

## A Recurrent Connectionist Model of Person Impression Formation

**Frank Van Overwalle**

*Department of Psychology  
Vrije Universiteit Brussel, Belgium*

**Christophe Labiouse**

*Belgian NFSR Research Fellow & Department of Psychology  
University of Liège, Belgium*

*Major findings in impression formation are reviewed and modeled from a connectionist perspective. The findings are in the areas of primacy and recency in impression formation, asymmetric diagnosticity of ability- and morality-related traits, increased recall for trait-inconsistent information, assimilation and contrast in priming, and discounting of trait inferences by situational information. The majority of these phenomena are illustrated with well-known experiments and simulated with an autoassociative network architecture with linear activation update and using the delta learning algorithm for adjusting the connection weights. All of the simulations successfully reproduced the empirical findings. Moreover, the proposed model is shown to be consistent with earlier algebraic models of impression formation (Anderson, 1981; Busemeyer, 1991; Hogarth & Einhorn, 1992). The discussion centers on how our model compares to other connectionist approaches to impression formation and how it may contribute to a more parsimonious and unified theory of person perception.*

Getting to know others socially often involves inferences about characteristics and traits of individuals. This is a crucial to social reasoning, as it allows one to go beyond the specific behavior of the individual and to generalize to similar events in the future and to similar people. How do we make trait inferences from observing an individual's behavior? How is this information processed and stored in memory? The purpose of this article is to attempt to gain insight into this process by taking a computational modeling perspective—in particular, a connectionist approach—to these questions.

In the recent past, the dominant view on the impression-formation process in social psychology was that people are intuitive statisticians who extract behavioral information by applying some kind of rule to them. Many of these models rely on an algebraic function to transform social information into abstract traits. Models of this type include the weighted averaging model of Anderson (1981), the step-by-step belief-adjustment model of Hogarth and Einhorn (1992), and the serial averaging strategy of Busemeyer (1991). Although supported by an impressive amount of empiri-

cal research, the most popular model, developed by Anderson, was criticized on the grounds that it lacked psychological plausibility, because it seems unlikely that people would perform all the necessary weighting and averaging calculation in their minds to arrive at an impression, and, for this reason, many researchers abandoned algebraic models altogether.

The main purpose of this article is to revive computational modeling of impression formation by adopting a connectionist framework to describe the perceivers' internal computations. This framework is loosely patterned after how the brain works—neurons spreading activation to other neurons and developing synaptic connections with each other—to develop computational models of human thinking. Earlier attempts of applying the brain metaphor in social psychology used neuron-like representation and activation spreading as its major principles (Hamilton, Katz, & Leirer, 1980; Hastie & Kumar, 1979) but did not provide any computational formalization of these ideas. However, the ever-increasing success of connectionist models in cognitive psychology that provide more precise computational implementations has led a number of authors to turn to these models in an attempt to develop connectionist models of diverse social psychological phenomena, including causal attribution (Read & Montoya, 1999; Van Overwalle, 1998), cognitive dissonance (Shultz & Lepper, 1996; Van Overwalle & Jordens, 2002), and group impression formation and

---

This research was supported by Grant OZR423 of the Vrije Universiteit Brussel to Frank Van Overwalle. We are grateful to Bob French for his suggestions and comments on an earlier version of this article.

Requests for reprints should be sent to Frank Van Overwalle, Department of Psychology, Vrije Universiteit Brussel, Pleinlaan 2, B-1050, Brussel, Belgium. E-mail: Frank.VanOverwalle@vub.ac.be

change (Kashima, Woolcock, & Kashima, 2000; Van Rooy, Van Overwalle, Vanhoomissen, Labiouse, & French, 2003).

Connectionist models have also been developed in the area of impression formation. Kunda and Thagard (1996) developed a parallel-constraint-satisfaction model that described how social stereotypes and individuating information constrained each other's meaning and jointly influenced impression formation of individuals. However, this model was static, lacking a learning mechanism by which novel impressions or stereotypes of groups and individuals could be developed and stored in memory. This shortcoming was addressed in the tensor-product model of impression formation by Kashima and Kerekes (1994), from which emerged various results in primacy and recency effects. In addition to this model, the recurrent model by Smith and DeCoster (1998) described how perceivers may use past knowledge of individuals or groups to make inferences to unobserved or novel characteristics about them. Both models incorporate a learning algorithm that allows the integration of old and novel information and the subsequent storage of the resulting impression in memory. The goal of this article is to extend the work of Smith and DeCoster by highlighting how the recurrent model can explain many additional phenomena in impression-formation research. In so doing, we hope to contribute to the theoretical integration of the connectionist approach to impression formation. We also formulate "postdictions" that have been confirmed by past research and, importantly, novel predictions that have been subsequently confirmed by recent research.

Many mainstream processes and findings on impression formation can be explained within a connectionist framework and, in many cases, explained better than by the algebraic or activation-spreading models developed in the past. What are the main characteristics of the connectionist models that accomplish this?

First, connectionist models exhibit emergent properties, such as prototype extraction, pattern completion, generalization, constraint satisfaction, and graceful degradation. (All of these are extensively reviewed in Rumelhart & McClelland, 1986, and Smith, 1996). It is clear that these characteristics are potentially useful for any account of impression-formation phenomena (see Smith & DeCoster, 1998). In addition, connectionist models assume that the development of internal representations and the processing of these representations are done in parallel by simple and highly interconnected units, contrary to traditional models in which the processing is inherently sequential. As a result, connectionist systems have no need for a central executive, thereby eliminating the requirement of explicit (central) processing of relevant social information, as assumed by previous theories. Consequently, information can, in principle, be processed in an implicit and automatic manner without recourse to

explicit conscious reasoning. This does not, of course, preclude people's being aware of the outcome of these preconscious processes.

Second, most neural networks (e.g., Kashima & Kerekes, 1994; Smith & DeCoster, 1998) are not fixed models but are able to learn over time, usually by means of a simple learning algorithm that progressively modifies the strength of the connections between the units making up the network. The fact that most traditional models in social psychology are incapable of learning is a significant restriction. Interestingly, the ability to learn incrementally puts connectionist models in broad agreement with developmental and evolutionary constraints.

Third, connectionist networks have a degree of neurological plausibility that is generally absent in previous statistical approaches to information integration and storage (e.g., Anderson, 1981; Busemeyer, 1991; Hogarth & Einhorn, 1992). Although it is true that connectionist models are highly simplified versions of real neural circuitry and processing, it is commonly assumed that they reveal a number of emergent processing properties that real human brains also exhibit. One of these emergent properties is the integration of long-term memory (i.e., connection weights), short-term memory (i.e., internal activation), and outside information (i.e., external activation). In short, there is no clear separation between memory and processing as there is in traditional models. Even if biological constraints are not strictly adhered to in connectionist models of social judgments, there is currently an outpouring of interest in biological implementation of social inference mechanisms (Adolphs & Damasio, 2001; Allison, Puce, & McCarthy, 2000; Cacioppo, Berntson, Sheridan, & McClintock, 2000; Ito & Cacioppo, 2001; Phelps et al., 2000) that parallel the increasing attention paid to neurophysiological determinants of social behavior. Other emergent properties of the connectionist approach are explained in more depth in the next section.

This article is organized as follows: First, we describe the proposed connectionist model in some detail, giving the precise architecture, the general learning algorithm, and the specific details of how the model processes information. In addition, a number of other less well-known emergent properties of this type of network are discussed. We then present a series of simulations using the same network architecture applied to a number of significantly different phenomena. These phenomena involve primacy and recency in impression formation, the asymmetric impact of ability- versus morality-related behaviors, memory advantages for inconsistencies, assimilation and contrasts in priming, and the effect of situational constraints on trait inferences.

Our review of empirical phenomena in the field is not meant to be exhaustive but is rather designed to illustrate how connectionist principles can be used to shed light on the processes underlying impression for-

mation. Although the emphasis of this article is on the use of a particular connectionist model to explain a wide variety of phenomena in social cognition, previous applications of connectionist modeling to social psychology (Kashima & Kerekes, 1994; Kunda & Thagard, 1996; Smith & DeCoster, 1998) are also mentioned. In addition, we compare and contrast our model with a number of different models. Finally, we discuss the limitations of the proposed connectionist approach and point to areas where further theoretical developments are underway or are needed.

### A Recurrent Model

Throughout this article, we use the same basic network model—namely, the recurrent linear autoassociator developed by McClelland and Rumelhart (1985; for an introductory text see McClelland & Rumelhart, 1988, pp. 161ff.; McLeod, Plunkett, & Rolls, 1998, pp. 72ff). This model is already familiar to a number of social psychologists studying person and group impression (Queller & Smith, 2002; Smith & DeCoster, 1998; Van Rooy et al., 2003), causal attribution (Read & Montoya, 1999), and attitude formation (Van Overwalle & Siebler, 2002). We decided to apply a single basic model to emphasize the theoretical similarities that underlie a broad variety of processes in person impression. In particular, we chose this model because it is capable of reproducing a wider range of phenomena than other connectionist models, such as feedforward networks (e.g., Van Overwalle, 1998; Van Overwalle & Jordens, 2002), parallel-constraint-satisfaction models (e.g., Kunda & Thagard, 1996), or tensor-product models (e.g., Kashima & Kerekes, 1994; Kashima et al., 2000).

We believe that one of the strengths of our approach is that—despite the great flexibility of connectionist models, which is sometimes seen as rendering them theoretically empty because, as models of human cognition, they are too powerful—we actually make little use of that flexibility here. As we demonstrate shortly, the parameters of our model are not varied at will to fit each different situation. Rather, only the learning rate and the assumed sequence of learning inputs vary from problem to problem. This makes the network directly testable because, at least in principle, assumptions about sequence of input information and speed of learning (e.g., depth of encoding) can be tested empirically. This stands in sharp contrast to parallel-constraint-satisfaction models, in which the key assumptions are not about potentially observable inputs but about unobservable internal structure, such as connection weights that are set by hand by the researchers but are not directly testable.

The autoassociative network can be distinguished from other connectionist models on the basis of its architecture (the elements of the model) and how the in-

formation is processed and consolidated in memory (the activation updates and the learning algorithm). We discuss these points in turn.

### Architecture

The architecture of the linear autoassociative network used in this article is illustrated in Figure 1. Its most salient property is that all nodes are interconnected with all of the other nodes. Thus, all nodes exchange (send out and receive) activation with each other, but not with themselves (i.e., there are no self-connections).

### Information Processing

In this type of recurrent network, processing information takes place in two phases. During the first activation phase, each node in the network receives activation from the environment. Because the nodes are interconnected, this activation is spread throughout the network in proportion to the weights of the connections to the other nodes. The activation coming from the other nodes is called the internal input (for each node, it is calculated by summing all activations arriving at that node). This activation is further updated during a number of cycles through the network. Together with the external input, this internal input determines the final pattern of activation of the nodes, which reflects the short-term memory of the network. Typically, activations from external sources are bounded between  $-1$  and  $+1$ , although the activation levels within the network may grow beyond these bounds (e.g., because the external activation is augmented with the internal input). In addition, the initial weights are also bounded between  $-1$  and  $1$  and, like the activations, are allowed to grow beyond these bounds.

In the linear version of activation spreading in the autoassociator that we use here, the final activation at each cycle is the linear sum of the external and internal

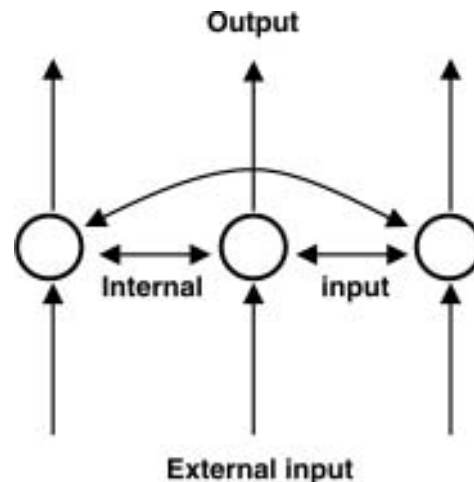


Figure 1. Generic architecture of an autoassociative recurrent network.

input. In nonlinear versions used by other social-psychology researchers (Read & Montoya, 1999; Smith & DeCoster, 1998), the final activation is determined by a nonlinear combination of external and internal input (typically, a sigmoid function). During our simulations, however, we found that the linear version with a single internal updating cycle often reproduced the observed data slightly better for reasons that we discuss later. Therefore, we used the linear variant of the autoassociator for all the reported simulations.

### Consolidation in Memory

After the first activation phase, the recurrent model enters the second learning phase in which the short-term activations are consolidated in long-term weight changes of the connections. Basically, these weight changes are driven by the error between the internal activation generated by the network and the external input received from outside sources. This error is reduced in proportion to a learning rate that determines how fast the network changes its weights (between .10 and .35 for all simulations reported). This error-reducing mechanism is known as the *delta learning algorithm* (McClelland & Rumelhart, 1988).

Thus, when the network overestimates the external input of a node, this means that this node received too much internal input from the other nodes through their connections. To adjust this, the delta algorithm decreases the weights of these connections. Conversely, when the network underestimates the external input, this means that it has received too little internal input and the weights are increased. These weight changes allow the network to better approximate the external input. Thus, the delta algorithm strives to match the internal predictions of the network as closely as possible to the actual state of the external environment and stores this information in the connection weights (see Appendix A for more details).

### Basic Emergent Connectionist Principles

Before moving on to the social phenomena of interest, it is essential to briefly discuss the basic principles or mechanisms that drive many of our simulations. These principles are emergent properties of the delta-learning algorithm and include acquisition, competition, and diffusion. Some of these principles have already been documented in prior social connectionist work (Van Overwalle, 1998; Van Overwalle & Van Rooy, 1998, 2001, 2003; Van Rooy et al., 2003). However, because they are essential for understanding our examples, we describe these principles first and discuss their application for impression formation in more detail later during the simulations.

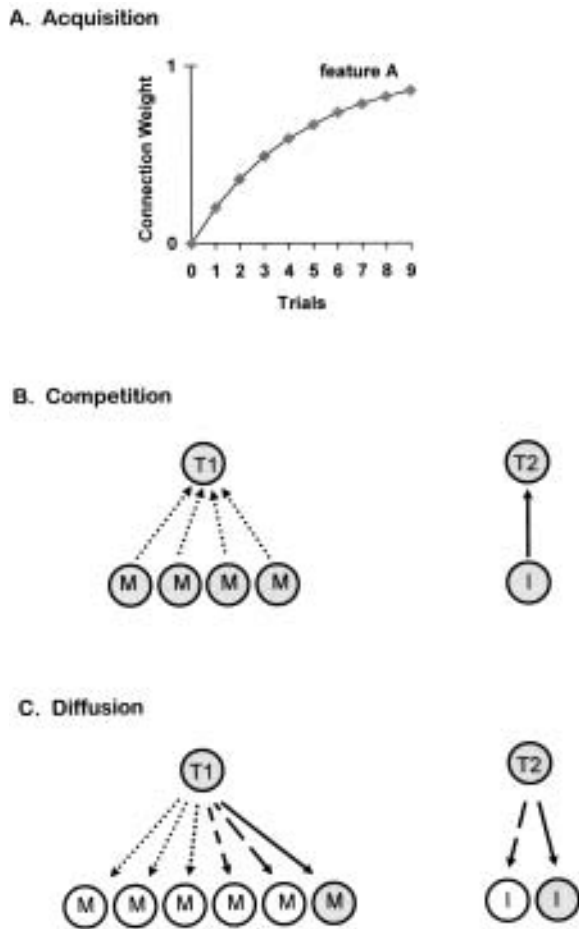
### Acquisition Property and Sample Size Effect

The acquisition property involves sample size effects that have been documented in many areas of social reasoning. For instance, when receiving more supportive information, people tend to hold more extreme impressions about other persons (Anderson, 1976, 1981), make more extreme causal judgments (Baker, Barbier, & Vallée-Tourangeau, 1989; Försterling, 1992; Shanks, 1985, 1987, 1995; Shanks, Lopez, Darby, & Dickinson, 1996), make more polarized group decisions (Ebbesen & Bowers, 1974; Fiedler, 1996), endorse hypotheses more firmly (Fiedler, Walther, & Nickel, 1999), make more extreme predictions (Manis, Dovalina, Avis, & Cardoze, 1980), and agree more with persuasive messages (Eagly & Chaiken, 1993).

One of the most striking characteristics of connectionist models using the delta algorithm is that learning is modeled as a gradual online process of adjusting existing knowledge to novel information. This characteristic has already been exploited in the earlier associative learning models that preceded connectionism, such as the popular Rescorla–Wagner (Rescorla & Wagner, 1972) model of animal conditioning and human contingency judgments.

The Rescorla–Wagner (Rescorla & Wagner, 1972) model predicts that when a cue (i.e., conditioned stimulus) is followed by an effect (i.e., unconditioned stimulus), the organism incorporates this information, thereby resulting in a stronger cue–effect association and a more vigorous response when the cue is present. In humans, this also results in stronger judgments of the causal influence of the cue (see Baker et al. 1989; Shanks, 1985, 1987, 1995; Shanks et al., 1996; Van Overwalle & Van Rooy, 2001b). Likewise, the delta algorithm predicts that the more information that is received about the joint presence of an actor or stimulus and a trait category, the stronger their connection weight will become. This results in a pattern of increasing weights as more information is processed, that is, a sample size effect (see illustration in Figure 2A). Most earlier algebraic models of impression formation also predict this gradual increase of the strength of judgments (Anderson, 1981; Bussemeyer, 1991; Hogarth & Einhorn, 1992).

How are online learning and a sample size effect achieved in connectionist models? Given that connection weights are initially set to zero (or any arbitrary low-scale value), the effect is that in the initial phases of learning, the connection weights are relatively small and often inaccurate, and gradually grow more accurate (stronger or weaker, positive or negative) as more information is received. The reason for this incremental learning is that the error in the delta algorithm is only gradually minimized as a function of the learning rate. Even when the covariation between a feature and



**Figure 2.** Graphical illustration of the principles of (A) acquisition (with learning rate 0.20), (B) competition, and (C) diffusion. T = trait; M = multiple features; I = infrequent features. Filled nodes are activated at a single trial; empty nodes are not activated. Full lines denote strong connection weights; broken lines denote moderate weights; dotted lines denote weak weights.

a category is perfect, the learning rate dictates that the weights connecting the two will increase by only a small fraction. Thus, it takes multiple repetitions of the same information before a strong weight corresponding to this covariance emerges.

Figure 2A, for example, depicts a system with a learning rate of 0.20. Feature A is always paired with a particular category (i.e., perfect correlation). Assuming that the connection weight associated with feature A is initialized to 0, a learning rate of 0.20 means that the initial error of underestimating this correlation is gradually corrected by increasing the connection weight with 20% of the error. As a consequence, the weight will gradually increase at each trial starting with 0.20 at the first trial to eventually reach a maximum value of +1 after a number of trials. Note that parallel-constraint-satisfaction models (Kunda & Thagard, 1996; Read & Marcus-Newhall, 1993; Shultz & Lepper, 1996) do not possess a learning algorithm and are therefore not able to make this obvious prediction.

It has been noted that, given a sufficient number of trials, the delta learning algorithm converges to the same predictions as conventional rule-based algebraic models of causal reasoning (Chapman & Robbins, 1990; Sarle, 1994; Van Overwalle, 1996). More important, using the same logic, it can be demonstrated that the delta algorithm also converges to Anderson's (1981) weighted averaging model (see Appendix B). That is, the delta algorithm predicts that in the initial phases of learning, a person impression gradually becomes stronger, as if the information is summed (e.g., Betsch, Plessner, Schwieren, & Gütig, 2001), but after more information, the impression is characterized as a weighted average of the information. In addition, it is easy to show that other algebraic models (Busemeyer, 1991; Hogarth & Einhorn, 1992) are mathematically identical to a simplified version of the delta algorithm applied in connectionist models, one that deals with only one cause at the time (for a proof, see Van Overwalle & Van Rooy, 2001b; Wasserman, Kao, Van Hamme, Katagiri, & Young, 1996). Because these models can be considered to be special cases of the more general delta algorithm, only with less computational power, we ignore them for the most part in the remainder of this article.

### Competition Property and Discounting

Another essential property of the delta algorithm is that it gives rise to competition between connections. This competition favors the more predictive or diagnostic features. The term *competition* stems from the associative learning literature on animal conditioning and causality judgments mentioned earlier (Rescorla & Wagner, 1972; Shanks, 1995) and should not be confused with other usages in the connectionist literature such as competitive networks (McClelland & Rumelhart, 1988). A typical example of competition is the phenomenon of discounting in causal attribution. When one cause acquires a strong causal weight, perceivers tend to ignore alternative causes (Hansen & Hall, 1985; Kruglanski, Schwartz, Maides, & Hamel, 1978; Rosenfield & Stephan, 1977; Van Overwalle & Van Rooy, 1998, 2001a; Wells & Ronis, 1982). In impression formation, information on the situational context in which a behavior occurred often (but not always) leads to discounting of trait inferences to the actor (Gilbert & Malone, 1995; Trope & Gaunt, 2000).

Competition arises naturally in associative learning models such as the Rescorla–Wagner model (Rescorla & Wagner, 1972), where it is known as “blocking.” In fact, one of the reasons for the widespread popularity of the Rescorla–Wagner model is that it was among the first conditioning models capable of predicting this property. As noted by several researchers (Read & Montoya, 1999; Van Overwalle, 1998), the delta algorithm makes similar predictions. In contrast, another well-known connectionist learning algorithm, the

Hebbian algorithm, used in the tensor-product model of Kashima and Kerekes (1994) does not possess this property and is therefore not able to make straightforward discounting predictions.

