

An Adaptive Connectionist Model of Cognitive Dissonance

Frank Van Overwalle and Karen Jordens

Vrije Universiteit Brussel, Belgium

This article proposes an adaptive connectionist model that implements an attributional account of cognitive dissonance. The model represents an attitude as the connection between the attitude object and behavioral-affective outcomes. Dissonance arises when circumstantial constraints induce a mismatch between the model's (mental) prediction and discrepant behavior or affect. Reduction of dissonance by attitude change is accomplished through long-lasting changes in the connection weights using the error-correcting delta learning algorithm. The model can explain both the typical effects predicted by dissonance theory as well as some atypical effects (i.e., reinforcement effect), using this principle of weight changes and by giving a prominent role to affective experiences. The model was implemented in a standard feedforward connectionist network. Computer simulations showed an adequate fit with several classical dissonance paradigms (inhibition, initiation, forced compliance, free choice, & misattribution), as well as novel studies that underscore the role of affect. A comparison with an earlier constraint satisfaction approach (Shultz & Lepper, 1996) indicates that the feedforward implementation provides a similar fit with these human data, while avoiding a number of shortcomings of this previous model.

More than 40 years ago, Festinger (1957) developed a theory of cognitive dissonance that became one of the most influential models in social psychology (Jones, 1985). Cognitive dissonance arises when there are inconsistencies between cognitions or elements of knowledge that people have about oneself, one's behavior, or the environment. This cognitive inconsistency generates psychological discomfort that motivates people to reduce it, for instance, by changing their beliefs, attitudes, or behavior. After Festinger's original formulation, numerous revisions or alternatives to cognitive dissonance theory have been advanced (see Harmon-Jones & Mills, 1999). Some revisions, like self-perception theory (Bem, 1972) and the attributional reformulation (Cooper & Fazio, 1984) propose that dissonance reduction is driven by people's attributions for their discrepant behavior and the situation in which it occurs. When no situational attribution can be made, people assume that their behavior reflects their true attitude. As a result, they change their attitude to attain consistency between their behavior and their attitude. Other, more recent revisions like self-consistency theory (e.g., Aronson, 1968) and self-affirmation theory (e.g., Steele, 1988) focus on the central role of the self in the cognitive dissonance process (see also Stone & Cooper, 2001).

Recently, a number of computational models have been formulated to account for dissonance phenomena (Sakai, 1999; Shultz & Lepper, 1996). For instance, Shultz and Lepper presented the consonance model, a constraint satisfaction connectionist model that reflects a person's representation of the experimental situation in which dissonance is aroused. In this model, cognitions about the discrepant behavior, justification, and evaluation are represented in separate nodes, and connection weights denote the causal implications between the cognitions, much like in automatic spreading activation models. Shultz and Lepper's novel contribution was that the consonance model can reach consistency automatically through the simultaneous satisfaction of multiple constraints imposed by the connections. However, an important limitation is that the connections themselves are not dynamically learned, but handset by the authors based on available evidence.

The aim of this article is to further advance connectionist modeling of cognitive dissonance by presenting an alternative connectionist model in which the connections between cognitions are automatically developed, without intervention from the experimenter. The idea that the connections are developed and adjusted by the model itself makes the present approach drastically different from the consonant model and involves an entirely different set of basic assumptions on how the mind works.

Constraint satisfaction models reflect a view of the mind as a mechanism that maintains some equilibrium,

We are grateful to Dirk Van Rooy and Christophe Labiouse for their helpful suggestions on earlier versions 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

and cognitive dissonance is seen basically as a process of rationalizing someone's discrepant decisions and behaviors. In contrast, our connectionist approach reflects a view of the mind as an adaptive learning mechanism, where cognitive dissonance is seen as a relatively rational process in which people seek causal answers for why they think, feel or behave inconsistently. These answers, we assume, drive the development of dissonance and people's attempts to reduce it. The ability to learn puts our approach in general agreement with evolutionary pressures that shaped the brain and that allowed for more flexible responses to the demands of the environment. Thus, cognitive dissonance emerges from general cognitive processes that are otherwise quite adaptive.

