1999). Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature, 401*, 699-701.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review, 93*, 239-257.

Trope, Y., & Alfieri, T. (1997). Effortfulness and flexibility of dispositional inference processes. *Journal of Personality and Social Psychology, 73*, 662-675.

Trope, Y., & Cohen, O. (1989). Perceptual and inferential determinants of behavior-correspondent attributions. *Journal of Experimental Social Psychology, 25*, 142-158.

Trope, Y. & Gaunt, R. (1999). A dual-process model of overconfident attributional inferences. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 161-178). New York: Guilford Press.

Trope, Y., & Gaunt, R. (2000). Processing alternative explanations of behavior: Correction or integration? *Journal of Personality and Social Psychology, 79*, 344-354.

Trope, Y., Cohen, O., & Maoz, Y. (1988). The perceptual and inferential effects of situational inducements on dispositional attributions. *Journal of Personality and Social Psychology, 55*, 165-177.

Trope, Y., & Liberman, A. (1993). Trait conceptions in identification of behavior and inferences about persons. *Personality and Social Psychology Bulletin, 19*, 553-562.

Trope, Y., & Liberman, A. (1996). Social hypothesis-testing: Cognitive and motivational mechanisms. In E.T. Higgins & A.W. Kruglanski (Eds.), *Social Psychology: Handbook of Basic Principles*. NY: Guilford Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293-315.

Tyler, L. K., & Moss, H. E. (1998). Going, going, gone…? Implicit and explicit tests of conceptual knowledge in a longitudinal study of semantic dementia. *Neuropsychologica, 36*, 1313-1323.

Tzelgov, J. (1997). Specifying the relations between automaticity and consciousness: a theoretical note. *Consciousness and Cognition, 6*, 441-451.

Uleman, J. S., Newman, L. S., & Moskowitz, G. B. (1996). People as flexible interpreters: Evidence and issues from spontaneous trait inference. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 29, pp. 211-279). San Diego: Academic Press.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *European Journal of Neuroscience, 11*, 1239-1255.

Wagner, A. D., Schacter, D. L., Rotte, M., Kootstaal, W., Maril, A., Dale, A. M., Rosen, B. R., & Buckner, R. L. (1998). Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science, 281*, 1188-1191.

Waldmann, M. R., & Holyoak, K. J. (1992) Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*, 222-236.

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., Santos, M. M., Thomas, C. R., & Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science, 10*, 119-125.

Wang, G., Tanaka, K., & Tanifuji, M. (1996). Optical imaging of functional organization in the monkey inferotemporal cortex. *Science, 272*, 1665-1668.

Webster, D. M. (1993). Motivated augmentation and reduction of the overattribution bias. *Journal of Personality and Social Psychology, 65,* 261-271.

Weiskrantz, L., & Saunders, R. C. (1984). Impairments of visual object transforms in monkeys. *Brain, 107*, 1033-1072.

Wharton, C. M., & Grafman, J. (1998). Deductive reasoning and the brain. *Trends in Cognitive Sciences, 2*, 54-59.

Whitehead,A.N. (1911). *An introduction to mathematics*. London: Williams and

Norgate.

Wilshire, B. (1982). *Role playing and identity: The limits of theatre as metaphor*. Bloomington, IN: Indiana University Press.

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology, 47*, 237-252.

Winter, L., Uleman, J. S., & Cuniff, C. (1985). How automatic are social judgments? *Journal of Personality and Social Psychology, 49*, 904-917.

Zimbardo, P. B. (1969). The human choice: Individuation, reason, and order versus deindividuation, impulse and chaos. In W. J. Arnold & D. Levine (Eds.), *Nebraska symposium on motivation* (pp. 237-307). Lincoln, NE: University of Nebraska Press.