How does this property work in connectionist models? Competition is driven by the connections linking multiple determinants (actors, situational circumstances, and so on) with a category node (e.g., implied trait; see upward arrows in Figure 2B). The key mechanism is that the activation of the trait node is determined by the sum of the activations received from all other determinants. The left panel in Figure 2B depicts multiple possible determinants of trait T1. Given these multiple activations, this leads to overactivation of trait T1 and increased (negative-valued) error in the delta algorithm, and therefore blocks or even decreases any further growth of the connection weights to T1. In contrast, in the right panel of Figure 2B, the connection of a single determinant with T2 can grow unhampered until it reaches asymptote, because there are no other activations that compete and may slow down acquisition.

### **Diffusion Property and Memory for Inconsistent Information**

Still another property of the delta algorithm is responsible for the weakening of connections when a single trait node is connected to many behavior nodes that are only occasionally activated. This property is introduced to explain enhanced recall for inconsistent as compared to consistent behavioral information in impression formation (Hastie & Kumar, 1979). Enhanced recall for inconsistent behavior has been traditionally explained in terms of a spreading-activation model of memory, in which inconsistent information is more deeply processed so that it develops stronger lateral connections with other behavior nodes (Hastie & Kumar, 1979). Other researchers argued that enhanced memory for unique information is due to a fan effect, whereby a given amount of activation is divided between the connections fanning out to other nodes. The more numerous the connections, the less activation each one gets (Anderson, 1976). However, the diffusion property is a fundamentally different mechanism. Whereas fan-out causes a division of activation, diffusion involves the division of weights. In the associative learning and connectionist literature, this is a novel property that—to our knowledge—was not detected or mentioned earlier.

How does this diffusion principle explain enhanced recall of inconsistent information? In contrast to the competition property, the diffusion effect is driven by trait→behavior connections (see downward arrows in Figure 2C). The basic mechanism is that during learning about behaviors performed by an actor, each node that reflects a specific behavior is activated only once together with the trait node and remains inactive while other behaviors are activated with the same trait node.

This long period of inactivation results in a weakening of the trait→behavior connections. This is not due to spontaneous decay of the connection weights. Rather, the mechanism is that each behavior node (except the last) is inactive at some moment in learning after having been active during previous learning. This inactivity of the behavior node is unexpected by the network and therefore leads to a weakening of the trait→behavior connections. This process continues for all the behavior nodes (except the last activated), resulting in an overall weakening of the trait→behavior connections.

To illustrate, after observing a first behavior, the trait→behavior connection gains some strength (by the acquisition property). When the second behavior is presented, the first behavior node is inactive while the trait node is still active, and consequently the strength of the first trait→behavior connection is reduced. This reduction continues for all the behaviors that follow (see left panel of Figure 2C for a schematic illustration), resulting in the weakest weight for the first connections. However, compared to many consistent behaviors that imply trait T1, inconsistent behaviors that imply the opposite trait T2 are, by definition, smaller in number. Hence, there is less often inactivation and thus weakening of inconsistent trait→behavior connections (with T2) than of consistent connections (with T1). This unequal weakening or diffusion therefore leads to better recall for inconsistent information.

## **Overview of the Simulations**

### **Simulated Phenomena**

We applied the three emergent connectionist processing principles to a number of classic findings in the social cognition literature. For explanatory purposes, we replicated a well-known experiment that illustrates a particular phenomenon. Table 1 lists the topics of the simulations to be reported and the relevant empirical data that we attempted to replicate, as well the major underlying processing principle responsible for producing the data in the simulation. Although not all relevant data in impression formation can be addressed in a single article, we are confident that we have included some of the most relevant phenomena in the current literature.

Essentially the same methodology was used throughout the simulations. The particular conditions and trial orders of the focused experiments were reproduced as faithfully as possible, although sometimes minor changes were introduced to simplify things (e.g., fewer trials than in the actual experiments). For each simulation, the autoassociative network was run 50 times (i.e., simulating 50 participants) with a randomized or fixed trial order (as in the real experiment), and the results were then averaged over the 50 runs.

**Table 1.** Overview of the Simulated Person Impression Topics and the Underlying Properties

Topic	Findings	Property
Person Impression Formation		
1. Online Integration	More extreme judgments given more evidence on positive or negative features	<i>Acquisition</i> of actor→trait weights
2. Serial Position Weights	The last item in a series has most impact on trait inferences (recency). In addition: <ul style="list-style-type: none"> <li>• Recency attenuates after longer lists of items</li> <li>• Primacy if only a final trait inference is given</li> </ul>	<i>Competition</i> : Stronger context discounts impact of inconsistent item on trait <i>Acquisition</i> of trait→actor link sends inconsistent trait activation to actor and so reduces learning
Inferring Behavior–Correspondent Traits		
3. Asymmetric Cues	High ability and low morality behaviors are more diagnostic for an actor’s traits than low ability and high morality	<i>Acquisition</i> : Skewed distribution of ability and morality behaviors
Recall of Behavioral Information		
4. Inconsistent Behavior	Recall is better for trait-inconsistent behaviors	Less <i>diffusion</i> of infrequent trait→behavior links
Priming		
5. Assimilation and Contrast	Priming with <ul style="list-style-type: none"> <li>• a trait leads to assimilation of that trait</li> <li>• an exemplar leads to contrast away from the implied trait</li> </ul>	<i>Acquisition</i> : Additional trait activation is linked to the actor <i>Competition</i> : Exemplar→trait link competes with actor→trait link
Discounting by Situational Constraints		
6. Integration of Situational Information	Discounting of a trait given situational information, especially if this information is more salient or applicable	<i>Competition</i> : Discounting of actor→trait link given a stronger situation→trait link
7. Discounting and Sample Size	Discounting of an actor’s trait when there is more evidence on an alternative actor	<i>Acquisition</i> of alternative→trait link which leads to <i>competition</i> against target→trait link

### Architecture of the Network

The concepts of interest in the simulations such as actors, traits, behaviors, and so on are each represented by a single node. This is a localist encoding, whereby each node reflects a “symbolic” concept. In contrast, in a distributed encoding, as used by Kashima and Kerekes (1994), and Smith and DeCoster (1998), a concept is represented by a pattern of activation across an array of nodes, none of which reflect a symbolic concept but rather some subsymbolic micro-feature of it (Thorpe, 1994). We acknowledge that localist encoding lacks biological plausibility, because it implies that each concept is stored in a single processing unit and, except for explicit differing levels of activation, is always perceived in the same manner by the network. However, this localist coding scheme was chosen as a simplifying assumption in our attempt to demonstrate the power of our model. We show at the end of this article that when distributed representations are used, they yield approximately the same results.

**Representation of traits.** We assume that behaviors or actors are naturally categorized by at least one of two opposing trait categories (i.e., represented by two nodes given localist encoding). For instance, someone’s performance may be characterized as “stupid” or “intelligent.” More fine-grained categorizations, such as different levels of intelligence, are also possible. However,

perceivers in an experimental setting are sometimes forced to make one-dimensional judgments, for instance, when the experimental instructions call for participants to make ratings on a single intelligence or likeability scale. When such demands dominate the impression task—especially when participants have to give unidimensional ratings repeatedly—we assume that perceivers are likely to represent their judgment along a single integrative conceptual category (i.e., represented by a single node given localist encoding). This single unidimensional representation is used only in the first two simulations discussed shortly, but it is critical to reproduce some of the phenomena of interest.

How might such a novel unitary trait concept be represented in human memory? Research in neuropsychology has revealed that traces of novel episodic events or concepts are stored in the hippocampus, and, recently, connectionist modelers have begun to model these processes (e.g., O’Reilly & Munakata, 2000, pp. 287–293; O’Reilly & Rudy, 2001). The basic idea is that novel information or concepts consist of a unique combination of existing features, and this configuration is temporarily stored in a hippocampal layer by representing each unique event or concept by an internal representation (e.g., a limited set of nodes) that is connected with its constituting features. However, because such a detailed network is beyond the scope of this article, in our simulations, we simply added a single

node to represent an integrative one-dimensional representation of a trait. To elucidate the most basic learning principles, we made use only of this integrative trait node and ignored the two opposing trait categories (i.e., represented by two nodes in a localist encoding). Although it is very likely that these opposing trait categories continued to play a role in learning because they seem most natural, adding them in the simulation did not meaningfully alter the results (if anything, it slightly improved the fit with the observed data).

**Context nodes.** In practically all simulations that involved a judgment about an actor, the target actor was accompanied by a general context or other comparison persons. This serves as a standard of comparison, to judge the level of a trait that an actor possesses. This is a crucial process in social cognition. Because there are no objective standards for judging people's behaviors and opinions, perceivers need another, social standard of comparison in their judgments. This idea was perhaps best elaborated in attribution theory (Kelley, 1967). An abundance of research has documented that attributions to an actor depend on low consensus, or the degree to which the behavior of the person differs from that of other comparison persons. If the behavior is similar, we do not make attributions to the person but rather to some external circumstances or context (Kelley, 1967; Van Overwalle, 1997). Other persons thus provide a standard of comparison, often internalized as norms for different target categories (e.g., sex, social groups, and so on). Comparisons can be made against general norms or specific comparison persons, depending on task instructions and the availability of specific individuals to compare with. In other words, including a context is necessary to determine whether the actor is responsible for his or her behavior and is thus required for making trait inferences. Indicating how often a behavior occurs in general allows us to establish what the relevant social norms or standards are. How much the actor's behavior deviates from this norm is indicative of the actor's underlying trait.

What connectionist mechanism allows a context to serve as a standard of comparison in trait inferences? The underlying mechanism is the principle of competition. If the context develops a strong connection with a trait because it is paired equally (or more) often with a trait-implicating behavior than with the actor, it tends to compete against the actor→trait connections, leading to weaker trait inferences. In contrast, when the context develops a weaker connection with the trait because it is less often paired with the behavior than with the actor, it cannot discount the actor→trait connection, leading to stronger trait inferences.

Competition by the context plays a role in all our simulations of trait judgments. We used a variety of contextual factors. In some experiments, the context was implicit (e.g., only trait adjectives described the

person), so that we used a general context node without specifying in great detail what it represented. In these cases, the context might reflect a variety of aspects from task instructions to (unspecified) cues in the actor's behavioral environment (Simulations 1 and 2). In other experiments, the context was explicitly manipulated and was defined by specific actor exemplars with detailed meanings or behaviors, which we used in Simulations 5, 6, and 7.

### Activation and Learning Parameters

All parameters of the autoassociative model that influence the spreading of activation were kept fixed for all simulations (cf. McClelland & Rumelhart, 1988; see Appendix A for the technical details of our simulations). In contrast, we did not impose a common learning rate for all the experiments simulated because of the different contexts, measures, and procedures used in them. Rather, we selected a learning rate value that provided the highest correlation with the observed data of each simulation, after examining all admissible parameter values (see Gluck & Bower, 1988; Nosofsky, Kruschke, & McKinley, 1992).

Variations in the learning rate are assumed to arise from differences in attention to the task. These differences can originate from modulations in basic arousal and behavioral activation (e.g., sleep–wake cycle); responsiveness to novel, affect-laden, motivation-relevant, or otherwise salient stimuli; or responsiveness to task-specific attentional focus and voluntary control of exploring, scanning, and encoding information. Research on the neurological underpinnings of attention suggest that general arousal is driven by lower level nuclei and pathways from the brain stem, whereas basic features of the stimuli are detected by the thalamus and related subcortical nuclei (e.g., amygdala, basal ganglia). In contrast, task-specific attention and voluntary control is most likely modulated by supervisory executive centers in the prefrontal neocortex (LaBerge, 1997, 2000; Posner, 1992). We felt that this wide range of sources of attention justified our allowing the learning rate to vary from simulation to simulation.

Further, in several simulations, variations in learning rate stem predominantly from conscious control over one's attentional focus. For instance, instructing participants to memorize behavioral details or by giving them an additional cognitive task (Simulations 4 and 7) causes them to turn their attention away from trait inferences. Some connectionist researchers have begun to model these higher level voluntary control and attention processes (e.g., O'Reilly & Munakata, 2000, pp. 305–312, 379–410). The basic idea of their approach is that activation plays a major role in maintaining and switching attention (through dopamine-based modulation), resulting in greater accessibility and impact of the internal representations. When

information is actively maintained, it is immediately accessible to other parts of the system and constantly influences the activation of other representations. It is assumed that task instructions directly impact on the activation of internal representations.

Because such central executive “subnetworks” are beyond the scope of this article, to simulate reduced attention to the target task, we simply hand-set overall learning to a lower rate to reproduce the idea that encoding and learning was hampered and more shallow. For instance, when instructed to memorize behaviors or by giving participants a secondary task, we assumed that they would be less attentive to making trait inferences, so that the rate of development of trait-relevant connections is diminished. Although we could have manipulated the activation of the input nodes rather than the speed of learning to simulate reduced attention (which gives similar results), for reason of parsimony in the manipulation of parameters we varied only the learning rate (for a similar approach, see Kinder & Shanks, 2001).

In some cases, apart from a general learning rate, the context node received a separate learning rate for all context→trait connections. Because, as noted earlier, the context was not always sharply defined in the experiments on which the simulations were based, it was unclear what constituted its representation. For instance, it might consist of fewer or more relevant features than other representations in the network, and this might lower or increase, respectively, the speed of learning about the context. Nor was it clear how much attention participants would pay to contextual features as compared to other information. Rather than making ad hoc arbitrary assumptions about the content or encoding of contextual features, for reasons of consistency with our overall simulation approach we estimated a separate learning rate for the context that fitted most closely with the observed data.

Varying the learning rate as described previously does not violate the locality principle of connectionism, which says that each connection weight should be able to be updated using locally available information from associated nodes. That is because the learning rate only affects the general speed of learning in the network, not how much and in which direction weight adaptation should occur, which is uniquely determined by local information according to the delta-learning algorithm. Generally, the selected learning rates were quite robust. In other words, increasing or decreasing this parameter had little substantial effect on the simulations. Only when the original learning rate was already high ( $\geq .25$ ) did increasing the rate further become problematic, because the weights became too large and became unstable. Moreover, this would give too much impact to novel information and would result in a complete neglect of earlier information.

## Dependent Measures

At the end of each simulated experimental condition, to simulate the empirical dependent measures, test trials were run by prompting certain nodes of interest (i.e., turning on their activation), and the resulting output activation in other nodes was recorded. For instance, to test trait inferences, the actor node was turned on and the resulting activation of the trait node (without any additional external activation) was read off. Similar test procedures for the other dependent variables are explained and motivated in detail for each simulation.

Our predictions were verified by comparing the resulting test activations with observed experimental data. Given that the resulting activation values and experimental results are difficult to compare quantitatively, we examined only the general pattern of activations and projected them visually onto the observed data (i.e., we re-scaled the obtained test activations by linear regression with a positive slope). In addition, we report statistical tests between conditions of interest. All tests involved between-subjects analyses of variance (ANOVAs) or unpaired *t* tests, unless otherwise noted. These tests would be impossible in some simulations because a fixed order of trials prevents variability in the results. To avoid this and to add realism to our simulations, in all the localist encodings, we added to the default starting weights a random value ranging between  $-0.1$  and  $+0.1$ .

## Impression Formation

Most processes of impression formation can be thought of as categorization. That is, in making trait inferences on the basis of feature information or based on the behaviors of an actor, the social perceiver attempts to decide to which trait category the person belongs. The categorization of diverse information into meaningful trait concepts or categories that contain similar features, roles, or behaviors promotes cognitive economy and organization, thus enabling us to go beyond the given information and to plan our behavior and interaction with social agents accordingly.

In recent approaches, categorization is most often described in terms of either a prototype or an exemplar approach. According to the prototype approach, learners abstract a central tendency of each category and then classify instances according to their similarity to the category’s central prototype (e.g., Rosch, 1978). In contrast, no such average or typical prototype is assumed in the exemplar approach, in which categorizing depends on the similarity of the given object to a sample of memory traces of category exemplars (Fiedler, 1996; Hintzmann, 1986; Medin & Schaffer, 1978; Nosofsky, 1986; Smith & Zárate, 1992). In this recurrent approach, like in most connectionist models, categorization is per-

formed by prototype extraction, that is, by developing connections between the person and various trait prototypes. The stronger a person's connection with a particular trait category, the more the person is thought to be a member of that category and to possess that trait.

In the following simulations, we illustrate how a recurrent network can model impression formation without recourse to explicit arithmetical calculations, as assumed in algebraic models, with two experiments that grew out of Anderson's (1981) weighted averaging model. In these experiments, participants typically receive a series of trait adjectives or behavior descriptions about an actor and are requested to make overall trait or likability impressions of that person (e.g., Anderson, 1981; Asch, 1946; Kashima & Kerekes, 1994). We focus on representative findings from this research that reflect

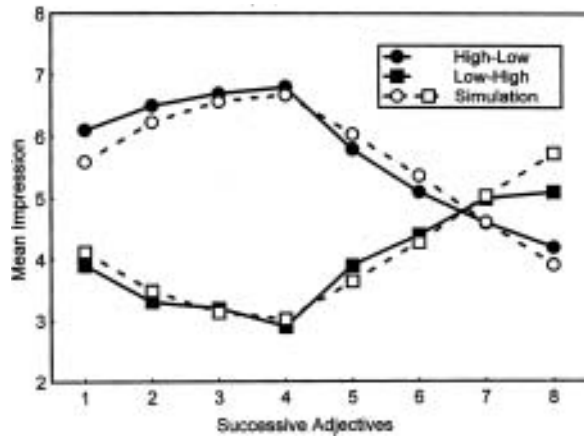
- how impressions are integrated online and grow stronger or weaker after participants are given information bearing on the positive (high) or negative (low) status of a trait, and
- when initial or final pieces of information are weighted more heavily in making an impression.

In this experimental paradigm, judgments are requested along a single dimensional rating scale from the start of the experiment and at several intervals during the experiment, so that we assume a single trait representation (by a single trait node).

**Simulation 1: Online Integration**

We first consider an experiment by Stewart (1965) in which an actor was described by four adjectives implying a high trait (e.g., "talkative") and were followed by four adjectives implying the opposite (or low) trait (e.g., "reticent"). Some participants received high trait information about an actor in the first half of the experiment and low trait information in the second half, whereas other participants received the reverse low-high order. In the continuous condition that we simulate here, after each adjective, participants had to rate the actor on a likability scale that ranged from *highly unfavorable* to *highly favorable*. In line with the predictions of Anderson's (1981) weighted averaging model, Stewart (1965) documented that when high trait information was given, the likability ratings went up, and when low trait information was given, likability went down (see Figure 3). In addition, he also documented a recency effect; that is, the later information had somewhat more impact on the final judgment than the earlier information, as can be seen from the crossover at the rightmost part of the figure.

**Simulation.** Stewart's (1965) experiment was modeled using a network architecture consisting of an



**Figure 3. Online integration of trait-implying information: Observed data from Stewart (1965) and simulation results (general learning rate = .32, for context = .08). The human data are from Figure 1 in "Effect of Continuous Responding on the Order Effect in Personality Impression Formation," by R. H. Stewart, 1965, *Journal of Personality and Social Psychology*, 1, pp. 161-165. Copyright 1965 by the American Psychological Association. Adapted with permission.**

actor node connected to a trait node and an additional context node that reflects situational constraints (e.g., social and group norms) or other experimental context variables (e.g., instructions). This context node guarantees a smooth acquisition curve in agreement with Stewart's data, and, without it, the acquisition pattern is more rigid.

What we want to demonstrate here is how the trait-implying information is applied to build up an impression of a specific actor by changing the weights linking the actor with the trait according to the acquisition principle. Therefore, the simulation starts from the assumption that the likability implied by the adjectives has already been learned and recruited from semantic and social knowledge. This assumption is elaborated in more detail in the next section (Simulation 3). Specifically, we assume here that adjectives associated with likeability are denoted by an activation value of +1 for the trait node, whereas adjectives associated with unlikability have an activation value of -1 (this is equivalent to Anderson's scale values).

Table 2 schematically depicts a list of the trials to simulate the information given in Stewart's (1965) experiment. When the actor is described by a positive (high) trait, the trait node is activated by a value of +1 and the connection weight is increased according to the acquisition principle of the delta algorithm. In contrast, when the actor is described by a negative (low) trait, the trait node is activated by a value of -1 and the weight is decreased according to the acquisition principle. After each adjective, the actor node in the network is prompted and the resulting activation of the trait node indicates what trait the actor conveys (see bottom "test" panel of the table).

**Table 2.** *On-line Integration of Trait-Implying Information (Simulation 1)*

	Actor	Context	Trait
Condition 1: High–Low Presentation Order			
# 4 High trait	1	1	+1
# 4 Low trait	1	1	–1
Condition 2: Low–High Presentation Order			
# 4 Low trait	1	1	–1
# 4 High trait	1	1	1
Test			
Trait of Actor	1	0	?

Note: Schematic representation of the experimental design of Stewart (1965). High = adjective implies trait, Low = adjective implies opposite trait, # = number of trials. Cell entries denote external activation. The simulation was run separately for each condition.

**Simulation results.** The simulation using the recurrent network was run with 50 “participants” (i.e., 50 different simulation runs) and a fixed trial order. The results with learning rate 0.32 (0.08 for the context) are shown in Figure 3. Recall that the simulation data were re-scaled by a linear regression to make them directly comparable to the observed data. As can be seen, there is a close fit between the simulation and the empirical data, which strongly suggests that online integration in impression formation is adequately captured by the acquisition principle of our recurrent connectionist model, much like Anderson’s (1981) weighted averaging model. A repeated-measures ANOVA showed that the differences between trial order were significant in both the high–low condition,  $F(7, 343) = 3853.21, p < .001$ , and the low–high condition,  $F(7, 343) = 3344.23, p < .001$ , and simple  $t$  tests confirmed that the differences between all adjacent trials in each condition were significant,  $p < .001$ .