In addition, the present approach has a higher degree of neurological plausibility that is absent in the consonant model. Although it is true that connectionist models are highly simplified versions of real neurological circuitry and processing, it is commonly assumed that they reveal a number of emergent processing properties that real human brains also exhibit. One of the most typical properties of adaptive models is the integration of long-term memory (i.e., adaptation of connection weights) and short-term memory (i.e., activation in the network). There is no clear separation between learning and processing, as there is in the consonant model.

Our basic idea that cognitive dissonance reduction is driven by a rational process in which the causal understanding of thoughts, feelings, and behaviors plays a major role, is largely inspired by the attributional reformulation advocated by Cooper and Fazio (1984), although we diverge from them on some central points. We first briefly discuss this attributional approach and then present our connectionist formulation.

An Attributional Approach

Cooper and Fazio (1984) posited that dissonant behavior creates negative arousal, and that this arousal motivates a causal search for the nature of the emotion and its cause. Individuals try to understand and justify their discrepant behavior ("Why did I behave this way?") and their concurrent feelings ("Why do I feel this way?"). Cooper and Fazio argued that when the discrepant behavior is attributed to one's own responsibility, then pressure to change one's attitudes occurs. In contrast, when external demands (e.g., payment or threat by the experimenter) provide sufficient justification for engaging in the dissonant behavior, then dissonance reduction will not occur. For example, when the arousal is mistakenly attributed to some external source (e.g., a placebo pill), no need to modify one's attitude is experienced (e.g., Zanna & Cooper, 1974). Cooper and Fazio therefore

concluded that "dissonance has precious little to do with the inconsistency among cognitions per se, but rather with the production of a consequence that is unwanted" (p. 234).

We concur with Cooper and Fazio (1984) that people's attempt to causally understand and justify their dissonant behavior and emotions is at the root of the creation and reduction of dissonance. However, our model differs in a number of respects. As we see shortly, we view the attributions to the attitude object as central rather than attributions of one's responsibility, we emphasize the role of affect during dissonance and neglect arousal, and we focus on unexpected outcomes rather than unwanted outcomes.

To represent dissonant cognitions, behaviors, and emotions in the network, we follow the three-component view on attitudes (Rosenberg & Hovland, 1960). Specifically, we define an attitude as manifesting itself through its causal connections in memory between the cognitive representation or belief about the attitude object and two types of responses: the behavioral tendencies that characterize the interaction with the attitude object and feelings about this interaction (Ostrom, Skowronski, & Nowak, 1994). The intensity of an attitude is defined by the strength of these connections. This makes sense intuitively, because these connections reflect to what extent the attitude object causes a person to approach or avoid the object and to feel positive or negative about this (e.g., "This toy looks so attractive that it must be fun playing with it"). Consequently, increasing or decreasing one's causal attribution (i.e., connection) for discrepant behavior or affect to the attitude object is equivalent to increasing or decreasing the attitude itself. This view implies that attitudes can be changed by inducing people to change their habitual behavior or affect, as is typically the case in dissonance experiments. Given that these experiments are concerned mainly with discrepant behavior and affect, the multitude of other cognitive beliefs that a person holds about an attitude object's features is left unspecified in the current network.

Our adoption of the three-component view on attitudes illustrates a first difference with the attributional model of Cooper and Fazio (1984). Although they argued that internal attributions of responsibility for a discrepant behavior are a necessary precondition for dissonance arousal to occur, in the proposed model, we focus instead on attributions to the attitude object. Thus, for instance, when a low incentive provides insufficient justification for engaging in an aversive behavior, then attributions are made to the counterattitudinal object ("The object must be better than I thought").