Of particular interest is the crossover at the end of training. The difference between the two conditions at the most recent trials was significant,  $t(98) = 44.11, p < .001$ . This reflects a recency effect whereby the most recently presented adjectives win over the earlier presented adjectives. It is interesting to note that increasing the learning rate would produce an even stronger recency effect, as it would force novel information to have more impact than older information. Together, the results suggest that the revision and adjustment of person impressions is an online acquisition process whereby novel information often “overwrites” older information previously stored in the connection weights.

### Simulation 2: Serial Position Weights

As a second example, we considered research in which disconfirmatory information is given during a single specific position in a series of trials. By comparing the effect of disconfirmatory with confirmatory information at the same position in the trial series (de-

noted as *serial position*), one can estimate the weight each trait takes at a given position (Anderson, 1979; Anderson & Farkas, 1973; Busemeyer & Myung, 1988; Dreben, Fiske, & Hastie, 1979; Kashima & Kerekes, 1994). Early disconfirmatory trait information might be important in crystallizing an impression (primacy effect), whereas late information might be influential because it sheds new light on traits presented earlier (recency effect).

The bulk of research suggests that when participants give their trait ratings continuously after each adjective is presented, item weights are relatively equal in all but the last position, at which point they rise sharply. This reflects a recency effect that was also observed in the previous simulation. However, it is important to note that this recency effect attenuates after more trait information is given and processed. That is, after been given only a few pieces of trait-implying information, disconfirmatory information has a stronger recency effect than when given a lot of trait-implying information. It is as if increasing the amount of confirmatory information shields the perceiver from the disconfirmatory information. In contrast, when trait ratings are given only once after the whole series of information is presented, then primacy is more likely (for reviews, see Hogarth & Einhorn, 1992; Kashima & Kerekes, 1994).

In a typical experiment by Dreben et al. (1979), participants read about several actors, each described by four behaviors that implied the same high (H) or low (L) extreme of a trait (e.g., HHHH, HHHL, HHLH, ..., LLLL). In the continuous condition, the participants were requested to make a likability rating about the actor after each behavioral description on a scale with endpoints labeled *most likable* to *least likable*. In the final condition, this rating was made after all behavioral descriptions of an actor were presented. Serial position was measured by comparing the judgments of each item list with another list that had the same set of high or low items except for one opposite item (i.e., with a behavior conveying the opposite trait). This inconsistent item was positioned at the first, second, third, or last position of the series so as to provide a means of measuring the weight of the item in the first, second, third, or fourth position of the list. For instance, if we present behavioral items implying the same (high or low) trait by the same characteristic x, y, or z, then the weight of serial position 1 is measured by the mean difference in the ratings between all item lists Hxyz and Lxyz. Similarly, the weight of serial position 4 is measured by the mean difference between all item lists xyzH and xyzL. The results were as expected (see Figure 4). When given concurrent ratings, a recency effect appeared that became weaker at the end of the list, as indicated by the dotted line depicting attenuation of recency. Conversely, after a single final rating, a primacy effect was observed.

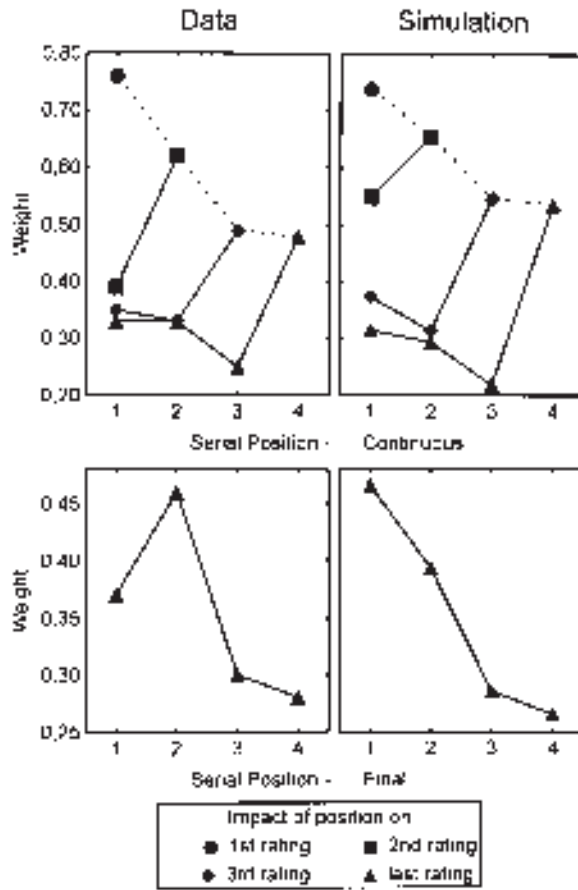


Figure 4. Serial position weight: Observed weight curves from Dreben, Fiske, and Hastie (1979; left panels) and simulation (right panels) of attenuation of recency given continuous responding (top; general learning rate = .25, for context = .025) and primacy given final responding (bottom; general learning rate = .25, for context = .00). The human data are from Figure 1 in “The Independence of Evaluative and Item Information: Impression and Recall Order Effects in Behavior-Based Impression Formation,” by E. K. Dreben, S. T. Fiske, & R. Hastie, 1979, *Journal of Personality and Social Psychology*, 37, pp. 1758–1768. Copyright 1979 by the American Psychological Association. Adapted with permission.

**Simulation.** To simulate the experiment of Dreben et al. (1979), we used the same recurrent architecture as before. The learning history is schematically listed in Table 3. The table depicts only the fourth serial position, and the other serial positions were simulated by changing the order of the inconsistent trial in the item lists. The likability rating was simulated by priming the actor node and reading off the activation of the likability node after presentation of each trial. The serial position weight was measured just as in the original experiment, by calculating the mean difference of the resulting likability activations between items lists that differed only on a single trait.

**Simulation results.** The recurrent simulation was run with 50 participants and a fixed order of trials.

Table 3. Serial Position Weights (Simulation 2)

	Actor	Context	Trait
Item List			
# 3 High or Low	1	1	+1 or -1
# 1 High <sup>a</sup>	1	1	+1
Item List with Opposite Trait			
# 3 High or Low (same as above)	1	1	+1 or -1
# 1 Low <sup>a</sup>	1	1	-1
Test			
Trait of Actor	1	0	?

Note: Schematic representation of the experimental design of Dreben et al. (1979), High = behavior implies high extreme of a trait; Low = behavior implies opposite extreme of trait; # = number of trials; Cell entries denote external activation; Initial weights were set at .05 (for distributed coding .02). The simulation was run separately for each item list.

<sup>a</sup>This trial is presented at position 1, 2, 3, or 4 of the series (here it is shown at position 4), and the resulting test activation averaged across all target trait lists and across all opposite trait lists are subtracted from each other to measure the serial position weight.

The results with learning rate .25 (and .025 for the context) are shown in the top panel of Figure 4. The recurrent network was clearly able to reproduce the predicted recency effects (indicated by the solid lines). For each likeability rating, *t* tests confirmed that the weights were significantly higher at the last position than at previous positions,  $t_s(98) = 2.37-12.21, p < .05$ . More important, there was also attenuation of recency, as a one-way ANOVA showed that the weights at the last serial position (indicated by the dotted line) decreased significantly from ratings given at the beginning (first) to the end (fourth) of the item list,  $F(3, 196) = 9.26, p < .001$ .

As we have seen in the previous simulation, recency is a natural consequence of acquisition in a connectionist network, because later information tends to overwrite earlier information stored in the connection weights. The more crucial question is how the network attained *attenuation* of recency. Here context plays a crucial role. Because the context is always paired with the implied trait, the context→trait weight becomes increasingly stronger given more information and hence competes more strongly against the person→trait weight. Thus, when inconsistent items are provided later in a series, they tend to be discounted more by the increasing (confirmatory) impact of the context. Stated somewhat differently, a robust impression that is the consequence of earlier information in the same context makes the perceiver more resistant to change his or her impression in the face of one disconfirmatory item. This explanation differs from Anderson’s reasoning based on a distinction between item-specific and abstract aspects of impression formation (Anderson & Farkas, 1973; Dreben et al., 1979) but is similar to the tensor-product simulation by Kashima and Kerekes (1994) in its emphasis on the role of the judgmental context.

Let us now turn to the primacy effects that are typically revealed when impression judgments are given at the end of the series of trials rather than continuously (Anderson, 1979; Hogarth & Einhorn, 1992). Perhaps the simplest way to obtain a basic primacy effect is by reducing the learning rate of the context node to zero or by ignoring the context altogether (i.e., by setting the external context activation to zero). The bottom panel of Figure 4 shows the results of the simulation when the learning rate is set to zero. A one-way ANOVA showed that the differences between trial orders were significant,  $F(3,196) = 29.32, p < .001$ .

How did recency disappear by reducing the learning rate or external activation of the context? An analysis of the simulation suggests the following explanation. After a few trials, the person node and the trait node develop strong connections with each other in both directions (cf. the acquisition principle). When a disconfirmatory trial is presented, the negative (or disconfirmatory) activation of the trait node spreads to the person node and reduces the net activation of the person node. (This does not happen when a context is present, because that context sends positive activation to the actor node so that a reduction of net activation in the person node is prevented.) The same logic applies to the trait node. The positive external activation of the person node spreads to the trait node and reduces the negativity of the trait activation. As a consequence of the reduced net activation of the person and trait nodes, there is little adjustment of the person→trait connection. In other words, in the recent trials, the opposing external activations of the person and the trait tend to cancel each other out, resulting in little learning and no recency. This reduction is stronger at later trials when the person and the trait have developed stronger connections. Note that this mutual cancellation is only possible if a single integrative trait node rather than two independent opposite traits is assumed; otherwise this primacy effect would not happen.

A more direct way to simulate primacy is by having a higher learning rate of the actor node at the beginning of the item list that then gradually decreases. This explanation shares with Anderson's (1981) attention decrement hypothesis the notion that there is the most attention paid to and the most uptake of information during the earliest trials, thus allowing little impact of information presented later. Such a procedure was followed in the tensor-product simulation of primacy by Kashima and Kerekes (1994). However, because this forces a primacy effect to occur rather than having it emerge from underlying processing mechanisms, it leaves open the question of the origin of a decrement in attention.

In sum, based on our connectionist simulations, we can explain the different effects of continuous and final judgments by differences in learning or encoding of the context. We assume that when perceivers are made more accountable of their impression by requesting judg-

ments continuously, information uptake and processing becomes more careful, taking into account more of the contextual cues in the actor's situation or experimental set-up. This might be less the case when a trait judgment is requested only once, resulting in more negligence of the context and hence primacy rather than recency.

**Discussion and further research.** Our recurrent model reproduced both recency and primacy effects. Although Kashima (Kashima & Kerekes, 1994; see also Busemeyer & Myung, 1988) pointed out that attenuation of recency could not be simulated with a feedforward network (which is correct) or even with a recurrent network (Kashima et al., 2000), we have demonstrated here on the contrary that it can be reproduced easily with a recurrent connectionist model. The tensor-product model proposed by Kashima and Kerekes was also able to replicate these effects, but they required additional assumptions such as a changing context after each judgment to obtain attenuation of recency. This was not needed in our model, as the presence of the context was sufficient.

We suggest that encoding contextual cues in the presence of the actor allows later information to have more impact (recency), whereas ignoring the context is responsible for building up an impression very quickly (primacy). Furthermore, we suggest that attenuation of recency was due to a growing discounting of inconsistent information by the context. There is some support for our hypothesis that recency is driven by greater attention to the information (including the context). This comes from research demonstrating that conditions that induce increased accuracy or motivation result in more recency as opposed to primacy. A series of studies by Kruglanski (Freund, Kruglanski, & Shpitzajzen, 1985; Heaton & Kruglanski, 1991; Kruglanski & Freund, 1983) documented that conditions that promote accurate judgments reduce primacy. Similarly, Gannon, Skowronski, and Betz (1994) observed that depressives who are more motivated to process information show enhanced recency.

Future research might be able to provide additional support for the predictions of the recurrent network. According to our network, giving more attention to the context to explain the actor's behavior should increase recency as opposed to primacy. Moreover, the more constant the context is, the more the attenuation of recency might be expected, because an unchanging contextual core increases its strength and hence tends to discount more inconsistent information.<sup>1</sup>

<sup>1</sup>Dreben et al. (1979) also reported serial position effects on recall. Regardless of the response conditions (continuous or final), they found primacy of the first two items and recency of the last item. Recall of behavioral information can be simulated by extending the network (including the original opposite traits) with nodes that reflect unique behavioral information (see also Simulation 4). This approach was able to reproduce the reported primacy effect but not the recency effect.

### **Inferring Traits From Correspondent Behavior**

In the previous simulations, we ignored the issue of how information on behaviors or features of an actor is used to make a trait judgment. We simply assumed that the implied trait or evaluative meaning that these behaviors and features carry was directly applied to the actor. Although this simplification might characterize the most important aspects of person-impression processes, there are two questions that remain unanswered. First, are all behaviors equally diagnostic for traits; that is, are some behaviors applied more readily than others in making a trait inference? Second, can this process of inferring associations between behaviors and traits be modeled by a recurrent network? In the next simulation, we attempt to address these questions.

#### **Simulation 3: Asymmetry in Inferences of Ability and Morality**

A remarkable finding in the research on dispositional inferences is that the diagnosticity of a particular behavior varies according to the content of the inference domain. In the domain of ability, for example, because perceivers typically expect that a high level of performance could have been attained only by someone with a disposition of high ability, an actor's high level of performance should prompt an inference of high ability. In contrast, a low level of performance should create greater uncertainty for the perceiver as it may indicate both low and high ability, as even someone with high ability may occasionally fail. This pattern is completely reversed for inferences of morality. Because perceivers expect that a low level of moral conduct could have been attained only by someone with a disposition of low morality, an actor's immoral act should prompt an inference of low morality. In contrast, a high level of moral conduct could indicate both low and high morality, because even immoral persons behave morally most of the time. These behavior–trait expectations lead to an asymmetry in the diagnosticity of ability- and morality-related behaviors. High performance is more diagnostic for inferences of ability, whereas low moral conduct is more diagnostic for inference of morality. Moreover, extreme behaviors are generally seen as more diagnostic of people who have extreme traits, whereas moderate behaviors may be characteristic of both extreme and moderated traits (Lupfer, Weeks, & Dupuis, 2000; Reeder, 1997; Reeder & Fulks, 1980; Reeder & Spores, 1983; Skowronski & Carlston, 1987; Wojciszke, Brycz, & Borkenau, 1993; for reviews, see Reeder & Brewer, 1979; Skowronski & Carlston, 1989).

Where do these asymmetric behavior–trait expectations come from? One of the explanations that has received considerable empirical support is the cue-diagnosticity interpretation of Skowronski and

Carlston (1989). In line with our perspective, these authors assume that behavioral cues are used to assign an actor to one or more trait categories. Behaviors that strongly suggest one trait category (e.g., dishonest) over alternative categories (e.g., honest) are said to be more diagnostic. The asymmetry is assumed to come from differential associations with ability and morality trait categories. According to Skowronski and Carlston, extremely immoral actors may rob banks, but they may also help an old woman to cross the street. On the other hand, moral actors may lie about their age, but they will never rob banks. In contrast, in the ability domain, an outstanding high jumper will sometimes clear 7 ft and sometimes fail to do so. On the other hand, a poor high jumper will probably never clear 7 ft. Thus, immoral behaviors are more often associated with immorality than moral behaviors are associated with morality, whereas competent behaviors are more often associated with high ability than incompetent behaviors are associated with low ability.

One possible interpretation of the cue-diagnosticity notion might result from the semantic meaning of ability and morality traits. This argument relies on the semantics of high or low ability and morality to infer what behaviors are most likely. However, this leaves unanswered the question of how this semantic knowledge was learned in the first place. From a developmental perspective, it seems more likely that when toddlers learn to correct a semantic error, such as an overgeneralization (e.g., calling all male adults “Daddy”), they do so not because they were explicitly told the correct meaning, but rather because they experienced first-hand the circumstances under which children call someone “Daddy.” Therefore, a more interesting interpretation is that the asymmetric strength of the behavior–trait associations is due to the asymmetric distribution of prior relevant observations in the morality and ability domain. These observations might be direct or indirect (e.g., when other people relate their own experiences or observations) and may only later become incorporated in the meaning of ability and morality traits.

The aim of the next simulation is to demonstrate that a skewed distribution during prior learning may lead to differential diagnosticity, as revealed in a study by Skowronski and Carlston (1987, Experiment 1). In this study, participants were provided with descriptions of positive and negative behaviors reflecting five different levels of intelligence and morality. Examples of extreme morality behaviors were robbing a store (low) and returning a lost wallet (high); examples of extreme ability behaviors were failing most exams (low) and teaching at a university (high). For each behavior, participants were asked to estimate to what extent “would a [trait] person ever [behavior]?” (Skowronski & Carlston, 1987, p. 691), with possible traits including dishonest, honest, intelligent, and stupid.

**Table 4.** *Asymmetry in Ability and Morality Trait Inferences (Simulation 3)*

	Behavioral Features										Traits			
	A++	A+	A0	A-	A--	M++	M+	M0	M-	M--	A+	A-	M+	M-
High Ability Trait														
# 6	1	1	1	0	0	0	0	0	0	0	1	0	0	0
# 7	0	1	1	0	0	0	0	0	0	0	1	0	0	0
# 6	0	0	1	0	0	0	0	0	0	0	1	0	0	0
# 5	0	0	1	1	0	0	0	0	0	0	1	0	0	0
# 4	0	0	1	1	1	0	0	0	0	0	1	0	0	0
Low Ability Trait														
# 0	1	1	1	0	0	0	0	0	0	0	0	1	0	0
# 3	0	1	1	0	0	0	0	0	0	0	0	1	0	0
# 4	0	0	1	0	0	0	0	0	0	0	0	1	0	0
# 5	0	0	1	1	0	0	0	0	0	0	0	1	0	0
# 4	0	0	1	1	1	0	0	0	0	0	0	1	0	0
High Morality Trait														
# 6	0	0	0	0	0	1	1	1	0	0	0	0	1	0
# 7	0	0	0	0	0	0	1	1	0	0	0	0	1	0
# 6	0	0	0	0	0	0	0	1	0	0	0	0	1	0
# 3	0	0	0	0	0	0	0	1	1	0	0	0	1	0
# 0	0	0	0	0	0	0	0	1	1	1	0	0	1	0
Low Morality Trait														
# 2	0	0	0	0	0	1	1	1	0	0	0	0	0	1
# 3	0	0	0	0	0	0	1	1	0	0	0	0	0	1
# 4	0	0	0	0	0	0	0	1	0	0	0	0	0	1
# 5	0	0	0	0	0	0	0	1	1	0	0	0	0	1
# 4	0	0	0	0	0	0	0	1	1	1	0	0	0	1
Test														
Trait A+	0	0	?	?	?	0	0	0	0	0	1	0	0	0
Trait A-	?	?	?	0	0	0	0	0	0	0	0	1	0	0
Trait M+	0	0	0	0	0	0	0	?	?	?	0	0	1	0
Trait M-	0	0	0	0	0	?	?	?	0	0	0	0	0	1

*Note:*—Schematic representation of the experimental design of Skowronski and Carlston (1987, Experiment 1), A = ability, M = morality, ++ = extremely positive, + = positive, 0 = neutral, - = negative, -- = extremely negative. # = number of trials. Cell entries denote external activation. All learning trials were presented in an order randomized for each run.

**Simulation.** A possible (simplified) learning history that may reflect participants’ acquisition of prior relevant knowledge on behavior–trait occurrences is illustrated in Table 4. The figure lists five levels of ability and morality behavioral features together with the implied high or low traits. Extreme behaviors are characterized by a configuration of extreme to neutral features, moderate behaviors by a configuration of moderate and neutral features, and neutral behaviors by a configuration of only neutral features. Thus, for instance, extremely competent behavior does not only involve neutral features, such as writing up your research, or moderate features, such as having your article accepted, but also extreme features like publishing the article in a top journal (see first row in the table). In contrast, moderate and neutral competent behaviors involve only lower-level features (see rows 2 and 3).

To be realistic, because extreme behaviors are less likely than moderate behaviors, we began by setting the frequencies of the extreme trait-consistent behaviors equal to those of the neutral behaviors. In addition, to reflect the common finding that perceivers typically hold positive expectations about people, we also assumed that the behaviors for high ability and high mo-

rality traits were more frequent (i.e., 6, 7, 6, 5, 4, from high to low) than those for the low traits (i.e., 4, 5, 4, 3, 2, from low to high). The realism of this learning history is reflected in the fact that the overall distribution reveals many more neutral than moderate or extreme features in people’s behavior and many more moderate than extreme features.

The asymmetric distribution during learning is revealed by the fact that an actor with high ability will most often perform well to very well but will sometimes perform moderately or even occasionally very poorly. For instance, a top researcher will most often publish in top journals but occasionally in lower-ranked journals as well. In contrast, an actor with low ability will most often perform poorly or sometimes moderately but never extremely well. In the table, nondiagnostic behaviors were introduced by setting the frequencies much lower: For extreme behaviors, the frequencies were set to zero, and for moderate behaviors, the frequencies were set to 3. In the morality domain, this asymmetry is reversed. Apart from this, the simulation was quite robust for changes in the nonzero frequencies of Table 4, such as setting the nonzero frequencies to other smooth distributions or even setting them all equal.