Another important difference with Cooper and Fazio's (1984) attributional perspective is that the proposed model gives a more prominent and proximal role to affective outcomes. Cooper and Fazio suggested that

dissonance creates arousal, which in turn serves as the instigator of an attributional interpretation. However, we assume that the affective experience itself is subjected to an attributional analysis. This assumption is consistent with extant appraisal and attributional theories of emotion that minimize the mediating role of physiological arousal (Frijda, 1986; Ortony, Clore, & Collins, 1988; Roseman, 1991; Smith & Ellsworth, 1985; Weiner, 1986), with mood-as-information theories that hypothesize that affect serves as a source of information in making judgments and inferences (Schwarz, 1990) and with research focusing on dissonance as an emotional state of discomfort rather than physiological arousal (Elliot & Devine, 1994; Higgins, Rhodewalt, & Zanna, 1979; Losch & Cacioppo, 1990). This different perspective allows explanation of some atypical effects (e.g., reinforcement) that were hereto unexpected by the original (Festinger, 1957) or attributional theory (Cooper & Fazio).

A Feedforward Implementation

To implement our attributional account of cognitive dissonance, we applied an adaptive network approach that stores long-term attitude changes in connection weights without supervision of a central executive. Although there are many of such adaptive connectionist networks (e.g., Read & Montoya, 1999; Smith & DeCoster, 1998), to simplify the exposition of the most important properties that drive dissonance, we used the simple but very powerful standard feedforward network architecture with the Widrow–Hoff or delta learning algorithm (McClelland & Rumelhart, 1988; Van Overwalle, 1998). The delta learning algorithm is responsible for changing the weight of the connections. More interesting, this algorithm is formally identical to the Rescorla–Wagner (1972) formulation of animal conditioning and has been applied in recent research on human causal learning and categorization (e.g., Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988; Shanks, 1991; Van Overwalle, 1998; Van Overwalle & Van Rooy, 1998; for reviews, see Allan, 1993; Shanks, 1993) and, more generally, on several issues in social cognition (Read & Montoya, 1999; Smith & DeCoster, 1998; Van Overwalle, Labiouse, & French, 2001; Van Rooy, Van Overwalle, Vanhoomissen, Labiouse, & French, 2002). Connections in a feedforward network are predictive or causal. For example, they reflect how much a cause predicts or explains an outcome (cf., Mutare, Arcediano, & Miller, 1996).¹

¹ It is instructive to note that the present feedforward implementation of cognitive dissonance can easily be “upgraded” with very similar results to a more complex recurrent architecture used in earlier modeling of social cognition (Read & Montoya, 1999; Smith & DeCoster, 1998; Van Overwalle et al., 2001; Van Rooy et al., 2002).

How does our feedforward network account for changes in attitude connections and dissonance reduction? In essence, because our network employs an adaptive learning algorithm, the connections that link causes (including the attitude object) with outcomes are adjusted online as new information on their co-occurrences are received and processed. This information can be based on one’s own direct experiences and observations as well as on indirect communication or observational modeling (i.e., witnessing other people’s experiences), although indirect information might potentially have less impact. The delta learning algorithm strives to reduce the error between what the network expects based on prior information and the current information. Thus, we concur with Lord (1992) that dissonance has much in common with the Rescorla–Wagner (or delta) learning algorithm that is driven by the “discrepancy between the expected and obtained reward” (p. 341), and thus focuses on reducing the unexpected (see also Festinger, Riecken, & Schachter, 1956). This dynamic feature is illustrated in subsequent sections with a simplified version of a prohibition experiment by Freedman (1965). The whole experiment is discussed in more detail later.

In the example, we focus on the conditions in Freedman’s (1965) experiment in which children were forbidden to play with an attractive toy under either mild or severe threat of punishment. Freedman found that most of the children did not play with the toy and derogated the forbidden toy more under mild than severe threat. Although this occurred only when the children were under surveillance of an adult, for the sake of clarity of exposition, we ignore this factor in the example. Freedman’s results are consistent with the attributional account that severe threat provides more justification for not playing with an attractive toy than the mild threat. With the aid of this example, we first discuss the architecture of the network, that is, how causes and outcomes are represented and connected, and then turn to the learning mechanism that allows the connections to dynamically develop and adjust themselves, leading to lasting changes in attitude.

Representation of Cognitions: Network Architecture

In the present implementation, we focus on the following causes and outcomes. First, we assume that various causes may be responsible for the outcomes, including the attitude object (e.g., toy) and several additional external pressures (e.g., threat) imposed by the experimenter. Second, we assume that two types of outcomes need a causal explanation, notably, the person’s behavior (e.g., playing with the toy) and his or her concurrent emotions (e.g., being happy while playing).