Trait-inconsistent behaviors were considered to be most diagnostic by Skowronski and Carlston (1987) and therefore constituted the most important test of their cue-diagnosticity hypothesis. As an example of trait-inconsistent behavior, participants were asked whether an honest person would ever engage in an immoral behavior. As can be seen in the bottom panel of Table 4, these ratings were simulated by prompting the trait and then reading off the output activation of the inconsistent behavior. The reverse direction of testing that would make more sense theoretically would involve asking for the trait implied by the behavior (but this was not used in the experiment). This would be simulated by prompting the behavior and testing for the output activation of the trait, and this procedure worked equally well in the simulation.

**Simulation results.** The learning history and testing prompts depicted in Table 4 were run for 50 participants, each with a different random trial order. Figure 5 shows the results with a learning rate of .10. As can be seen, the simulation results clearly conform to the empirical data obtained by Skowronski and Carlston (1987), and the predicted interaction is significant,  $F(1,196) = 79.98, p < .001$ . Inconsistent low performance is more probably given high ability than is high performance given low ability,  $t(98) = 8.42, p < .001$ . Conversely, an inconsistent high moral behavior is more probably given low morality than is low moral conduct given high morality,  $t(98) = 4.18, p < .001$ .

The underlying connectionist principle that produces these results is low acquisition of the behaviors low in diagnosticity. Recall that testing occurred for trait-inconsistent behaviors. Because high performance behaviors never occurred with low ability, low

ability is not a good predictor of (inconsistent) high performance. In contrast, because both high and low performance behaviors occur with high ability, high ability is a relatively good predictor of (inconsistent) low performance. The reasoning is analogous for morality behaviors and for the reverse connections from behaviors to traits. Simply put, the network does not learn things that it was never taught.

**Extensions and further research.** The same learning history also reproduced Skowronski and Carlston's (1987) straightforward prediction and finding that trait ratings are higher for moderately inconsistent behaviors than for extremely inconsistent behaviors. This was tested in a simulation by comparing the resulting output activation of extreme positive or negative behaviors (as shown by "?" in the bottom panel of Table 4) by the resulting output activation of moderate behaviors (not shown).

With respect to trait-consistent behaviors, Skowronski and Carlston (1987) reported that there were no diagnosticity effects, although earlier work indicates that typical behaviors are usually very diagnostic of the implied trait (e.g., Anderson, 1981; Cantor & Mischel, 1977). Consistent with these latter findings, our model does predict diagnostic effects for trait-consistent behaviors in the same direction as trait-inconsistent behavior. Thus, high performance is predicted to be more diagnostic of high ability than is low performance of low ability; and low moral conduct is more diagnostic of low morality than is high moral conduct of high morality.

Another issue is how prior learning about the different levels and diagnosticity of ability and morality behaviors (as simulated previously) can be used to make a trait inference about a particular actor. Our assumption is that this prior learning on behavior–trait associations is stored in semantic memory and that novel behavioral information automatically spreads to the implied trait. There is considerable evidence demonstrating that traits are automatically and immediately inferred upon reading behavioral information (for a review, see Uleman, 1999). This activated trait is then paired with the actor, resulting in the acquisition of a connection between the actor and the implied trait. Skowronski and Carlston (1987) investigated this process in a second experiment. Participants read information about actors engaging in five different levels of ability- or morality-related behaviors and then judged how intelligent or moral each actor was. As one might expect, the results documented a linear decrease in trait inferences when going from extremely high to extremely low levels of the behaviors. Our recurrent model can easily reproduce this linear relation using the same learning history from Table 4, extended with an actor node.

In the simulations that follow, we do not further implement explicitly the prior learning of behavior–trait association but rather assume—as before—that the im-

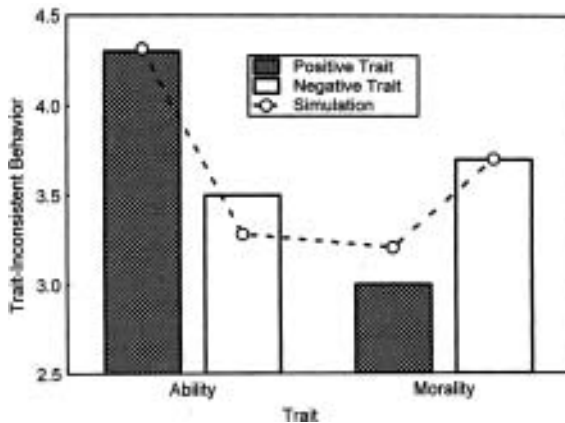


Figure 5. Asymmetric cues of ability and morality: Observed data from Skowronski and Carlston (1987, Experiment 1) and simulation results (learning rate = .10). The human data are from Figure 1 in “Social Judgment and Social Memory: The Role of Cue Diagnosticity in Negativity, Positivity and Extremity Biases,” by J. J. Skowronski & D. E. Carlston, 1987, *Journal of Personality and Social Psychology*, 52, pp. 689–699. Copyright 1987 by the American Psychological Association. Adapted with permission.

plied trait is automatically activated from semantic memory. This allows us to focus on other person-impression phenomena of interest.

### Memory for Behavioral Information

When we learn about others from our own observations, we infer traits from the behaviors that are associated with them. In addition to the question of how we make these inferences, as addressed in the previous section, it is also important to understand how these behavioral observations are stored in memory. An intriguing finding is that inconsistent or unexpected behavioral information about an actor is often better recalled than information that is consistent with the dominant trait expectation (for a review, see Stangor & McMillan, 1992). Thus, we better recall a hooligan helping an older woman cross the street than a nurse performing the same act.

Hastie (1980, Hastie & Kumar, 1979) reasoned that the inconsistent information requires an extra cognitive effort to explain and to make sense of the inconsistency and is therefore elaborated more deeply. This leads to extra links between the inconsistent information and other locations in memory and, thereby, to better recall. Hastie supported this interpretation by research indicating that inconsistent information leads to more causal elaborations of the behavioral sentences. However, these sentence elaborations were explicitly requested from the participants after the initial phase of impression formation was over. It is thus not clear whether they were generated spontaneously during initial encoding or only constructed after the request (cf. Nisbett & Wilson, 1977). Moreover, recent research questioned the assumption that inconsistent behaviors are more strongly associated with other behavioral information about the actor, because support for this assumption was based on flawed measures of associative strength (Skowronski & Gannon, 2000; Skowronski & Welbourne, 1997). It was, therefore, concluded that “associative linkages may not provide the only mechanism, and may not even provide the primary mechanism, for incongruity effects in recall” (Skowronski & Gannon, 2000, p. 17).

#### Simulation 4: Higher Recall for Inconsistent Information

Can our connectionist principles account for the enhanced memory of inconsistent information without recourse to explicit elaborative processes or associations between behaviors? Yes, and to illustrate this, we simulate a well-known experiment by Hamilton et al. (1980, Experiment 3). Participants read information concerning several fictional actors. For each actor, they read a list of 10 consistent and 1 inconsistent behav-

ioral descriptions about that actor, after which they had to recall as many behavioral sentences as possible. The consistent descriptions conveyed common, everyday behavior (e.g., read the newspaper, cleaned up the house), whereas the inconsistent description included a violent behavior (e.g., lost his temper and hit a neighbor, insulted his secretary without provocation). Half of the participants were instructed to form an impression of the actor, and the other half were told to memorize the behavioral information. After a distracter task, the participants were asked “to list as many of the behavior descriptions as they could remember” (p. 1053). The recall data (see Figure 6) documented that under impression-formation instructions, participants were more likely to recall inconsistent items, whereas this difference disappeared under memory instructions.

**Simulation.** To understand enhanced memory for inconsistent behavioral information, consider a network architecture with an actor node connected to two opposing trait nodes. One trait node reflects a neutral trait as conveyed by the descriptions of everyday (non-violent) behavior, whereas the other trait reflects a violent trait implied by the inconsistent violent behavior. In addition, the behavioral descriptions are represented each by a separate node that reflects the particular behavioral exemplar. In sum, each description is represented by two nodes that reflect the categorical trait implied by the behavior as well as the individual behavioral exemplar. Table 5 provides a schematic description of Hamilton et al.’s (1980) experiment with 10 consistent behaviors and 1 inconsistent behavior.

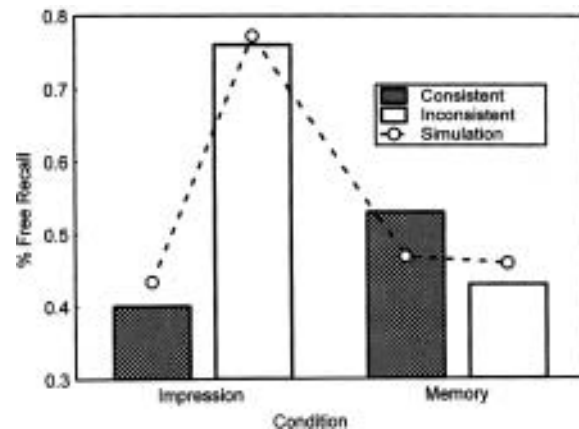


Figure 6. Recall of inconsistent behavioral information after impression-formation and memory instructions: Observed data from Hamilton, Katz, and Leirer (1980, Experiment 3) and simulation results (learning rate given impression instructions = .27, given memorizing instructions = .027). The human data are taken from Table 2 in “Cognitive Representation of Personality Impressions: Organizational Processes in First Impression Formation,” by D. L. Hamilton, L. B. Katz, & V. O. Leirer, 1980, *Journal of Personality and Social Psychology*, 39, pp. 1050–1063. Copyright 1980 by the American Psychological Association.

**Table 5.** *Memory for Inconsistent Information (Simulation 4)*

	Actor	Traits		Behaviors						
		Neutral	Violent	Consistent			Inconsistent			
Item list										
# 1 Consistent	1	1	0	1	0	..	0	0	0	0
# 1 Consistent	1	1	0	0	1	..	0	0	0	0
...										
# 1 Consistent	1	1	0	0	0	..	1	0	0	0
# 1 Consistent	1	1	0	0	0	..	0	1	0	0
# 1 Inconsistent	1	0	1	0	0	..	0	0	0	1
Test										
Recall										
Consistent	1	0	0	?	?	..	?	?	?	0
Inconsistent	1	0	0	0	0	..	0	0	0	?
Unbiased (and Biased) Recognition <sup>a</sup>										
Consistent	?	0 (?)	0	1 <sup>b</sup>	1 <sup>b</sup>	..	1 <sup>b</sup>	1 <sup>b</sup>	1 <sup>b</sup>	0
Inconsistent	?	0 (?)	0	0	0	..	0	0	0	1

*Note:* Schematic representation of the experimental design of Hamilton, Katz, and Leirer (1980, Experiment 3). There were 10 consistent items overall. Although the inconsistent item was given at specific positions in the series, a random order is simulated here (like in most similar studies). Cell entries denote external activation; # = number of trials. All learning trials were presented in an order randomized for each run. The learning rate was reduced to 10% under memorizing instructions.

<sup>a</sup>Between parentheses is the coding for recognition biased in the direction of consistent traits.

<sup>b</sup>Only one exemplar node was activated (with value +1) at a time, and the resulting output activation of all of them was averaged.

As can be seen in the table, each behavior was activated together with the associated trait and the actor node. As predicted by the diffusion principle, however, many of the consistent behaviors are not activated when the consistent trait is present, resulting in a negative learning error for these behaviors and weaker trait→behavior connections. Thus, the more behaviors confirm the consistent trait, the less indicative each behavior becomes for that trait. Because there are more consistent behaviors than inconsistent behaviors, diffusion is particularly strong for consistent behavior. As a result, the trait→behavior connections are weaker for consistent as opposed to inconsistent behavior. It is important to note that because the actor is always active together with the traits, the learning error that causes increased diffusion for consistent behaviors impacts not only on the trait→behavior connections but also on the actor→behavior connections.

In contrast, in the memorizing condition, perceivers are less motivated to form a unified trait impression of the actor. As discussed earlier, we assumed that this would result in a much shallower encoding of actor and trait information, which was simulated by setting the learning rate to 10% of its original value. As a result, all connection weights between the actor or trait and the behaviors would sharply decrease.

To simulate recall of the behavioral episodes, we activated the actor node and read off the resulting activation at each behavioral episode (see bottom panel of the table). This priming procedure assumes that the actor is most strongly available in memory and used as a cue to recall the behaviors. The resulting activation reflects the probability that any of the behavioral exemplars would be recalled. Because research suggests that

participants may use also the actor’s traits as cue (e.g., Hamilton, Driscoll, & Worth, 1989), we also ran the simulations with traits as additional primes to retrieve the behavioral information, together with the actor prime. These simulations gave very similar results.

**Simulation results.** Figure 6 shows the results of the recurrent simulation with 50 participants each receiving a different random trial order. The learning rate under impression instructions was .27, and, as mentioned earlier, under memorizing conditions it was set to 10% of this value (or .027) to reflect shallower encoding. The expected interaction was significant,  $F(1,196) = 5.72, p < .05$ . As can be seen, the simulation replicated the basic finding that inconsistent information was better recalled than consistent information under impression-formation instructions,  $t(98) = 2.33, p < .05$ . However, under memorizing instructions, this difference disappeared,  $t < 1$ . The diffusion principle suggests that higher recall of inconsistent behavioral information is primarily due to relatively stronger trait→behavior and actor→behavior links of unique or infrequent behavioral information. Thus, this connectionist account emphasizes the direct connections from a particular person or trait to behavioral exemplars. This account is consistent with Skowronski’s (Skowronski & Gannon, 2000; Skowronski & Welbourne, 1997) argument that associations between behaviors are not responsible for better recall of inconsistent behaviors, in contrast to earlier suggestions by Hastie (1980) and Srull (1981).

**Extensions and further research.** This network model makes a number of additional “postdictions” (i.e., a “prediction” of something that was already

known independently). The model predicts better recall for items at the end of a list than at the beginning of a list, because diffusion impacts more on earlier information than later information. This has been confirmed by research (Srull, Lichtenstein, & Rothbart, 1985, Experiments 5 and 6). In addition, the model predicts less recall advantage when the number of inconsistent items increases, because this results in more diffusion of the inconsistent items. This prediction has also been confirmed (Hastie & Kumar, 1979, Experiment 3; Srull, 1981, Experiments 1–3; Srull et al., 1985, Experiment 3). However, research has shown that a recall advantage remains even when the number of consistent and inconsistent items is equal, and inconsistency is manipulated by providing advanced trait expectations about the actor (Hastie & Kumar, 1979, Experiment 3). The model also predicts this result if an advanced learning phase is inserted in which the actor is first paired with the consistent trait.

Support for Hastie's (1980) alternative suggestion of effortful generation of elaborations came from studies that found decreased recall for inconsistent behaviors when mental resources were limited by reducing answering time, by making the task more complex, or by adding distracter tasks (Bargh & Thein, 1985; Hamilton et al., 1989; Macrae, Hewstone, & Griffiths, 1993; Stangor & Duan, 1991). However, these results can be easily simulated with our connectionist network by simply assuming that load decreased the encoding of the behavioral episodes or even all information (e.g., by reducing the learning rate to 10% of its original value). This suggests that poorer encoding of conceptual information, rather than less inconsistency reduction and elaboration, might have been responsible for the reduced recall of inconsistent information (for a similar view, see Pandelaere & Hoorens, 2002). Again, there seems to be no need to postulate explicit elaborations to explain higher recall of inconsistent behavior.

Other conditions under which less cognitive effort is spent on the impression-formation task and that typically reveal no enhanced memory for inconsistent recall can be explained in a similar manner. For instance, when an impression or stereotype is formed of a group of individuals rather than a single individual, there seems to be enhanced memory for stereotype-consistent behaviors (for a review, see Fyock & Stangor, 1994). This can be simulated by assuming a decreased learning rate, based on the idea that perceivers do not expect the same level of evaluative consistency in a group of people and therefore are less willing to invest cognitive effort in encoding an overall coherent impression. This results in a loss of enhanced memory of inconsistent behaviors, and, to the extent that the dominant stereotype biases retrieval when in doubt, a stereotype-consistency memory effect is observed.

This network can also simulate recognition measures when participants are presented with behaviors and

asked to assign each of them to the correct actor. Hence, in contrast to recall measures, recognition is simulated by testing the opposite behavior→actor direction. Typically, recognition is biased in the direction of consistent information, presumably because consistent traits guide recognition when the perceiver relies on guessing rather than on genuine memory traces (for a review, see Stangor & McMillan, 1992). This can be replicated by a recognition test that is biased by detecting more easily behaviors congruent with the consistent trait (see “?” for the consistent neutral trait in the bottom panel of Table 5). However, if this bias is removed in the simulation (by setting the activation to zero for all consistent and inconsistent traits), then inconsistent behaviors are again better recognized than consistent behaviors, in line with research showing that improved recognition sensitivity measures reveal better memory for inconsistent behavior (see Stangor & McMillan, 1992). It is important to note that, unlike recall, improved recognition of inconsistent behaviors is based on the competition between trait→actor and behavior→actor connections. When consistent behaviors are presented, the trait that they reflect becomes more strongly associated with the actor. This strong trait→actor connection competes against the weaker behavior→actor connections, so that these latter connections are discounted (see also Van Rooy et al., 2003).

Overall, it appears that the proposed model is broadly consistent with a relatively large spectrum of research findings. This suggests that the diffusion principle provides an interesting alternative hypothesis explaining increased recall for inconsistent information. Moreover, the competition property might play an additional role in the enhanced memory of inconsistent information when recognition measures are used.

### Assimilation and Contrast

An important feature of recurrent models is their capacity to generalize. A trained network exposed to an incomplete pattern of information fills in the missing information on the basis of the complete pattern learned previously. This property was convincingly demonstrated by Smith and DeCoster (1998, Simulations 1 and 2) who used a recurrent network very similar to this one. This generalization process can be seen as a type of assimilation in that past experiences influence how we perceive and interpret novel information that is similar or closely related to it. For instance, when seeing a photo of Hitler, we might immediately complete this image with activated memories on his aggressive wars, mass annihilation of Jews, and so on. There is abundant evidence showing that accessible knowledge like traits, stereotypes, moods, emotions, and attitudes is likely to result in the generalization to unobserved features.

Perhaps a more intriguing and unique property of a recurrent network is the creation of new emergent attributes by combining parts of existing attributes (see Smith & DeCoster, 1998, Simulation 3). Traditional theories of categorization assume that people use a single schema, stereotype, or knowledge structure to make inferences about a target person or a group. Even if multiple schemas are relevant, each of them is independently activated and applied. However, people can combine many sources of knowledge to construct new emergent properties to describe subtypes or subgroups of people. For instance, a militant feminist who is also a bank teller may become subtyped as a feminist bank teller with specific idiosyncratic attributes that are not those traditionally associated either with the militant feminist or the bank teller representations (Asch & Zukier, 1984; Smith & DeCoster, 1998). Previous connectionist models, such as constraint satisfaction models (Kunda & Thagard, 1996), were unable to model this process.

**Simulation 5: Assimilation with Traits, Contrast with Exemplars**

The abundance of assimilation effects in social cognition research may leave the idea that filling in unobserved characteristics is the default or most natural process. Thus, when primed with “violent,” we judge a nondescript or ambiguous target person as more hostile, and when primed with “friendly,” we judge that same target as less hostile. However, under some circumstances, the opposite effect may occur. Sometimes primed features may lead to contrast rather than assimilation.

For instance, when primed with the exemplar “Gandhi,” people may judge a target person as relatively more hostile, whereas primed with “Hitler,” they may judge the same target as relatively less hostile. Under these conditions, the exemplars Gandhi and Hitler serve as an anchor against which the target is judged and so lead to contrast effects. In sum, contextually (or chronically) primed information may not only serve as an interpretation frame that leads to assimilation in impression formation but also as a comparison standard that leads to contrast.

What produces assimilation or contrast? According to Stapel, Koomen, and van der Pligt (1997), trait concepts are more likely to serve to interpret an ambiguous person description (assimilation) because traits carry with them only conceptual meaning. On the other hand, exemplars—if sufficiently extreme—will be used as a comparison standard (contrast) because both the exemplar and the target are persons that can be compared to each other. An experiment by Stapel et al. (Experiment 3) confirmed this proposition. Participants were asked to form an impression of an ambiguous friendly or hostile actor. Before they were exposed to the description of the target, they were primed with traits (e.g., violent or nice) or with names of extreme exemplars (e.g., Hitler or

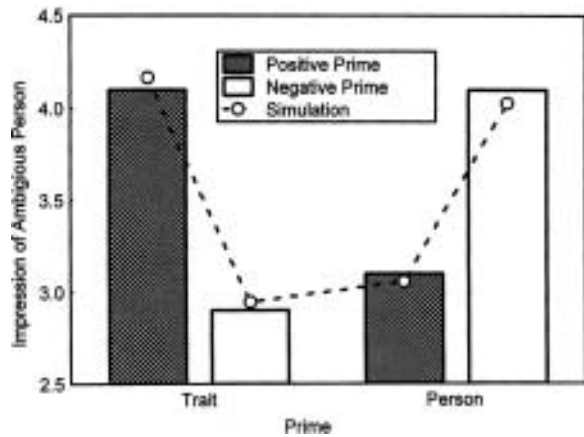


Figure 7. Assimilation and contrast effects after priming with a trait or person: Observed data from Stapel et al. (1997, Experiment 3) and simulation results (learning rate = .15). The human data are taken from Table 3 in “Categories of Category Accessibility: The Impact of Trait Concept Versus Exemplar Priming on Person Judgments,” by D. A. Stapel, W. Koomen, & J. van der Pligt, 1997, *Journal of Experimental Social Psychology*, 33, pp. 47–76. Copyright 1997 by Academic Press.