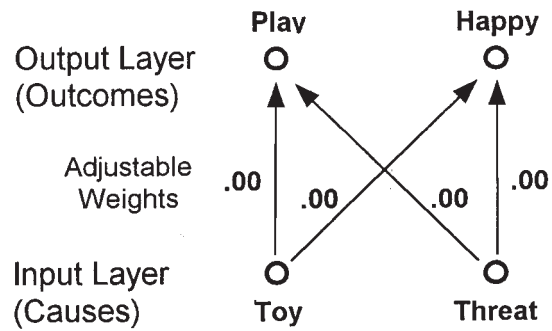
As illustrated in Figure 1A, in the feedforward network, nodes representing causes and outcomes are located in two different layers that are connected via adjustable connections. A first layer consists of input nodes representing the possible causes, and the second layer comprises output nodes representing the behavioral and affective responses anticipated by the network. The connections between input and output represent causal explanations, that is, how well the input determines and influences the output. The weight of the connections reflects the quality or strength of causal influence, or for attitude-object nodes, the intensity of the attitude. Activation in the network is spread from the input nodes to the output nodes through these connections (hence the term *feedforward*), consistent with the intuitive and scientific notion that causes precede and determine outcomes (cf. Mutate et al., 1996).²

**Processing Mechanism:
The Delta Algorithm**

As noted before, an important feature of our adaptive connectionist approach is that the weights are developed dynamically by virtue of the delta learning algorithm. Initially, all connections have zero weights (see Figure 1A) and eventually reach excitatory, inhibitory, or zero weight depending on the person's learning history. This will be demonstrated with a simple learning history from our example as listed in Table 1 (taken from the full-fledged and more realistic simulation history described later). For reasons of simplicity, we chose an example where the learning experiences with the toy precede those with threat, and where behavioral and affective outcomes always match so that they develop identical connection weights. However, the adjustment principles are identical in cases where behavior and affect do not match, because each outcome develops its own connection weights, and only when an attitude is retrieved from memory, their outcome activations are averaged to determine the attitude.

In general, the delta learning algorithm predicts that the more a cause and an outcome co-occur, the stronger their connections will develop until they reach asymptote (typically -1 and $+1$). Consequently, the learning history of our example shown in Table 1 will result in positive connections of toy with playing/affect out-

**A. Feedforward Network
(initial weights)**



**B. Feedforward Network
(after prior history)**

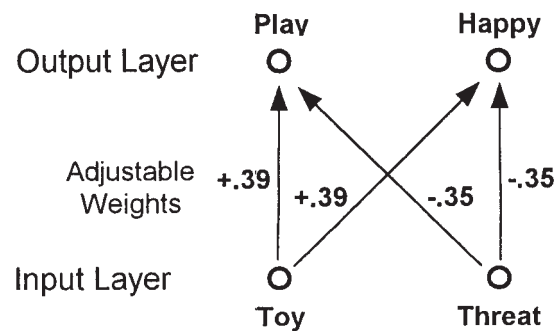


Figure 1. Specifications of the feedforward network model.

comes, and negative connections of threat. Because this learning mechanism provides a novel theoretical account of attitude change and dissonance reduction, we illustrate its workings in some more detail. Let us begin with the first learning trial of Table 1, in which the child plays happily with the attractive toy.

Step 1. When a causal factor is present (e.g., toy), its corresponding input activation is turned on to the default activation level of $+1$, while all other absent causes remain at zero resting activation. This input activation is then spread automatically to the output nodes in proportion to the weight of the connections. Because the connections are still zero, activating the toy node results in zero activation of the output nodes (input activation of $1 \times$ weight of 0).

Step 2. The activations received at the output nodes are linearly summed to determine their activa-

² Although causal attribution is sometimes interpreted as involving backward inferences from outcome to cause, we view attributions as the predictive influence of causes on outcomes. This is the typical interpretation in the associative literature on human causal induction. Moreover, Mutate et al. (1996) demonstrated that backward inferences from outcomes to causes are actually diagnostic inferences. They involve, for instance, the question of which symptom is most indicative of a disease. Thus, although a symptom may have little causal impact on a particular disease, it may constitute a very good diagnostic instrument to differentiate this disease from other diseases.

Table 1. *Simulated Learning Experiences of the Example*

Causal Factor	Frequency	Outcome	
		Behavior	Affect
Pre-experimental History			
Attractive Toy (T)	10	play	☺
T + Severe Threat (Th)	2	no	☺
T + Mild Threat (50% Th)	2	no	☺
Experimental Conditions			
Mild Threat: T + 50% Th	1	no	☺
Severe Threat: T + Th	1	no	☺

Note: Behavior is denoted by *no* when absent; Affect is denoted by ☺ for pleasant and ☹ for neutral.

tion. This output activation can be understood as representing the magnitude of the outcome anticipated by the network. In the example, given only the toy present and zero connection weights, both output nodes receive a total of zero activation.

Step 3. The actually observed outcome is represented by an external teaching signal that has activation of +1 when the outcome is present, zero when absent, and -1 when the opposite outcome is present (this is the typical coding in associative learning theory). Thus, in our example, playing with the toy is represented by a behavioral activation of +1 and not playing by zero. Likewise, experiencing happiness is represented by an emotional activation of +1, moderate affect by zero and unhappiness by -1. Note that these external activations may be inferred from one’s own thoughts and feelings as well as through direct observation or human conversation.

Step 4. The predicted outcome (output activation) is then compared with feedback about the actual occurrence of the outcomes (external activation). In the example, given that both outcomes are present at Trial 1 while the cause-to-outcome connections are zero, there is a large discrepancy or error between the predicted outcomes (output activations of 0) and the actual outcomes (external activations of +1). This error amounts to +1 for each output node. Thus, the network at this point seriously underestimates the magnitude of the behavioral and emotional reactions.

Step 5. Now we turn to the most crucial step in the delta algorithm. To maintain a faithful mental copy of reality, the feedforward network aims to minimize any discrepancy between predicted and actual outcome by adjusting the weights of the connections, in proportion to the magnitude of the error. When the outcome is underestimated, the connections are adjusted upward; when the outcome is overestimated, the connections are adjusted downward (for mathematical equations, see McClelland & Rumelhart, 1988, p. 93–95). In the present case, because the outcome is strongly underes-

timated, the connection between the toy and the outcomes will be adjusted upward.

How fast a person’s mental representation of a dissonant situation is brought into correspondence with reality is determined by a learning rate parameter, which typically varies between zero and +1. A high learning rate indicates that new information has strong priority over old information and leads to radical adjustments in the connection weights, whereas a low learning rate suggests conservative adjustments that preserve much of the knowledge in the weights acquired by old information. In the example, we set the learning rate to 0.30. This implies that only 30% of the error will be used to adjust the connection weights. Hence, the weight of the toy will be incremented by 0.30 (learning rate of 0.30 × error of +1) for each output node, so that after the first learning trial the toy will reach a weight of 0.30 (see Trial 1 in Figure 2).

Forming Attitudes through Learning

The delta learning algorithm is then applied throughout the whole history of the person (depicted in Table 1) by cycling through Steps 1 to 5 at each trial. The weight of the toy on the behavioral and affective output will gradually increase until it will reach the value of +0.97 at Trial 10. As can be seen, the increments become gradually smaller because the error between predicted and actual outcomes decreases.