Gandhi). Finally, they indicated their impression of the target by scoring five trait dimensions that implied either a high or low degree of hostility. A composite scale of these trait ratings demonstrated assimilation in the trait-priming condition but contrast in the exemplar-priming condition (see Figure 7).

**Simulation.** A recurrent network can simulate this combination of assimilation and contrast. As listed in Table 6, the network first builds up background knowledge about extreme exemplars like Gandhi and Hitler by linking them with friendly and hostile traits, respectively. The essential idea of the simulation is that during priming, the primed stimulus and the target description are temporarily activated together. This is represented by programming two learning trials. The first trial represents the priming condition and the second trial the description of the ambiguous target whereby the activation from the previous priming trial is left as the starting activation in addition to the external activation of the actor (see Table 6). We tested the impression of the target by prompting the target node and reading off the activation of the friendly node and the (reversed) activation of the hostility node.

**Simulation results.** Figure 7 depicts the results for 50 participants run in different random orders with a learning rate of .15. As can be seen, the simulation replicated the empirical findings reported in the study of Stapel et al. (1997). The predicted interaction was significant,  $F(1,196) = 797.33, p < .001$ . There was assimilation of the trait prime, as the rating of the ambiguous actor was higher after priming with a positive as opposed to negative trait,  $t(98) = 22.59, p < .001$ . Conversely, there was contrast away from the exemplar prime, as the rating was

**Table 6.** *Assimilation and Contrast (Simulation 5)*

	Actors			Traits	
	Target	Gandhi	Hitler	Friendly	Hostile
Prior Learning History					
# 10 Gandhi	0	1	0	1	0
# 10 Hitler	0	0	1	0	1
Condition 1: Priming Gandhi Exemplar					
# 1 Gandhi	0	1	0	0	0
# 1 Target description	1	= <sup>a</sup>	0	0	0
Condition 2: Priming Hitler Exemplar					
# 1 Hitler	0	0	1	0	0
# 1 Target description	1	0	= <sup>a</sup>	0	0
Condition 3: Priming Friendly Trait					
# 1 Friendly	0	0	0	1	0
# 1 Target description	1	0	0	= <sup>a</sup>	0
Condition 4: Priming Hostile Trait					
# 1 Hostile	0	0	0	0	1
# 1 Target description	1	0	0	0	= <sup>a</sup>
Test	1	0	0	?	-?

*Note:* Schematic representation of prior knowledge acquisition and experimental design of Stapel et al. (1997, Experiment 3). Cell entries denote external activation; # = number of trials. All trials during the prior learning history were presented in an order randomized for each run. The simulation was run separately for each condition.

<sup>a</sup>Activation from previous trial is left as starting activation for current trial.

lower after priming with a positive as opposed to negative prime,  $t(98) = 17.42, p < .001$ .

How was this result obtained? When a trait concept is primed, this activation spills over when the actor is presented, leading to stronger actor→trait connections through the principle of acquisition. As judging the actor's trait involves testing these actor→trait connections, this leads to the usual assimilation of the trait impression. In contrast, when an exemplar such as Hitler is primed, competition arises between this primed exemplar and the target exemplar in their connection to the hostile trait. Thus, competition arises between the (stronger) Hitler→trait connection and the target→trait connection, leading to discounting of the target→trait connection or a contrast effect.

**Extensions and further research.** This network makes an interesting prediction with respect to the impact of the extremity of exemplar and trait primes. According to the competition principle, extreme exemplars that serve as a comparison standard should cause more overestimation in the network and thus lead to stronger contrast. Similarly, as one might expect from the generalization (acquisition) property, extreme trait primes should lead to more assimilation. This prediction was supported by a recent study by Moskowitz and Skurnik (1999). In two experiments, they found that moderate exemplars (e.g., Kissinger) lead to less contrast than extreme exemplars (e.g., Hitler) and that moderate trait primes lead to less assimilation than extreme primes. This was produced in our simulations by replacing our friendly Gandhi exemplar (from Stapel et al., 1997) by a moderately hostile Kissinger exemplar (as used by Moskowitz &

Skurnik, 1999) linked with a moderate hostility trait (i.e., with activation .10) to obtain a moderate exemplar and priming the hostility trait by an activation of only .10 to simulate a moderate trait.

Interestingly, this latter simulation was also able to reproduce the additional finding of Moskowitz and Skurnik (1999) that cognitive interference (i.e., increasing task load or interrupting the current task) minimized the effects of trait assimilation but left the effects of exemplar contrasts relatively untouched. This was done by simulating decreased resources during priming by a decreased (10% of the original) learning rate. This eliminated assimilation effects but preserved the contrast effects, as found by Moskowitz and Skurnik (Experiments 3 and 4).

### Discounting by Situational Constraints

Most of the simulations that we have discussed so far rely on experimental paradigms in which trait-relevant information is provided about an actor in the form of traits or short behavior descriptions, and participants are to assume that this information is applicable to this person. As discussed earlier, however, social perceivers are much more sophisticated and often realize whether they should either utilize or disregard such information. When there are situational constraints that may have provoked the actor's behavior, or when there are many others behaving in similar ways, perceivers tend to discount the behavioral information and are less likely to make a correspondent trait inference (Gilbert & Malone, 1995; Gilbert, Pelham, & Krull, 1988; Trope, 1986; Trope & Gaunt, 2000).

The operation of this discounting process should be most evident when situational constraints are explicitly manipulated in the experiment. In the following simulations, we discuss such experimental paradigms in which complex behavioral scenarios about an actor are provided, including information about the situation, so that the perceiver must figure out whether the actor was the cause of the behavior before a correspondent trait inference can be made. Discounting by the context or situation involves the competition property, as we discussed before. This principle has already been demonstrated on mere causal attributions using a connectionist network with the delta learning algorithm (Read & Montoya, 1999; Van Overwalle, 1998) but not yet on trait inferences. The aim of the next section is to extend this approach also to trait inferences, to verify whether the underlying principles in causal attribution are generalizable to traits and to motivate further the use of context nodes as applied in the earlier simulations.

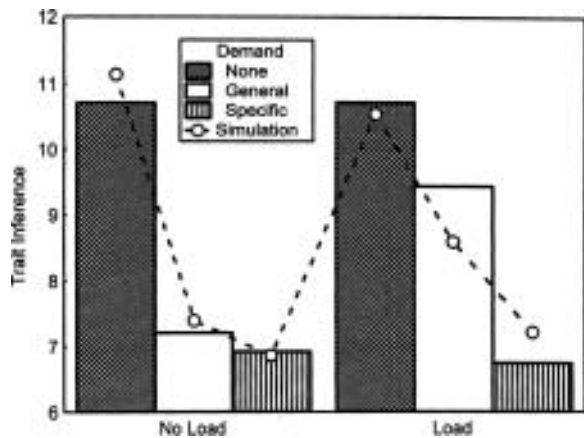
**Simulation 6: Situational Correction or Integration?**

One of the current debates in the literature on trait attributions concerns the process by which behavior–correspondent trait inferences are discounted given situational demands. In so-called correction theories, it is assumed that perceivers first automatically attribute the behavior to the correspondent trait and then discount the trait inference on the basis of situational information in a separate resource-dependent stage (Gilbert & Malone, 1995; Gilbert et al. 1988). In contrast, integration theories assume that situational information is used as an integral part of drawing trait inferences from behavior, whereby the weighting of the contribution of person and environmental factors “involves an iterative or even simultaneous evaluation of the various hypotheses before reaching a conclusion” (Trope & Gaunt, 2000, p. 353, see also Trope, 1986). It is evident that the connectionist perspective is more in line with the latter perspective, because network models typically allow many processes to occur in parallel without the need of separate and sequential process stages.

To demonstrate that discounting in trait inferences can be explained in part by parallel processing in a connectionist network, we focus on the work by Trope and Gaunt (2000). Trope and Gaunt replicated the well-known finding that situational information is often underutilized to discount trait inferences, especially under conditions of cognitive load. This finding has often been taken as evidence by correction theorists that the effortful correction stage was interrupted by the cognitive load manipulation (Gilbert & Malone, 1995). However, Trope and Gaunt argued that situational information might be underutilized because it is often less salient or applicable for the actor and that this rather than effortful correction might explain why

cognitive load often disrupts discounting. To support their view, they made situational information more cognitively salient, active, or applicable and found that, under these circumstances, situational information was used to discount trait inferences even under cognitive load (Trope & Gaunt, 2000).

In one of their experiments, Trope and Gaunt (2000, Experiment 3) provided a description of a teaching assistant who used strict criteria in grading an exam. Information about situational demands varied according to condition. In one condition, no demand was provided; in a general demand condition, participants were told that there was a university-wide requirement to use strict criteria in grading exams; and in a specific demand condition, participants were told that the professor of the exam gave specific instructions for the assistant to use strict criteria. Finally, the participants were asked to infer how strict the assistant was on a 13-point rating scale ranging from 1 (*not a strict person at all*) to 13 (*a very strict person*). In the cognitive load condition, the participants were asked to recall an eight-digit number during the task. The results showed that under no load conditions, discounting was applied. That is, the teaching assistant received less strict ratings given both general and specific demands. More crucially, in line with their integration prediction, Trope and Gaunt also documented that discounting was applied even under cognitive load when the demand was specific and directly applicable but not when it was general and less applicable (see also Figure 8). This demonstrates that discounting could not have been applied in a second effortful correction stage, as the cognitive load should have prevented any discounting in this stage. In the next simulation, we demonstrate that a recurrent approach can explain these findings by replicating Trope and Gaunt’s experiment.



**Figure 8. Integration of situational information: Observed data from Trope and Gaunt (2000, Experiment 3) and simulation results (learning rate given no load = .33; given load = .18). The human data are taken from Table 3 in “Processing Alternative Explanations of Behavior: Correction or Integration?” by Y. Trope & R. Gaunt, 2000, *Journal of Personality and Social Psychology*, 79, pp. 344–354. Copyright 2000 by the American Psychological Association.**

**Table 7.** *Integration of Situational Information (Simulation 6)*

	Actors			Traits	
	Teaching Assistant	University Administration	Professor	Strict	Lenient
Prior learning history					
# 5 Teaching assistant	1	0	0	1	0
# 10 Professor	0	0	1	1	0
# 8 University administration	0	1	0	1	0
# 2 University administration	0	1	0	0	1
Condition 1: No demand					
# 2 No demand	1	0	0	1	0
Condition 2: General demand					
# 2 University administration	1	1	0	1	0
Condition 3: Specific demand					
# 2 Professor	1	0	1	1	0
Test					
Trait of Teaching Assistant	1	0	0	?	-?

*Note:* Schematic representation of prior knowledge acquisition and experimental design of Trope and Gaunt (2000, Experiment 1). Cell entries denote external activation; # = number of trials. All trials during the prior learning history were presented in an order randomized for each run. The simulation was run separately for each condition. In the load condition, the learning rate was reduced to 50% in Conditions 1 to 3.

**Simulation.** Table 7 provides a schematic description of the learning history reflecting Trope and Gaunt's (2000) Experiment 3. As can be seen, the difference between specific and general demands was implemented in the prior learning history that participants brought with them. This learning history assumed that teaching assistants are typically strict in their grading, just like professors, although professors have more experience (i.e., more learning trials). In contrast, general academic instructions are less often followed (i.e., stricter grading was used most of the time, but sometimes lenient grading was applied as well). As a consequence, a specific demand by the professor should lead to stronger actor→trait connections than a general demand by the academic authority. (Different trial frequencies lead to similar simulation results as long as they preserve a similar proportion of strict versus lenient.)

The three demand conditions were then simulated in separate simulation runs in which we assumed two trials per condition. Cognitive load was simulated by reducing the learning rate in the three demand conditions to 50% of its original value. In contrast to some of the previous simulations, we assumed more trials and less reduction of learning rate because information consisted not of a simple trait adjective or trait-implicating behavior, but rather of more elaborated behavioral scenarios that presumably take somewhat longer and more attention to process. Finally, trait inferences were measured as before, by probing the actor node and reading off the trait nodes.

**Simulation results.** The simulation was run with 50 participants for each condition, with different random orders for each participant. The learning rate was .33 under no load conditions and 50% of this default under load conditions or .18. The results, depicted in

Figure 8, demonstrate that the recurrent network largely replicated the findings of Trope and Gaunt (2000). The predicted interaction between demand and cognitive load reached significance,  $F(2,294) = 4.08$ ,  $p < .05$ . More important, as predicted, specificity of demands had a differential effect under cognitive load than under no load. Although all demand conditions produced discounting,  $t(98) = 4.78$  to  $9.74$ ,  $p < .001$ , under no load, general and specific demands did not differ from each other,  $t(98) < 1$ , *ns*. In contrast, under load, general demands produced less discounted inferences than specific demands did,  $t(98) = 3.03$ ,  $p < .01$ .

How was discounting achieved in the simulation? This was due to the principle of competition. Recall that a strong actor→trait connection by the professor or administration was built up in the prior learning history. When these actors (i.e., their demands) are present, their strong connection competes against the assistant's actor→trait connection. Consequently, the assistant→trait connection is not increased further, reflecting discounting in comparison with a condition without any demand. Thus, it is important to realize that in this approach, when perceivers are aware of the situational demands, they never make a trait inference that is later reduced, as the two-stage model of Gilbert and Malone (1995) would predict. Rather, from the beginning of the parallel connectionist computation, the actor's trait inference is effectively discounted by the situational demand.

How was the crucial difference in discounting of general demands (by the administration) under no load and high load achieved in the simulation? The answer is rather straightforward. Because of the increased processing (that is, higher learning rate) in the no load condition, competition was applied more strongly, so that even the softer general demand led to a blocking of correspondent trait inferences in the no load condition.

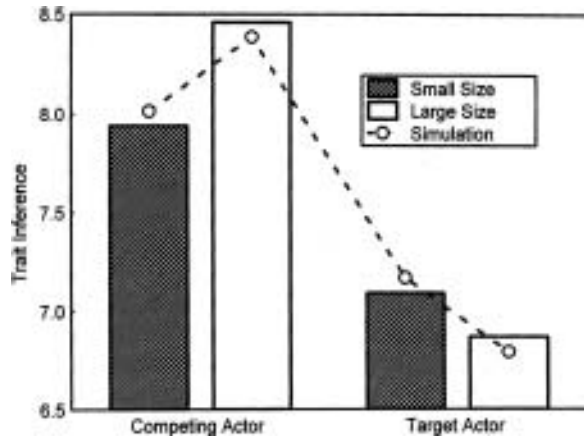
This was less so in the load condition, where there was a lower learning rate and thus less competition. These same principles used in our simulation also allowed replicating the results of the other experiments by Trope and Gaunt (2000).

**Simulation 7: Discounting and Sample Size**

An important assumption in the previous simulation was that stronger or more salient competing situational factors prevent the building up of trait inferences about the actor. This assumption is based on the combination of the principles of acquisition (to develop strong situation→trait connections during prior learning) and of competition (to induce discounting of the actor→trait connections). But is there more direct evidence for this assumption?

The combination of the principles of acquisition and competition provides an interesting test case to distinguish our approach from previous algebraic models of attribution (Cheng & Novick, 1992; Försterling, 1992) and impression formation (Anderson, 1981; Hogarth & Einhorn, 1992). These models would not predict that the increased frequency by an alternative actor or situational factors alone results in a greater discounting of a target actor (Van Overwalle & Van Rooy, 2001a). Alternative connectionist models also fail to make this prediction: the tensor-product model by Kashima and Kerekes (1994), because it does not possess the competition property, and the constraint satisfaction model by Kunda and Thagard (1996), because of the lack of the acquisition property.

To test this prediction, Van Overwalle (2001) combined differences in sample size with discounting. In particular, trait inferences were increased by increasing the frequency of an actor’s behaviors, and this was expected to induce greater discounting of the trait inferences of a target actor. Participants read several stories, each describing a competing actor who engaged in a particular behavior (e.g., “Stephan solved one or five questions during a quiz”). The competing actor engaged in this behavior only one time (i.e., small size condition) or five times (large size condition). Next, regardless of condition, participants read five descriptions in which the competing actor engaged in the same behavior together with a novel target actor (e.g., “Stephan and Walter worked together to solve another five questions”). After receiving this information, they had to rate the traits of the two actors. In our example, they had to rate how intelligent each contestant was on an 11-point scale ranging from 1 (*not at all intelligent*) to 10 (*very intelligent*; Van Overwalle, 2001b). Consistent with our connectionist prediction, but contrary to earlier models, the results revealed that a greater frequency (i.e., sample size) of the competing actor led not only to higher inferences of the implied trait to the competing actor, but also



**Figure 9. Discounting and sample size: Observed data from Van Overwalle (2001) and simulation results (general learning rate = .13, for competing actor = .08).**

to significantly greater discounting of these trait inferences for the novel target actor (see Figure 9). In other words, when Stephan solves more questions, he is seen as more intelligent, whereas Walter is seen as less intelligent. These results for trait inferences replicated similar findings for mere causal attributions following the same combined manipulation of sample size and competition (Van Overwalle & Van Rooy, 2001a).

**Simulation.** Table 8 shows the design of the simulation for this experiment (Van Overwalle, 2001). Because only one type of behavior was described in each of the stories, only a single trait node was provided (i.e., there was no behavior implying the opposite trait). As can be seen, the competing actor first engaged in a trait-implying behavior alone and then together with the target actor. The crucial difference between conditions is sample size, or the number of times (one or five) the competing actor engaged in the behavior alone.

**Simulation results.** The simulation was run with 50 participants, and trial order was fixed. Each sample size condition was run in a separate simulation. Figure 9 depicts the results for a learning rate of .13 (and .08 for the competing actor). As can be seen, the simulation replicated the empirical data. An ANOVA with sample size as a between-subjects factor and rating (target vs. competing actor) as repeated measures indicated that the predicted interaction was significant,  $F(1, 196) = 95.38, p < .001$ . Given a greater sample size, the simulation reveals stronger trait inferences of the competing actor,  $t(98) = 6.63, p < .001$ , and simultaneously weaker trait inferences of the target actor,  $t(98) = 6.60, p < .001$ . Thus, the simulation supports the unique prediction of the connectionist approach that the more often other people perform the same behavior, the less likely perceivers believe that the target actor possesses the correspondent trait. Note that when

**Table 8.** *Discounting in Function of Sample Size of the Competing Actor (Simulation 7)*

	Actors		Trait (e.g., intelligent)
	Competing	Target	
Condition 1: Small sample size			
# 1 Competing actor	1	0	1
# 5 Target and competing actor	1	1	1
Condition 2: Large sample size			
# 5 Competing actor	1	0	1
# 5 Target and competing actor	1	1	1
Test			
Competing actor	1	0	?
Target actor	0	1	?

*Note:* Schematic representation of the experimental design of Van Overwalle (2001), Cell entries denote external activation, # = number of trials. The simulation was run separately for each condition.

the order of the trials is randomized rather than blocked per condition (as it was in the simulation), the network still predicts discounting although to a lesser degree because the influence of the competing actor is built up in later trials and thus has less impact.<sup>2</sup>

**Extensions and further research.** As noted in the introduction, the competition principle illustrated here can be extended to explain the influence of covariation information (Kelley, 1967) on trait inferences. When only a few people engage in behaviors similar to the actor (i.e., low consensus), we tend to make stronger person attributions and correspondent trait inferences than if a lot of people engage in similar behaviors (i.e., high consensus). Alternatively, if the actor engages in the behavior only under specific circumstances (i.e., high distinctiveness), we tend to make more entity attributions and hence less correspondent trait inferences than when the actor behaves similarly under many different circumstances (i.e., low distinctiveness). Van Overwalle (1997, 2003) showed that the joint manipulation of these two covariation methods leads to trait ratings very similar to those obtained by Stewart (1965; see also Figure 3) where only simple trait descriptions of a single target were provided. These results could be simulated by a similar learning history as depicted for Simulation 1.

### Fit to Data and Model Comparisons

The simulations that we have reported all replicate empirical data and theoretical predictions reasonably well. However, it is possible that this fit is due to some procedural choices of the simulations rather than a

more general conceptual validity. The aim of this section is to demonstrate that changes in these choices generally do not invalidate our simulations. To this end, we explore a number of issues, including the localist versus distributed encoding of concepts and the specific recurrent network used. In addition, we also discuss how the recurrent approach compares to other connectionist network models. We do not discuss the algebraic models of Busemeyer (1991) and Hogarth and Einhorn (1992) as they are, in fact, simplified versions of the delta algorithm applied in connectionist models. We address each issue in turn.

### Distributed Coding

The first issue is whether the nodes in the autoassociative architecture encode localist or distributed features. As mentioned earlier, localist features reflect “symbolic” pieces of information; that is, each node represents a concrete concept. In contrast, in a distributed encoding, a concept is represented by a pattern of activation across an array of nodes, none of which reflect a symbolic concept but rather some subsymbolic microfeature of it (Thorpe, 1994). We used a localist encoding scheme to facilitate the understanding of the processing mechanisms underlying connectionism. However, localist encoding is far from biologically and psychologically realistic because it implies that each concept is stored in a single processing unit and, except for differing levels of activation, is always perceived in the same manner. Unlike such a localist encoding scheme, a distributed activation pattern allows for noisy or incomplete inputs to receive a fair amount of activation from similar inputs seen earlier (see Smith & DeCoster, 1998) and to sustain partial damage. Given the advantages of distributed coding, is it possible to replicate our localist simulations with a distributed representation?

To address this question, we reran all simulations with a distributed encoding scheme in which each concept (e.g., trait, behavior, situation, and so on) was repre-

<sup>2</sup>To simulate that both actors work independently rather than together on the quiz, one would have to assume that each of them is linked to a different behavior or outcome, each leading to the same intelligent trait. As such, no competition would arise and no discounting would be predicted.

**Table 9.** *Fit and Robustness of the Simulations, Including Alternative Encoding and Models*

Nr & Topic	Original Simulation	Distributed	Feedforward	Nonlinear Recurrent
1 Online integration	.98	.95	.97	.96
2a Weight–Continuous	.94	.89	.85 <sup>b</sup>	.81
2b Weight–Final	.71	.72	< 0 <sup>b</sup>	< 0 <sup>a,b</sup>
3 Asymmetric Cues	.96	.91	.91	.89
4 Inconsistent Behavior	.96	.96	.92	.82
5 Assimilation and Contrast	.99	1.00	.99	1.00
6 Situational Integration	.97	.96	.93 <sup>b</sup>	.96 <sup>a</sup>
7 Discounting and Size	.99	.99	.99	.99

*Note:* Cell entries are correlations between mean simulated values (averaged across randomizations) and empirical data. For the distributed encoding, we ran 50 “participants,” and each concept was represented by five nodes and an activation pattern drawn from a normal distribution with  $M$  = activation of the original simulation and  $SD$  = .20 (5 such random patterns for 10 “participants” were run and averaged) and additional noise at each trial drawn from a normal distribution with  $M$  = 0 and  $SD$  = .20. For the nonlinear auto-associative model, the parameters were:  $E = I = Decay = .15$ , and *internal cycles* = 9 (McClelland & Rumelhart, 1988). For all alternative models, we searched for the best fitting learning parameter.

<sup>a</sup>Number of internal cycles = 4

<sup>b</sup>Predicted pattern was not reproduced.

sented by a pattern of activation along five instead of a single node. All simulations were run with 50 participants, and each set of 10 participants received a different random pattern of activation for each concept to ensure that the simulation results generalized across activation patterns. For each participant and trial, random activation (i.e., noise) was added to this activation to simulate the imperfect conditions of perceptual encoding (see Table 9 for details). The fit with the observed data was measured by calculating the correlation between the observed and simulated means. These correlations are merely indicative, as the number of means (4 or more) is too few to obtain reliable differences between correlations. As a way of comparison, the correlation of the original localist simulations is also given.

As can be seen, all distributed simulations attained a good fit to data. In all cases, the pattern of results from the original localist simulations was reproduced. This suggests that the underlying principles and mechanisms that we put forth as being responsible for the major simulation results can be obtained not only in the more contrived context of a localist encoding, but also in a more realistic context of a distributed encoding.

**Feedforward Model**

In our discussion of the three properties of the delta algorithm, typically one direction of the connections was responsible for replicating the phenomena. Specifically, we focused on the connections as they were oriented from input (involving most often actors and behaviors) to output (involving trait categories) except for the diffusion property in Simulation 5, where this order was reversed. To illustrate that these input→trait connections are of primary theoretical importance, we reran the simulations with a feedforward pattern associator (McClelland & Rumelhart, 1988) that consists only of feedforward connections from input to trait (see also the arrangement of nodes from left to right in Tables 2 to 9).

As can be seen in Table 9, for most simulations, a feedforward architecture did almost as well as the original simulations. One major exception was the simulation of recency and primacy effects in serial position weights (Simulation 2). As noted earlier, the feedforward network is unable to replicate the critical finding of attenuation of recency in continuous judgments and robust primacy in final judgments. In addition, in Simulation 6, there was an unexpected difference between the general and specific demand conditions given no load, although as predicted, overall discounting was stronger in these two conditions than given high load. This confirms that for the majority of phenomena in person perception, one direction of the connections in the network was most crucial. This does not deny the fact that the additional lateral or backward connections play a role, although only a smaller one.

**Nonlinear Recurrent Model**

We also claimed earlier that a recurrent model with a linear updating activation algorithm and a single internal updating cycle (for collecting the internal activation from related nodes) was sufficient for reproducing the social phenomena of interest. This contrasts with other social researchers who used a nonlinear activation updating algorithm and many more internal cycles (Read & Montoya, 1999; Smith & DeCoster, 1998). Are these model features necessary or even preferable? To answer this question, we ran all our simulations with a nonlinear activation algorithm and 10 (i.e., 1 external and 9 internal) cycles.

As can be seen from Table 9, although the nonlinear model yielded an adequate fit, most simulations did not improve substantially compared to the original simulations. In the simulation on integration of situational information (Simulation 6), the number of internal cycles had to be reduced from nine to four to obtain meaningful results. The reason is that the nonlinear updating easily attenuates the competition property because it drives ac-

tivations that are too high (beyond +1) back to the default ceiling level of +1 and so prevents overestimation and competition to occur. By taking fewer internal cycles, this attenuation of competition is sometimes prevented. However, in the simulation on serial position weights (Simulation 2b), reducing the number of internal cycles to four or even one was not enough to obtain a primacy effect. The reason is similar, as the nonlinear updating drives the reciprocal reduction of activation by actor and trait given disconfirmatory items (which slowed down learning and caused primacy) back to normal +1 ceiling levels. Taken together, this suggests that the linear activation update algorithm with a single internal cycle is sufficient for simulating many phenomena in impression formation.<sup>3</sup> This should not come as a surprise. In recurrent simulations of other issues, such as the formation of semantic concepts, multiple internal cycles were useful to perform “cleanup” in the network so that after activating a perceptual input (e.g., hearing the word “cat”), the activations of the associated semantic concept were forced to eventually settle into “attractor” representations that had preestablished conceptual meaning (e.g., McLeod et al., 1998, pp. 145–148). Such a distinction between perceptual and conceptual levels was not made here, and, as a result, multiple internal cycles had no real function.

### **The Parallel-Constraint-Satisfaction Model**

We can be brief about the parallel-constraint-satisfaction network model developed by Kunda and Thagard (1996). Because this model lacks a learning algorithm, it has no acquisition property and is therefore incapable of replicating any of the simulations we presented. We see no way how this model can be amended, unless by major alterations that would definitely change the model drastically.

### **The Tensor-Product Model**

The tensor-product model is an important alternative connectionist approach to person and group impression formation and change (Kashima & Kerekes, 1994; Kashima et al., 2000). A major difference as compared to our recurrent model is that the tensor-product model uses a Hebbian learning algorithm. This type of learning has the significant disadvantage that it does not imply the competition property. Hence, social phenomena explained by this property such as contrast, situational correction, and discounting (Simulations 5 to 7) can presumably not be simulated with this model, at least not without additional assumptions. In addition, to simulate attenuation of recency in im-

pression formation, this model requires the ad hoc assumption of different context presentations before and after a judgment (Kashima & Kerekes, 1994). This assumption was not required with our simulations (see Simulation 2). For all other phenomena that we simulated, it appears to us that the tensor-product model might simulate most of them, although we are not sure about the diffusion property (Simulation 4). Of course, we do not have any idea as to how close the tensor-product model might fit the data and whether it might do so equally well as the recurrent model.

## **General Discussion**

In this article, we have presented an overview of a number of major findings in impression formation and have shown how they might be accounted for within a connectionist framework. This connectionist perspective offers a novel view on how information can be encoded, how it might be structured and activated, and how it can be retrieved and used for social judgment. This view differs from earlier theories in impression formation that relied on metaphors such as algebraic arithmetic (Anderson, 1981; Busemeyer, 1991; Hogarth & Einhorn, 1992), phase-like integration of information (Gilbert, 1989), or spreading activation and constraint satisfaction networks with fixed weights (Kunda & Thagard, 1996; Read & Marcus-Newhall, 1993; Shultz & Lepper, 1996). The problem is that these various metaphors give a rather inflexible, incomplete and fragmentary account of person-perception mechanisms.

In contrast, the connectionist approach proposed in this article, although relying on the same general autoassociative architecture and processing algorithm, has been used in such a way as to be applicable to a wide-ranging number of phenomena in impression formation. Moreover, we have shown that this model provides an alternative interpretation of earlier algebraic models (Anderson, 1981; Busemeyer, 1991; Hogarth & Einhorn, 1992). In addition, this model can also account for the learning of social knowledge structures. This involves not only the episodic relation between actors and their traits and behaviors (Hamilton et al., 1980), but also the more permanent semantic knowledge that relates behaviors to traits (Skowronski & Carlston, 1987, 1989). Hence, this approach potentially could be used to investigate the development among infants and children of the structures underlying social knowledge.

A basic assumption in our simulations of the development of semantic trait meaning is that traits are seen as prototypes in which strong associations between certain cues (behaviors) and categories (traits) are built from more frequent exposure to behavior–trait pairings. This assumption might seem questionable, because behaviors or attributes that are highly

<sup>3</sup>The simulation results by Smith and DeCoster (1998, Simulations 1 to 3) could also be obtained with linear updating of activation.

prototypical of trait categories are also very rare and may never even have been encountered before. After all, how often do we have exposure to someone who behaves extraordinarily honestly or dishonestly? Hence, one might argue that these extreme or idealized trait prototypes are not simply retrieved from memory when making a judgment but instead are constructed as needed. However, in contrast to this idea, research has demonstrated that extraordinary features not seen previously are judged more atypical and are categorized less quickly than extraordinary features seen earlier (Nosofsky, 1991). Thus, some limited exposure to extraordinary instances is important in making extreme trait inferences; otherwise people may be tempted to classify extraordinary exemplars to different categories, such as, for instance, UFOs or heroes, to which normal exemplars do not belong. Moreover, note that in our simulations (e.g., Simulation 3), an extreme trait inference was not viewed solely as stemming from higher behavioral frequency (in fact, the frequencies were set equal to neutral behaviors), but also as involving behaviors with a wider ranging configuration of features. Thus not only feature frequency, but also the configuration of many features (some of which are typical and others of which are extreme), was sufficient to induce extreme trait inferences.

We have focused to a large extent on the model as a learning device, that is, as a mechanism for associating patterns that reflect social concepts by means of very elementary learning processes. One major advantage of a connectionist perspective is that complex social reasoning and learning can be accomplished by putting together an array of simple interconnected elements that greatly enhance the network's computational power and by incrementally adjusting the weights of the connections with the delta learning algorithm. We have demonstrated that this learning algorithm gives rise to a number of novel properties, among them the acquisition property that accounts for sample size effects, the competition property that accounts for discounting, and the diffusion principle that accounts for higher recall of inconsistent information. These properties are able to explain most of our simulations of social judgment and behavior. In contrast, introductory textbooks on the autoassociator (e.g., McClelland & Rumelhart, 1988; McLeod et al., 1998) emphasize other capacities of the autoassociator, including its content-addressable memory, its ability to do pattern completion (see also Smith & DeCoster, 1998), and its fault and noise tolerance.

### Implications

What are the implications of this work for theories of impression formation? The key contribution of this article is that a wide range of phenomena was simulated with the same overall network model (differing

only in the learning rate parameter and the learning history), suggesting that these phenomena are based, at least during early processing, on the same fundamental information-processing principles. Providing a common framework for these different phenomena will hopefully generate further research and extend to new areas of social psychology usually seen as too different to be brought under a single theoretical heading. In addition, not only can this model account for prior empirical data, it can also generate new hypotheses that can be tested in a classical experimental setting. We briefly discuss some potential questions and research issues that emerge from this model.

**Knowledge acquisition.** To what extent is the learning history assumed in our simulations correct? What mechanisms and architectural considerations are necessary to preserve the network's knowledge base? How does prior (trait) knowledge interact with novel (behavioral) information? One of the suggestions made earlier is that semantic behavior–trait associations stored in semantic memory are spontaneously applied when novel behavioral information is received. This suggestion is in line with the bulk of research on spontaneous trait inferences (see Uleman, 1999). Perhaps answers to these questions can also be obtained by laboratory replications of the assumed learning histories that should reveal an equivalence with the prior knowledge of participants and similar effects on trait inferences as we have reviewed here.

**Automatic versus conscious reasoning.** Our approach does not make a clear and explicit distinction between automatic and conscious processing, or between implicit and explicit processing. Quite often, setting learning rate to a lower or higher (default) level made it possible to simulate this distinction, suggesting that automatic and conscious processing is perhaps, in part, a matter of slow or shallow versus fast and deep learning of information. This differential level of learning gives rise to a differential emphasis on, for example, prior information versus novel information and may result in different judgments. Some researchers (e.g., Smith & DeCoster, 2000) have proposed a distinction between two processing modes: a slow-learning (connectionist) pattern-completion mode and a more effortful (symbolic) mode that involves explicit symbolically represented rules and inferences. Other theorists have suggested, in line with our approach, that such sharp distinction is not necessary and that many social judgments—although differing in content—may share the same underlying process (e.g., Chun, Spiegel, & Kruglanski, 2002). Within the framework of a single connectionist network, differences between shallow and deep learning are possible by assuming that “explicit, conscious knowledge ... involves higher quality memory traces than implicit knowledge” (Cleeremans

& Jiménez, 2002, p. 21). This approach has also been taken to simulate differences between heuristic and central processes in attitude change (Van Overwalle & Siebler, 2002) and between implicit and explicit reasoning (Kinder & Shanks, 2001).

**Heuristics.** Although heuristics are typically viewed as rules-of-thumb that shortcut logical thinking and often result in biased judgments (Kahneman, Slovic, & Tversky, 1982), we argue that they actually provide a window to see how the brain—as a connectionist device—works. For example, the availability heuristic, invoked to explain why many judgments are biased by information about facts and arguments recently or frequently available in memory, can also be viewed from a connectionist framework as information that is recently primed or activated and that is spread automatically to other related concepts, influencing judgments about them (as we have seen in the assimilation simulation). Of course, this does not exclude the possibility that subjective experiences that may go together with activating memory traces, such as the ease to recall a given number of exemplars, may additionally influence how people utilize that activated information in further judgments (Schwarz et al., 1991). In addition, the representativeness heuristic that has been invoked to explain why categorization is often guided by a resemblance between concepts rather than by statistical base rates may in fact reflect the strength of the connections between a category and its members (as demonstrated in the simulation on the asymmetry of ability and morality related behavior). Finally, the anchoring and adjustment heuristic, originally proposed to explain why judgments are often biased toward an initial anchor, can be simply taken as a property of the delta learning algorithm, in which weight adjustments are often stronger initially (i.e., during anchoring) because of the greater error in the network, whereas later adjustments become increasingly smaller because the error is reduced. This approach might also explain insufficient adjustments in later phases of learning as due to decreased cognitive effort during integration of situational information.

### Limitations and Future Directions

Given the breadth of impression formation, we inevitably were not able to include many other interesting findings and phenomena. Perhaps the most interesting area omitted involves group processes. Connectionist modeling may well help to explain how group identity is created, how perceptions of group homogeneity is changed, how accentuation of correlated features is enhanced, and how illusory correlation and unrealistic negative stereotypes of minority groups are developed. These questions are addressed in Van Rooy et al. (2003), using the same model as ours. These ap-

plications merely reflect our current thinking and will almost certainly be replaced by improved models in the future. We believe, however, that the essence of the approach proposed here will survive.

Although we have attempted to show that a connectionist framework can potentially provide a parsimonious account of a number of disparate phenomena in impression formation, we are not suggesting that this is the only valid means of modeling social cognitive phenomena. On the contrary, we defend a multiple-view position in which connectionism would play a key role but would coexist alongside other viewpoints. We think that strict neurological reductionism is untenable, especially in personality and social psychology, where it is difficult to see how one could develop a connectionist model of such high-level abstract concepts as “need for closure,” “prejudice,” “close relationships,” “motivation,” and the like.

These limitations suggest a number of possible directions for extending the connectionist approach. First, a critical improvement to our recurrent network might be the inclusion of hidden layers (McClelland & Rumelhart, 1988, p. 121–126) possibly with coarse coding of nodes (e.g., O’Reilly & Rudy, 2001) or exemplar nodes (e.g., Kruschke & Johansen, 1999) that may potentially increase its power and capacity, for instance, to process nonlinear interactions.

Second, a more modular architecture will almost certainly be necessary to produce a better fit of the model to empirical data. For example, one severe limitation of most connectionist models is known as “catastrophic interference” (McCloskey & Cohen, 1989; Ratcliff, 1990, see French, 1999, for a review), which is the tendency of neural networks to forget, abruptly and completely, previously learned information in the presence of new input. Although catastrophic interference has been observed when perceivers process novel information (see Simulation 1), it is untenable for a realistic model of long-term social cognitive processes, whereby prior knowledge—such as stereotypes—is often resistant to change in the presence of new information. In response to such observations, it has been suggested that to overcome this problem, the brain developed a dual hippocampal–neocortical memory system in which new (mainly episodic) information is processed in the hippocampus and the old (mainly semantic) information is stored and consolidated in the neocortex (McClelland, McNaughton, & O’Reilly, 1995; Smith & DeCoster, 2000). Various modelers (Ans & Rousset, 1997; French, 1997) have proposed modular connectionist architectures mimicking this dual-memory system with one subsystem dedicated to the rapid learning of unexpected and novel information and the building of episodic memory traces and the other subsystem responsible for slow incremental learning of statistical regularities of the environment and gradual consolidation of information learned in the first subsystem.

There is considerable evidence for the modular nature of the brain, in particular for the complementary learning roles of hippocampal and neocortical structures (McClelland et al., 1995), the predominant role of the amygdala in social judgment and perception of emotions (Adolphs, Tranel, & Damasio, 1998), and so forth. A dual memory representation raises the intriguing possibility that for a limited period of time, old trait knowledge as well as novel trait inferences may coexist in memory. It strikes us that the next step in connectionist modeling of social cognition will involve the exploration of connectionist architectures built from separate but complementary systems.

Third, as a special case of modularization, it will ultimately be necessary to incorporate factors such as attention, consciousness, and motivation that are important in social perception into an improved model. For the time being, attentional aspects of human information processing are not part of the dynamics of our network (variations were simply hand-coded as differences in learning), which focuses almost exclusively on learning and pattern association. However, there are recent developments that provide insights in how an attentional switching mechanism might be implemented. O'Reilly and Munakata (2000) suggested a network model of attention and motivation based on the idea that a specialized module residing in the prefrontal cortex would be capable of active maintenance of rapidly updateable activation, which would enable this module to provide a sustained, top-down biasing influence over processing elsewhere in the system. These actively maintained frontal cortex representations could guide behavior and judgment according to goals, motivation, and other types of internal constraints.

## Conclusion

Connectionist modeling of person-impression formation fits seamlessly into a multilevel integrative analyses of human behavior (Cacioppo et al., 2000). Given that cognition is intrinsically social, connectionism will ultimately have to begin to incorporate social constraints into its models. On the other hand, social psychology will need to be more attentive to the biological underpinnings of social behavior. Social and biological approaches to cognition can therefore be seen as complementary endeavors with the common goal of achieving a clearer and deeper understanding of human behavior. We hope that connectionist accounts of social cognition will provide the common ground for this exploration.

## References

- Adolphs, R., & Damasio, A. (2001). The interaction of affect and cognition: A neurobiological perspective. In J. P. Forgas (Ed.), *Handbook of affect and social cognition* (pp. 27–49). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Adolphs, R., Tranel, D., & Damasio, A. (1998). The human amygdala in social judgment. *Nature*, *393*, 470–474.
- 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, J. R. (1976). *Language, memory and thought*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Anderson, N. H. (1979). Serial position curves in impression formation. *Journal of Experimental Psychology*, *97*, 8–12.
- Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic.
- Anderson, N. H., & Farkas, A. J. (1973). New light on order effect in attitude change. *Journal of Personality and Social Psychology*, *28*, 88–93.
- Ans, B., & Rousset, S. (1997). Avoiding catastrophic forgetting by coupling two reverberating neural networks. *Académie des Sciences de la vie*, *320*, 989–997.
- Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, *41*, 258–290.
- Asch, S. E., & Zukier, H. (1984). Thinking about persons. *Journal of Personality and Social Psychology*, *46*, 1230–1240.
- Baker, A. G., Barbier, M. W., & Vallée-Tourangeau, F. (1989). Judgments of a 2 × 2 contingency table: Sequential processing and the learning curve. *The Quarterly Journal of Experimental Psychology*, *41B*, 65–97.
- Bargh, J. A., & Thein, R. D. (1985). Individual construct accessibility, person memory, and the recall-judgment link: The case of information overload. *Journal of Personality and Social Psychology*, *49*, 1129–1146.
- Betsch, T., Plessner, H., Schwieren, C., & Gütig, R. (2001). I like it but I don't know why: A value-account approach to implicit attitude formation. *Personality and Social Psychology Bulletin*, *27*, 242–253.
- Busemeyer, J. R. (1991). Intuitive statistical estimation. In N. Anderson (Ed.), *Contributions to information integration theory: Vol. 1. Cognition* (pp. 189–215). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Busemeyer, J. R., & Myung, I. J. (1988). A new method for investigating prototype learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 3–11.
- Cacioppo, J. T., Berntson, G. G., Sheridan, J. F., & McClintock, M. K. (2000). Multilevel integrative analyses of human behavior: Social neuroscience and the complementing nature of social and biological approaches. *Psychological Bulletin*, *126*, 829–843.
- Cantor, N., & Mischel, W. (1977). Traits as prototypes: Effects on recognition memory. *Journal of Personality and Social Psychology*, *35*, 38–48.
- Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory and Cognition*, *18*, 537–545.
- Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*, 365–382.
- Chun, W. Y., Spiegel, S., & Kruglanski, A. W. (2002). Assimilative behavior identification can also be resource dependent: The unimodel perspective on personal-attribution phases. *Journal of Personality and Social Psychology*, *83*, 542–555.
- Cleeremans, A., & Jiménez, L. (2002). Implicit learning and consciousness: A graded, dynamic perspective. In R. M. French & A. Cleeremans (Eds.), *Implicit learning and consciousness: An empirical, philosophical and computational consensus in the making* (pp. 1–40). East Sussex, England: Psychology Press.
- Dreben, E. K., Fiske, S. T., & Hastie, R. (1979). The independence of evaluative and item information: Impression and recall order effects in behavior-based impression formation. *Journal of Personality and Social Psychology*, *37*, 1758–1768.
- Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. San Diego, CA: Harcourt Brace.