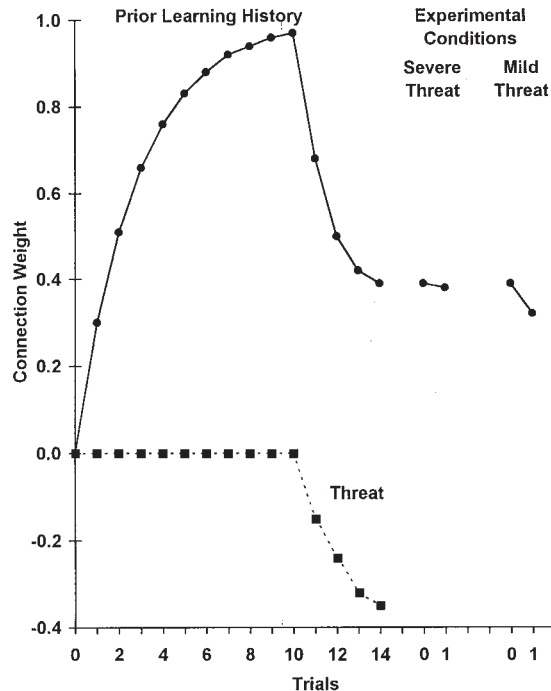


Figure 2. Changes in connection weights after each trial in the prior learning history (Trials 0–14 left) and in each of the experimental conditions (Trials 0–1 middle and right).

However, from Trial 11 onward (see Table 1), threat is combined with the toy, which prevents the child to play and feel happiness. As can be seen in Figure 2, this situation gradually increases the negative weights of threat and also reduces the positive weights of toy. Note that to simulate mild threat, the threat node was activated to only one half its default level (+0.5). The state of the network at the end of the prior learning history is illustrated in Figure 1B. The connection weights with the behavioral and affective nodes are identical in our example because, as noted earlier, (not) playing and (un)happiness always co-occur.

In sum, the repeated exposures to cause–outcome pairings allow the network to incrementally adjust its connections and to anticipate more and more accurately which behavioral and affective outcomes will occur on the basis of the causes present at input (which include the attitude object).

A Principle of Dissonance Reduction

The central hypothesis of our proposal is that the discrepancy in the network between expected and actual outcomes (actions and affect) reflects cognitive dissonance, while the adjustments in the connection weights (determined by the delta algorithm) reflect dissonance reduction through attitude change. Among the several possible sources of cognitive dissonance that Festinger (1957) outlined in his book, one of them is close in spirit to the present proposal. Festinger stated that behavior is guided by accurate information about the environment and the self, and that dissonance can arise when this information disconfirms cognitions or expectations (Festinger et al., 1956). Therefore, any discrepancy between one's predictions (based on relevant input) and one's behavior or emotion would be psychologically disturbing to the person and will be avoided. As Festinger (1957) noted:

Elements of cognition correspond for the most part with what the person actually does or feels or with what actually exists in the environment. In the case of opinions, beliefs, and values, the reality may be what others think or do; in other instances the reality may be what is encountered experientially or what others have told him. But ... persons frequently have cognitive elements which deviate markedly from reality. ... Consequently, the major point to be made is that *the reality which impinges on a person will exert pressures in the direction of bringing the appropriate cognitive elements into correspondence with that reality.* (p. 11, original italics)

Our idea of dissonance as error in the network reflects a novel, perhaps somewhat counterintuitive view on the imbalance in the structure or relationships among the relevant cognitions. Dissonance is captured in our model by the fact that as long as no attitude adjustments

are made, activation of the causal nodes will always create error at output (see also Lord, 1992). The amount of this error (averaged over all output nodes) can be taken as measure of the degree of dissonance. Our model thus specifies how an initially dissonant system can evolve to achieve greater coherence and less dissonance. By changing weights and decreasing the network's error, the network effectively strives for cognitive consistency and reduces cognitive dissonance.

Memory and Retrieval of Attitudes

The temporary information at each learning trial is encoded in the activation of the nodes in the network, whereas long-term causal knowledge is encoded in the connection weights. Part of this long-term knowledge reflects the attitude toward the object. As noted earlier, an attitude is reflected in the long-term connections linking the attitude object to the behavioral and affective output, and attitude intensity is derived from the weight of these connections. Specifically, to measure attitude strength in the network, the attitude object is activated and the outcome activation in the behavioral and affective nodes received through these connections is read off and then averaged to keep between the standard -1 and $+1$ activation range. In the present model, the connection weights from the attitude object and the output activations stemming from them are identical (which facilitates the discussion of the network's workings). However, in other networks (e.g., Read & Montoya, 1999; Smith & DeCoster, 1998), these two may differ slightly.

In the example, given the state of the network after prior learning (see Figure 1B), the strength of the attitude is reflected by the behavioral (+0.39) and emotional (+0.39) connections given the toy, which amounts to a mean outcome activation or positive attitude of +0.39 toward the toy. The moderate intensity of the attitude makes sense intuitively, because an attitude reflects an accumulated history of positive experiences (e.g., happy play) as well as negative experiences (e.g., prohibited play) that are associated with the object. In the example, the number of positive trials exceeds the negative trials, so that the resulting attitude is moderately positive.

Changing Attitudes after Experimental Treatment

So far, we have illustrated that the feedforward network can mimic a person's learning history and resulting connection weights. Can it also explain the change in attitudes observed in various dissonance experiments? Recall that the child did not play with the attractive toy under both severe and mild threat, but that most derogation for the toy was found under mild

threat. How does the proposed model explain when and how much dissonance reduction will occur?

Given severe threat, the activation received at the output nodes from the threat node (-0.35) and from the toy node (+0.39) are almost in balance, so that the summed output activation (+0.04) approaches zero. Hence, the network expects that the child will not play and will experience neutral feelings. This is what actually happened in the experiment, so that the adjustment in the experimental trial is negligible (see Figure 2). In contrast, given mild threat, the threat node is activated at one half its default level so that the activation received from this node is only weakly negative ($0.50 \times -0.35 = -.18$), which together with the output activation of the toy node (+0.39) results in a summed output activation of +0.21. Hence, the network predicts some amount of playing and happy feelings, which did not occur in the experimental situation. In intuitive terms, the network does not provide an adequate account of why children refrained from playing. The overestimation (of playing) in the network results in a downward correction (see Figure 2). Thus, the feedforward network predicts little attitude change after severe threat, and more derogation after mild threat. This is exactly what was observed in Freedman's experiment.

In sum, the proposed feedforward implementation makes similar predictions as Cooper and Fazio's (1984) attributional approach to dissonant behavior and extends their approach to affective responses. It provides a plausible account of how attributions are developed and how lasting attitude changes are stored in connection weights.

In the example, we introduced some simplifications to focus on the critical mechanisms at work in the feedforward model. Among these simplifications were the neglect of some experimental factors, a fixed learning order, and the assumption that causes

and outcomes during prior learning and the experiment are entirely identical. In the simulations to follow, these simplifications will be removed to enhance the realism of the simulations.

Simulations of Dissonance Paradigms

Our selection of simulations was guided by an earlier exposition by Shultz and Lepper (1996) in which they simulated classic, highly reliable paradigms of prohibition (Freedman, 1965), initiation (Gerard & Mathewson, 1966), forced compliance (Linder, Cooper, & Jones, 1967) and free choice (Shultz, Léveillé, & Lepper, 1999; see also Brehm, 1956). The common theme in these paradigms is that participants comply to do something that is discrepant with their own attitudes under the external inducement of threat, payment, or other experimental inducements. The findings typically reveal the counterintuitive result that the less external justification (e.g., threat or payment) for engaging in the discrepant behavior, the more participants tend to change their attitudes in line with their behavior. Because we want to demonstrate that the feedforward model can simulate these phenomena at least equally well as the consonant model, we also present a feedforward simulation of these four paradigms (see Table 2 Simulations 1-3 & 5). The results that we report show that our feedforward model can reproduce these results as well as the consonance model.

To underscore the crucial role of an attributional analysis of affective experiences in generating dissonance that is crucial in our feedforward perspective, we added simulations of paradigms dealing with the misattribution of affective labels given to a placebo pill (Higgins, Rhodewalt, & Zanna, 1979; see Simulation 7), as well as novel experiments of mood manipulation