- Ebbesen, E. B., & Bowers, R. J. (1974). Proportion of risky to conservative arguments in a group discussion and choice shifts. *Journal of Personality and Social Psychology, 29*, 316–327.
- Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environment. *Journal of Personality and Social Psychology, 103*, 193–214.
- Fiedler, K., Walther, E., & Nickel, S. (1999). The auto-verification of social hypotheses: Stereotyping and the power of sample size. *Journal of Personality and Social Psychology, 77*, 5–18.
- Försterling, F. (1992). The Kelley model as an analysis of variance analogy: How far can it be taken? *Journal of Experimental Social Psychology, 28*, 475–490.
- French, R. (1997). Pseudo-recurrent connectionist networks: An approach to the “sensitivity–stability” dilemma. *Connection Science, 9*, 353–379.
- French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences, 3*, 128–135.
- Freund, T., Kruglanski, A. W., & Shpitzajzen, A. (1985). The freezing and unfreezing of impressionality: Effects of the need for structure and the fear of invalidity. *Personality and Social Psychology Bulletin, 11*, 479–487.
- Fyock, J., & Stangor, C. (1994). The role of memory biases in stereotype maintenance. *British Journal of Social Psychology, 33*, 331–343.
- Gannon, K. M., Skowronski, J. J., & Betz, A. L. (1994). Depressive diligence in social information processing: Implications for order effects in impressions and for social memory. *Social Cognition, 12*, 263–280.
- Gilbert, D. T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thoughts: Limits of awareness, intention, and control* (pp. 189–211). New York: Guilford.
- Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*, 21–38.
- 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.
- Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227–247.
- Hamilton, D. L., Driscoll, D. M., & Worth, L. T. (1989). Cognitive organization of impressions: Effects of incongruency in complex representations. *Journal of Personality and Social Psychology, 56*, 925–939.
- Hamilton, D. L., Katz, L. B., Leirer, V. O. (1980). Cognitive representation of personality impressions: Organizational processes in first impression formation. *Journal of Personality and Social Psychology, 39*, 1050–1063.
- Hansen, R. D., & Hall, C. A. (1985). Discounting and augmenting facilitative and inhibitory forces: The winner takes all. *Journal of Personality and Social Psychology, 49*, 1482–1493.
- Hastie, R. (1980). Memory for behavioral information that confirms or contradicts a personality impression. In R. Hastie, T. M. Ostrom, E. B. Ebbesen, R. S. Wyer, D. L. Hamilton, & D. E. Carlston (Eds.), *Person memory: The cognitive basis of social perception* (pp. 155–177). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Hastie, R., & Kumar, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology, 37*, 25–38.
- Heaton, A. W., & Kruglanski, A. W. (1991). Person perception by introverts and extraverts under time pressure: Effects of need for closure. *Personality and Social Psychology Bulletin, 17*, 161–165.
- Hintzmann, D. L. (1986). “Schema abstraction” in a multi-trace memory model. *Psychological Review, 93*, 411–428.
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1–55.
- Ito, T. A., & Cacioppo, J. T. (2001). Affect and attitudes: A social neuroscience approach. In J. P. Forgas (Ed.), *Handbook of affect and social cognition* (pp. 50–74). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgments under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.
- Kashima, Y., & Kerekes, A. R. Z. (1994). A distributed memory model of averaging phenomena in person impression formation. *Journal of Experimental Social Psychology, 30*, 407–455.
- Kashima, Y., Woolcock, J., & Kashima, E. S. (2000). Group impression as dynamic configurations: The tensor product model of group impression formation and change. *Psychological Review, 107*, 914–942.
- Kelley, H. H. (1967). Attribution in social psychology. *Nebraska Symposium on Motivation, 15*, 192–238.
- Kinder, A., & Shanks, D. R. (2001). Amnesia and the declarative/nondeclarative distinction: A recurrent network model of classification, recognition, and repetition priming. *Journal of Cognitive Neuroscience, 13*, 648–669.
- Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay inferences: Effects on impressionality, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology, 19*, 448–468.
- Kruglanski, A. W., Schwartz, S. M., Maides, S., & Hamel, I. Z. (1978). Covariation, discounting, and augmentation: Towards a clarification of attributional principles. *Journal of Personality, 76*, 176–189.
- Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1083–1119.
- Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*, 284–308.
- LaBerge, D. (1997). Attention, awareness and the triangular circuit. *Consciousness and Cognition, 6*, 149–181.
- LaBerge, D. (2000). Networks of attention. In M. S. Gazzaniga (Ed.), *The new cognitive neuroscience* (pp. 711–724). Cambridge, MA: MIT Press.
- Lupfer, M. B., Weeks, M., & Dupuis, S. (2000). How pervasive is the negativity bias in judgments based on character appraisal? *Personality and Social Psychology Bulletin, 26*, 1353–1366.
- Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology, 23*, 77–87.
- Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. *Journal of Personality and Social Psychology, 38*, 231–248.
- McClelland, J., McNaughton, B., & O’Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and the failures of connectionist models of learning and memory. *Psychological Review, 102*, 419–457.
- McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology, 114*, 159–188.
- McClelland, J. M., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs and exercises*. Cambridge, MA: Bradford.
- McCloskey, M., & Cohen N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation, 24*, 109–165.
- McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to connectionist modeling of cognitive processes*. Oxford, England: Oxford University Press.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.
- Moskowitz, G. B., & Skurnik, I. W. (1999). Contrast effects as determined by the type of prime: Trait versus exemplar primes initiate processing strategies that differ in how accessible constructs are used. *Journal of Personality and Social Psychology, 76*, 911–927.

- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57.
- Nosofsky, R. M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory and Cognition*, 19, 131–150.
- Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18, 211–233.
- O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.
- O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, 108, 311–345.
- Pandelaere, M., & Hoorens, V. (2002). *The role of behavior categorization in act frequency estimation processes*. Manuscript submitted for publication.
- Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, S., Gatenby, C., Gore, J. C., et al. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, 12, 729–738.
- Posner, M. I. (1992). Attention as a cognitive and neural system. *Current Directions in Psychological Science*, 1, 11–14.
- Queller, S., & Smith, E. (2002). Subtyping versus bookkeeping in stereotype learning and change: Connectionist simulations and empirical findings. *Journal of Personality and Social Psychology*, 82, 300–313.
- Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97, 285–308.
- Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65, 429–447.
- Read, S. J., & Montoya, J. A. (1999). An autoassociative model of causal reasoning and causal learning: Reply to Van Overwalle's critique of Read and Marcus-Newhall (1993). *Journal of Personality and Social Psychology*, 76, 728–742.
- Reeder, G. D. (1997). Dispositional inferences of ability: Content and process. *Journal of Experimental Social Psychology*, 33, 171–189.
- 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.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–98). New York: Appleton-Century-Crofts.
- Rosch, E. H. (1978). Principles of categorization. In E. H. Rosch & B. B. Lloyds (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Rosenfield, D., & Stephan, W. G. (1977). When discounting fails: An unexpected finding. *Memory and Cognition*, 5, 97–102.
- Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations*. Cambridge, MA: MIT Press.
- Sarle, W. S. (1994). *Neural networks and statistical models*. Proceedings of the nineteenth annual SAS users group international conference.
- Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, 61, 195–202.
- Shanks, D. R. (1985). Forward and backward blocking in human contingency judgment. *Quarterly Journal of Experimental Psychology*, 37b, 1–21.
- Shanks, D. R. (1987). Acquisition functions in contingency judgment. *Learning and Motivation*, 18, 147–166.
- Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology*, 48a, 257–279.
- Shanks, D. R., Lopez, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation* (Vol. 34, pp. 265–311). New York: Academic.
- Shultz, T., & Lepper, M. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review*, 2, 219–240.
- Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity and extremity biases. *Journal of Personality and Social Psychology*, 52, 689–699.
- Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, 105, 131–142.
- Skowronski, J. J., & Gannon, K. (2000). Raw conditional probabilities are a flawed index of associative strength: Evidence from a single trait expectancy paradigm. *Basic and Applied Social Psychology*, 22, 9–18.
- Skowronski, J. J., & Welbourne, J. (1997). Conditional probability may be a flawed measure of associative strength. *Social Cognition*, 15, 1–12.
- 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. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: Simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology*, 74, 21–35.
- Smith, E. R., & DeCoster, J. (2000). Associative and rule-based processing: A connectionist interpretation of dual-process models. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 323–338). London, England: Guilford.
- Smith, E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, 99, 3–21.
- Srull, T. K. (1981). Person memory: Some tests of associative storage and retrieval models. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 440–463.
- Srull, T. K., Lichtenstein, M., & Rothbart, M. (1985). Associative storage and retrieval processes in person memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 316–345.
- Stangor, C., & Duan, C. (1991). Effects of multiple task demands upon memory for information about social groups. *Journal of Experimental Social Psychology*, 27, 357–378.
- Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin*, 111, 42–61.
- Stapel, D. A., Koomen, W., & van der Pligt, J. (1997). Categories of category accessibility: The impact of trait concept versus exemplar priming on person judgments. *Journal of Experimental Social Psychology*, 33, 47–76.
- Stewart, R. H. (1965). Effect of continuous responding on the order effect in personality impression formation. *Journal of Personality and Social Psychology*, 1, 161–165.
- Thorpe, S. (1994). Localized versus distributed representations. In M. A. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 949–952). Cambridge, MA: MIT Press.

- Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review*, *93*, 239–257.
- Trope, Y., & Gaunt, R. (2000). Processing alternative explanations of behavior: Correction or integration? *Journal of Personality and Social Psychology*, *79*, 344–354.
- Uleman, J. S. (1999). Spontaneous versus intentional inferences in impression formation. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 141–160). New York: Guilford.
- Van Overwalle, F. (1996). The relationship between the Rescorla–Wagner associative model and the probabilistic joint model of causality. *Psychologica Belgica*, *36*, 171–192.
- Van Overwalle, F. (1997). Dispositional attributions require the joint application of the methods of difference and agreement. *Personality and Social Psychology Bulletin*, *23*, 974–980.
- Van Overwalle, F. (1998). Causal explanation as constraint satisfaction: A critique and a feedforward connectionist alternative. *Journal of Personality and Social Psychology*, *74*, 312–328.
- Van Overwalle, F. (2001). *Discounting and augmentation of dispositional and causal attributions*. Manuscript submitted for publication.
- Van Overwalle, F. (2003). Acquisition of dispositional attributions: Effects of sample size and covariation. *European Journal of Social Psychology*, *33*, 515–533.
- Van Overwalle, F., & Jordens, K. (2002). An adaptive connectionist model of cognitive dissonance. *Personality and Social Psychology Review*, *3*, 204–231.
- Van Overwalle, F., & Siebler, F. (2002). *A connectionist model of attitude formation and change*. Manuscript submitted for publication.
- Van Overwalle, F., & Van Rooy, D. (1998). A connectionist approach to causal attribution. In S. J. Read & L. C. Miller (Eds.), *Connectionist models of social reasoning and social behavior* (pp. 143–171). New York: Lawrence Erlbaum Associates, Inc.
- Van Overwalle, F., & Van Rooy, D. (2001a). How one cause discounts or augments another: A connectionist account of causal competition. *Personality and Social Psychology Bulletin*, *27*, 1613–1626.
- Van Overwalle, F., & Van Rooy, D. (2001b). When more observations are better than less: A connectionist account of the acquisition of causal strength. *European Journal of Social Psychology*, *31*, 155–175.
- Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C., & French, R. (2003). A recurrent connectionist model of group biases. *Psychological Review*, *110*, 536–563.
- Wasserman, E. A., Kao, S.-F., Van Hamme, L., Katagiri, M., & Young, M. E. (1996). Causation and association. *The Psychology of Learning and Motivation*, *34*, 207–264.
- Wells, G. L., & Ronis, D. L. (1982). Discounting and augmentation: Is there something special about the number of causes? *Personality and Social Psychology Bulletin*, *8*, 566–572.
- Wojciszke, B., Brycz, H., & Borkenau, P. (1993). Effects of information content and evaluative extremity on positivity and negativity biases. *Journal of Personality and Social Psychology*, *64*, 327–335.

## Appendix A

### The Linear Autoassociative Model

In an autoassociative network, features and categories, or causes and outcomes, are represented in nodes that are all interconnected. Processing information in this model takes place in two phases. In the first phase, the activation of the nodes is computed, and in the sec-

ond phase, the weights of the connections are updated (see also McClelland & Rumelhart, 1988).

**Node activation.** During the first phase of information processing, each node in the network receives activation from external sources. Because the nodes are all interconnected, this activation is then spread throughout the network, where it influences all other nodes. The activation coming from the other nodes is called the internal input. Together with the external input, this internal input determines the final pattern of activation of the nodes, which reflects the short-term memory of the network.

In mathematical terms, every node  $i$  in the network receives external input, termed  $ext_i$ . In the autoassociative model, every node  $i$  also receives internal input  $int_i$ , which is the sum of the activation from the other nodes  $j$  (denoted by  $a_j$ ) in proportion to the weight of their connection, or

$$int_i = \sum (a_j \times w_{ij}) \quad (1)$$

for all  $j \neq i$ . Typically, activations and weights range between  $-1$  to  $+1$ . The external input and internal input are then summed to the net input, or

$$net_i = E \times ext_i + I \times int_i \quad (2)$$

where  $E$  and  $I$  reflect the degree to which the net input is determined by the external and internal input, respectively. Typically, in a recurrent network, the activation of each node  $i$  is updated during a number of cycles until it eventually converges to a stable pattern that reflects the network's short-term memory. According to the linear activation algorithm, the updating of activation is governed by the following equation:

$$\Delta a_i = net_i - D \times a_i \quad (3)$$

where  $D$  reflects a memory decay term. In our simulations, we used only one internal updating cycle and the parameter values  $D = I = E = 1$ . Given these simplifying assumptions, the final activation of node  $i$  reduces simply to the sum of the external and internal input, or:

$$a_i = net_i = ext_i + int_i \quad (3')$$

**Weight updating.** After this first phase, the autoassociative model then enters in its second learning phase, whereby the short-term activation is consolidated in long-term weight changes to better represent and anticipate future external input. Basically, weight changes are driven by the discrepancy between the internal input from the last updating cycle of the network and the external input received from outside sources,

formally expressed in the delta algorithm (McClelland & Rumelhart, 1988, p. 166):

$$\Delta w_{ij} = \varepsilon(ext_i - int_i)a_j \quad (4)$$

where  $w_{ij}$  is the weight of the connection from node  $j$  to  $i$ ,  $\varepsilon$  is a learning rate that determines how fast the network learns, and  $a_j$  is the current final activation of node  $j$ .

The presence of a feature or a category was typically encoded by setting the external input to +1 and -1 for opposite features or categories (lower values were also used; see appropriate tables); otherwise the external activation remained at resting level zero. The weights of the connections were updated after each trial. At the end of each simulation, the judgment of interest was tested by turning on the external input of the appropriate nodes and reading off the resulting activation of the nodes that represent the judgment of interest (see appropriate tables).

## Appendix B

### Anderson's Averaging Rule and the Delta Algorithm

This appendix demonstrates that the delta algorithm converges at asymptote to Anderson's (1981) averaging rule under two conditions. First, learning must have reached asymptote (i.e., after sufficient trials), and second, the relative weights in Anderson's model can be represented by the relative frequencies of person-trait pairings. Anderson's averaging rule of impression formation expresses a rating about a person as:

$$\text{rating} = \frac{\sum \omega_i s_i}{\sum \omega_i} \quad (5)$$

where  $\omega_i$  represents the weights and  $s_i$  the scale values of the trait.

This proof uses the same logic as Chapman and Robbins (1990) in their demonstration that the delta algorithm converges to the probabilistic expression of covariation. In line with the conventional representation of covariation information, person-impression information can be represented in a contingency table with two cells. Cell  $a$  represents all cases in which the actor is ascribed a focal trait, whereas cell  $b$  represents all cases in which the actor is ascribed the opposite trait. For simplicity, we use only two trait categories, although this proof can easily be extended to more categories.

In a recurrent connectionist architecture with localist encoding as used in the text, the target person  $j$  and the trait categories  $i$  are each represented by a node, which are connected by adjustable weights  $w_{ij}$ . When the target person is present, its corresponding

node receives external activation, and this activation is spread to each trait node. We assume that the overall activation received at the trait nodes  $i$  (or internal activation) after priming the person node reflects the impression on the person.

According to the delta algorithm in Equation 4, the weights  $w_{ij}$  are adjusted proportional to the error between the actual trait category (represented by its external activation  $ext$ ) and the trait category as predicted by the network (represented by its internal activation  $int$ ). If we substitute in Equation 4  $ext$  by Anderson's scale values ( $s_1$  for the focal trait and  $s_2$  for the opposite trait) and if we take the default activation for  $a_j$  (which is 1), then the following equations can be constructed for the two cells in the contingency table:

$$\text{For the a cell: } \Delta w_i = \varepsilon(s_1 - int) \quad (6)$$

$$\text{For the b cell: } \Delta w_i = \varepsilon(s_2 - int) \quad (7)$$

The change in overall impression is the sum of Equations 6 and 7, weighted for the corresponding frequencies  $a$  and  $b$ , in the two cells, or:

$$\Delta w_i = a[\varepsilon(s_1 - int)] + b[\varepsilon(s_2 - int)] \quad (8)$$

These adjustments will continue until asymptote, that is, until the error between actual and expected category is zero. This implies that at asymptote, the changes will become zero, or  $\Delta w_i = 0$ . Consequently, Equation 8 becomes:

$$\begin{aligned} 0 &= a[\varepsilon(s_1 - int)] + b[\varepsilon(s_2 - int)] \\ &= a[s_1 - int] + b[s_2 - int] \\ &= [a \times s_1 + b \times s_2] - [a + b] \times int \end{aligned}$$

so that:

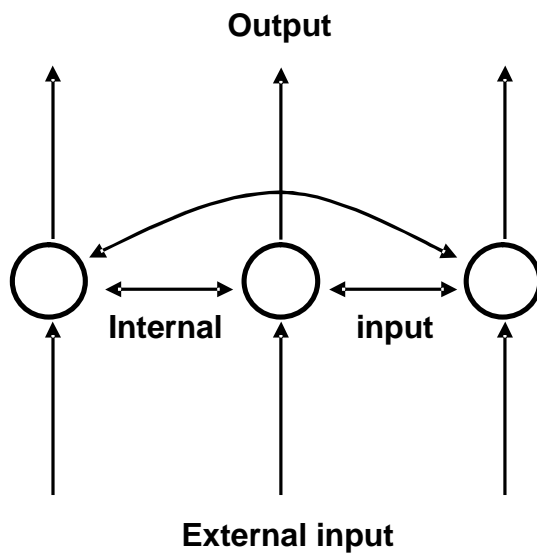
$$int = [a \times s_1 + b \times s_2] / [a + b]$$

Because the internal activation of the trait nodes reflects the trait impression on the person, this can be rewritten in Anderson's terms as:

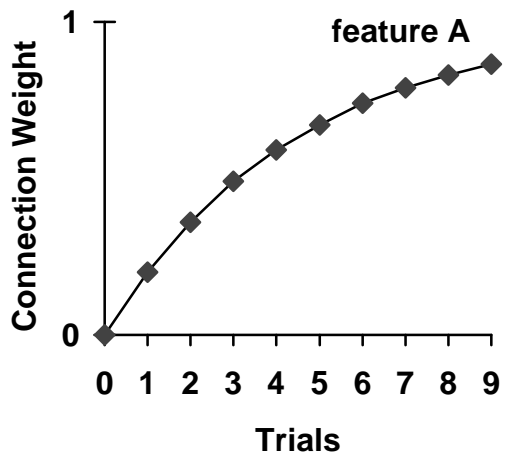
$$\text{impression} = \frac{\sum f_i s_i}{\sum f_i} \quad (9)$$

where  $f$  represents frequencies with which a person and the traits co-occur. As can be seen, Equation 9 has the same format as Equation 5. This demonstrates that the delta algorithm predicts a weighted averaging function at asymptote for making overall impression judgments, where Anderson's weights  $\omega$  are determined by the frequencies by which person and traits are presented together.

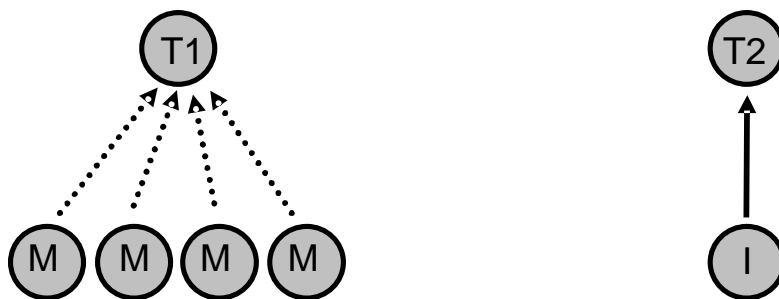
Figure 1



**A. Acquisition**



**B. Competition**



**C. Diffusion**

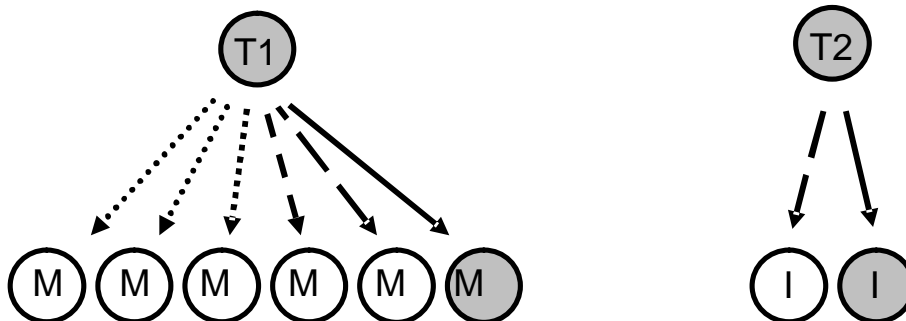
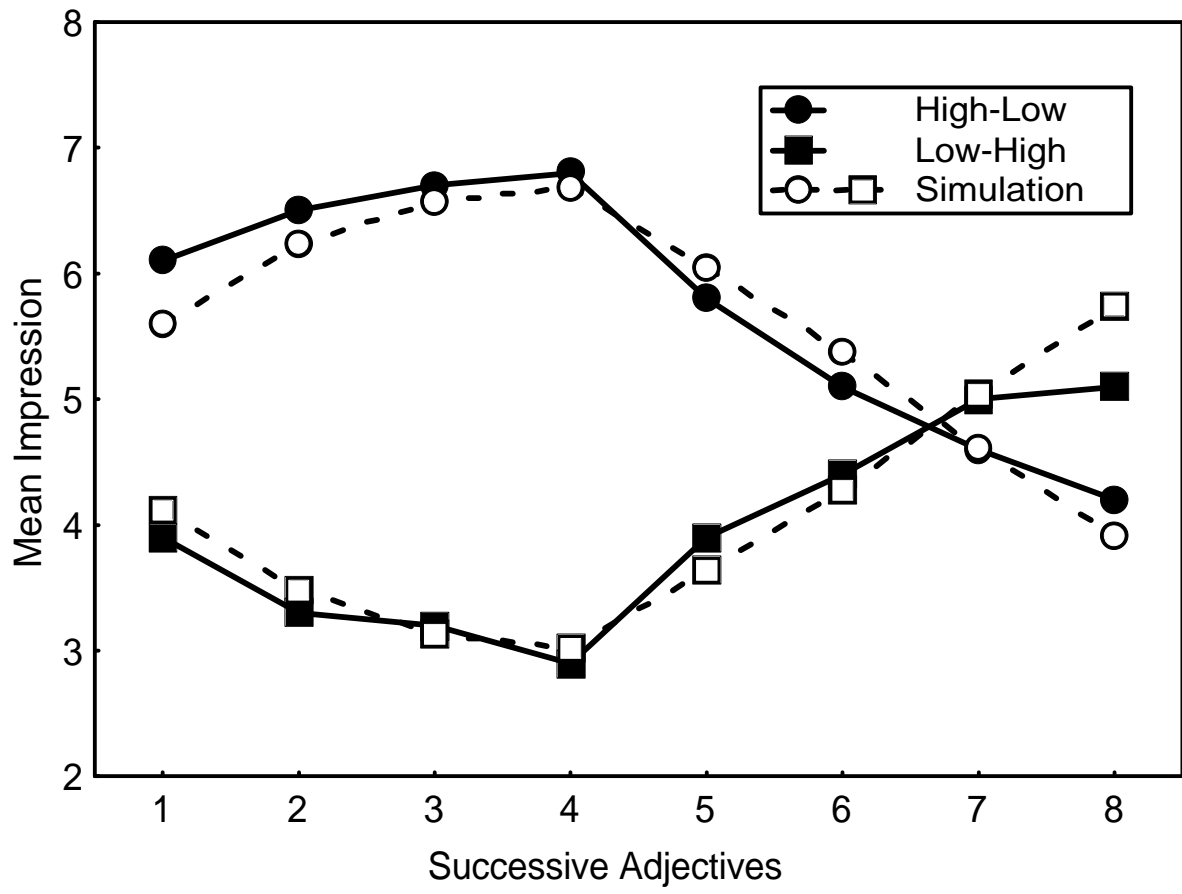


Figure 3



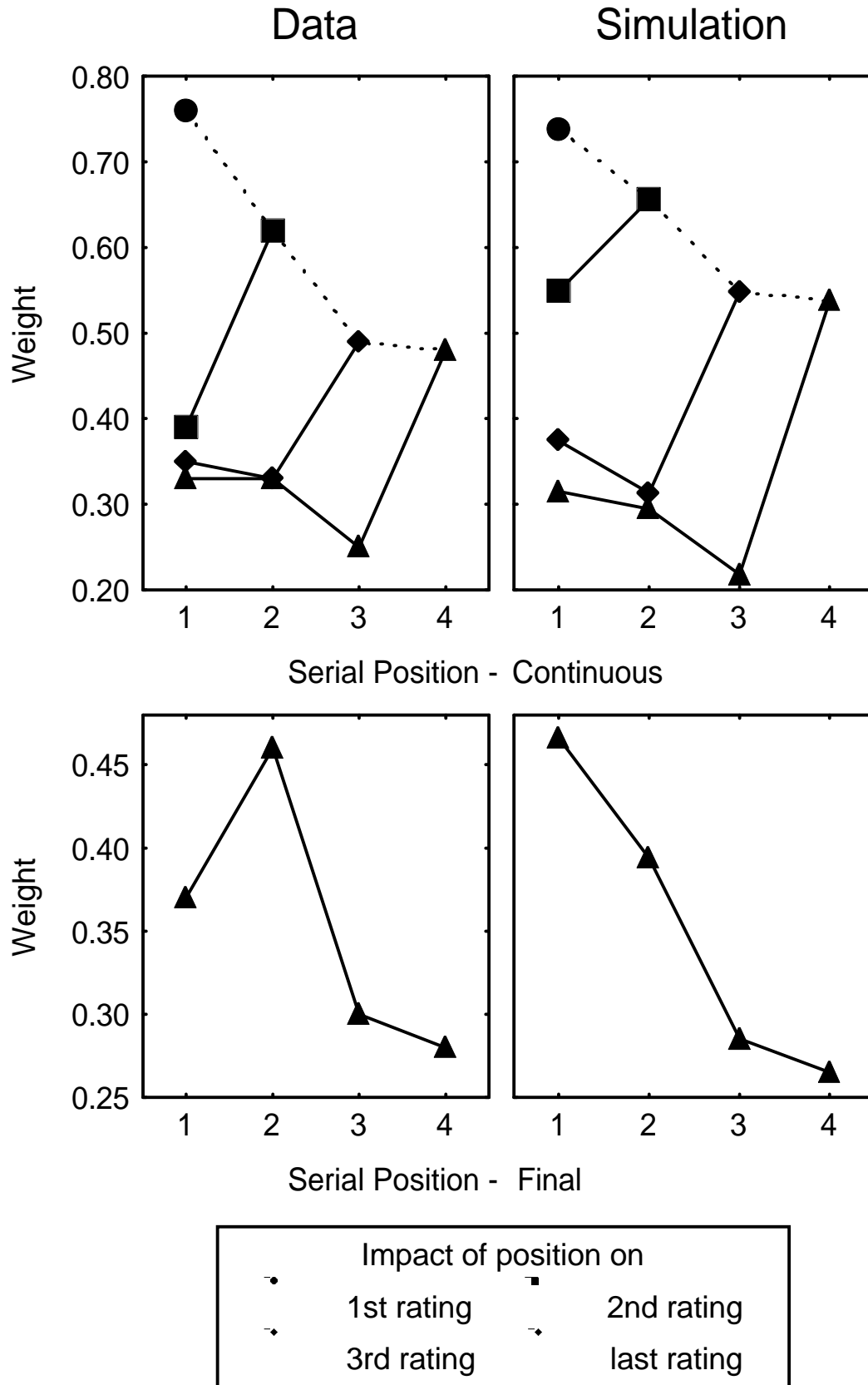


Figure 5

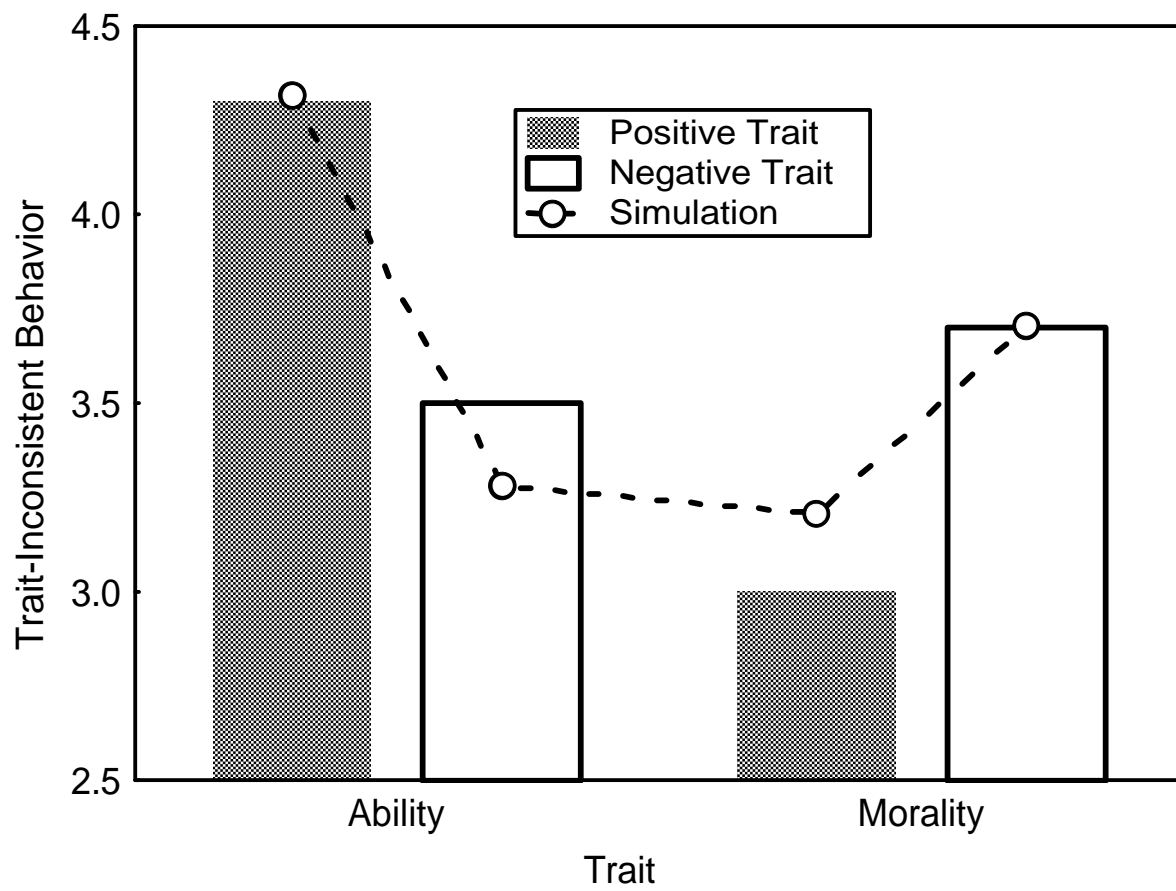


Figure 6

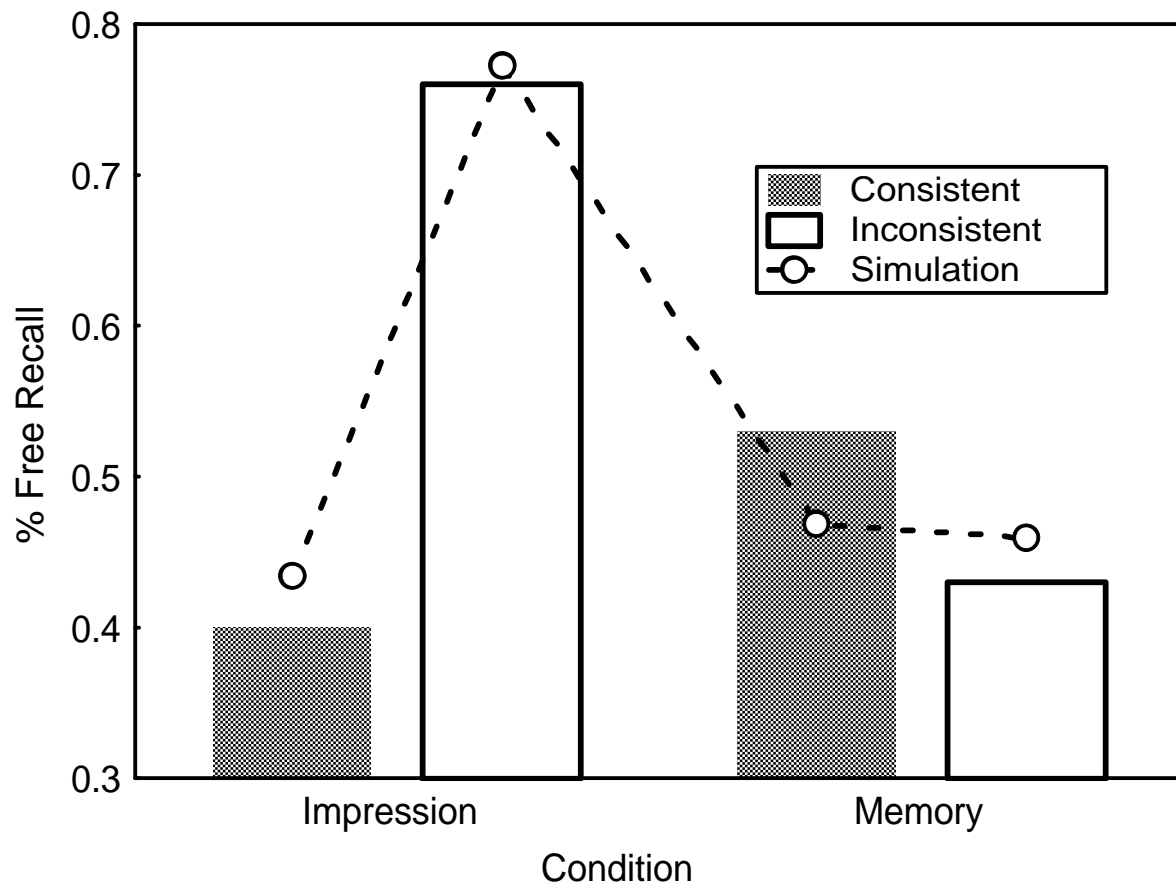


Figure 7

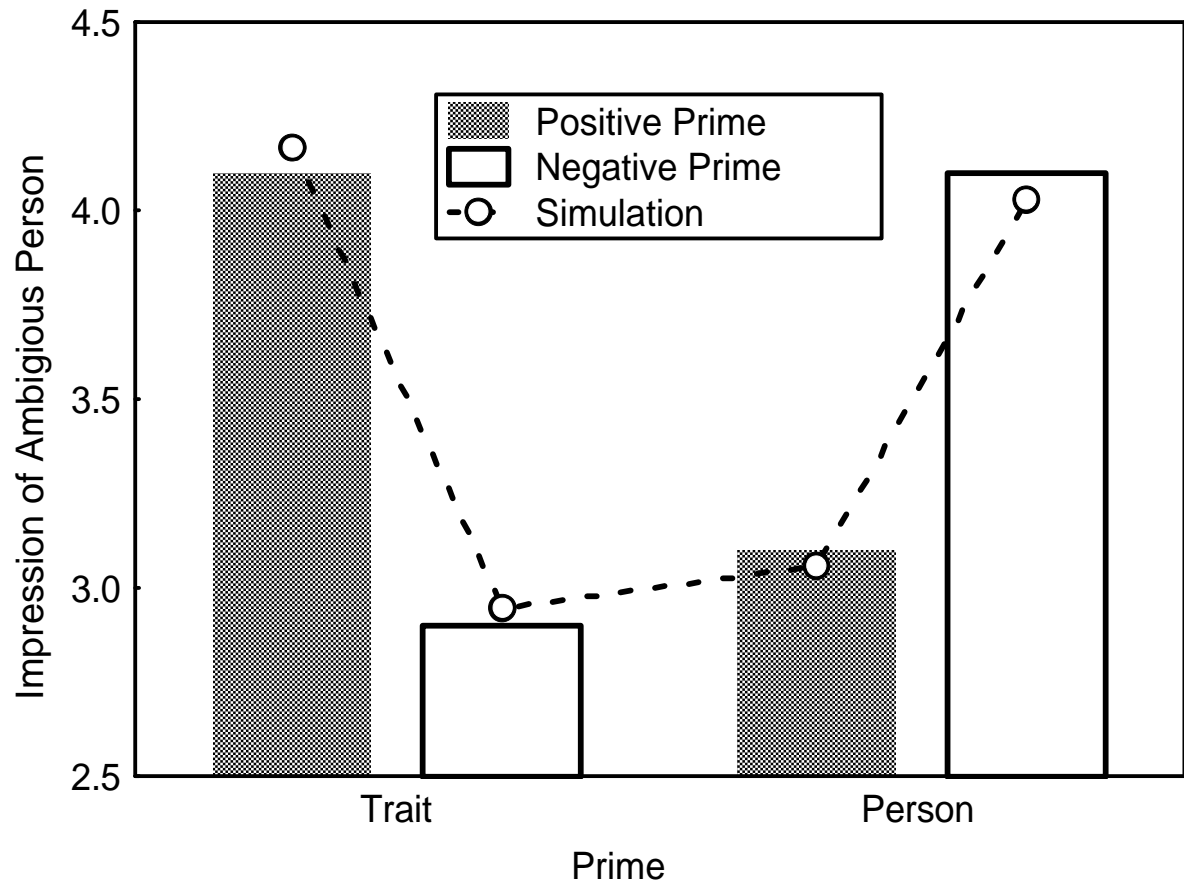


Figure 8

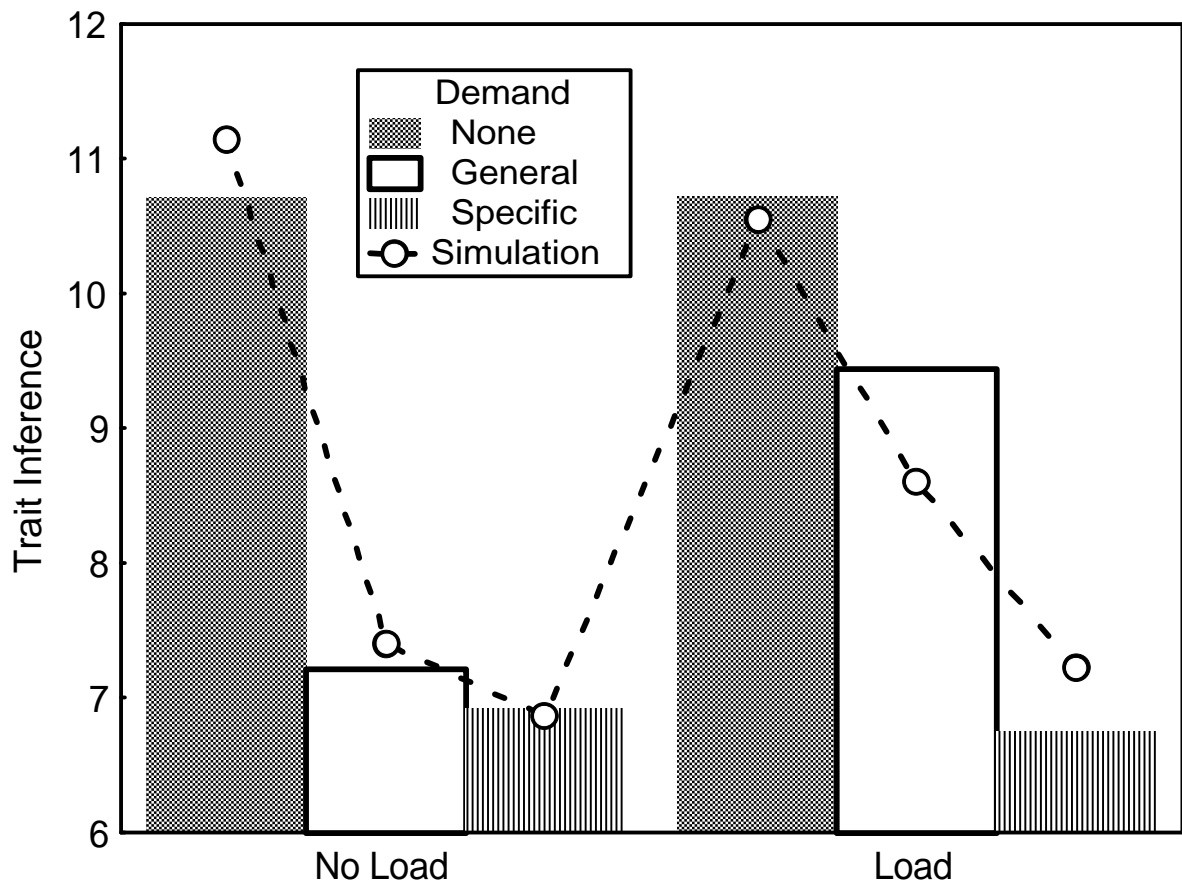


Figure 9