Table 2. Simulated Learning Experiences in Major Dissonance Paradigms

Causal factor	Frequency	Outcome	
		Behavior	Affect
1. Prohibition (Freedman, 1965)			
Pre-experimental History			
Attractive Toy (T)	20	play ^a	☺
T + Surveillance (S)	10	play	☹
T + Mild Threat (50% Th)	10	no	☹
T + Severe Threat (Th)	10	no	☹
T + S + 50% Th	5	no	☹
T + S + Th	5	no	☹
Experimental Conditions			
Nonsurveillance			
Mild Threat: T + 50% Th	1	no	☹ (3.02)
Severe Threat: T + Th	1	no	☹ (2.67)
Surveillance			
Mild Threat: T + S + 50% Th	1	no	☹ (2.79)
Severe Threat: T + S + Th	1	no	☹ (2.44)

(continued)

Table 2. (Continued)

Causal factor	Frequency	Outcome	
		Behavior	Affect
2. Initiation (Gerard & Mathewson, 1966)			
Pre-experimental History			
Attractive Group (G)	20	participate ^b	☺
G + Initiation Procedure (I)	10	participate	☺
G + Mild Shock (70% S)	10	no	☺
G + Severe Shock (S)	10	no	☹
G + I + 70% S	5	no	☺
G + I + S	5	no	☺
Experimental Conditions			
No Initiation			
Mild Shock: G + 70% S	1	participate	☺ (4.70)
Severe Shock: G + S	1	participate	☹ (1.76)
Initiation			
Mild Shock: G + I + 70% S	1	participate	☺ (5.38)
Severe Shock: G + I + S	1	participate	☺ (2.74)
3. Forced Compliance (Linder et al., 1967)			
Pre-experimental History			
Counterattitudinal Topic (T)	20	no	☺
T + Low Payment (20% \$)	10	write ^c	☺
T + High Payment (\$)	10	write	☺
T + Forced (F)	10	write	☹
T + 20% \$ + F	5	write	☹
T + \$ + F	5	write	☺
Experimental Conditions			
Choice			
Low Payment: T + 20% \$	1	write	☺ (3.78)
High Payment: T + \$	1	write	☺ (3.81)
No Choice			
Low Payment: T + 20% \$ + F	1	write	☹ (2.23)
High Payment : T + \$ + F	1	write	☺ (2.62)
4. Forced Compliance with Mood Induction (Jordens & Van Overwalle, 2001)			
Pre-experimental History (Same as Above)			
Experimental Conditions			
No Choice Without Mood Induction			
Low Payment: T + 20% \$ + F	1	write	☹
High Payment : T + \$ + F	1	write	☺
No Choice With Mood Induction			
Low Payment: T + 20% \$ + F	2	write	½☹ (☹+½☺)
High Payment: T + \$ + F	2	write	☹ (☺ + ☹)
5. Free Choice (Shultz et al., 1999)			
Pre-experimental History			
Attractive Poster (D)	20	chosen	☺
Unattractive Poster (U)	20	no	☺
Experimental Conditions			
Difficult High			
Chosen: D ₁ ^d	1	chosen	☺ (5.81)
Rejected: D ₂	1	no	☺ (3.48)
Easy (High and Low)			
Chosen: D ₁	1	chosen	☺ (6.48)
Rejected: U ₂	1	no	☺ (6.08)
Difficult Low			
Chosen: U ₁	1	chosen	☺ (3.15)
Rejected : U ₂	1	no	☺ (4.56)
6. Free Choice with Mood Induction (Jordens & Van Overwalle, 2002)			
Pre-experimental History (Same as Above)			
Experimental conditions			
Difficult High Without Mood Induction			
Chosen: D ₁ ^d	1	chosen	☺
Rejected: D ₂	1	no	☺

(continued)

Table 2. (Continued)

Causal factor	Frequency	Outcome	
		Behavior	Affect
Difficult High with Mood Induction			
Chosen: D ₁	2	chosen	⊕ (⊕+⊕)
Rejected: D ₂	2	no	⊖ (⊕+⊕)
7. Misattribution (Higgins et al., 1979)			
Pre-experimental History			
Counterattitudinal Topic (T)	20	no	⊖
T + Pleasant Drug (D ⁺)	10	no	⊖
T + Unpleasant Drug (D ⁻)	10	no	⊖
T + Forced (F)	10	write ^c	⊖
T + D ⁺ + F	5	write	⊖
T + D ⁻ + F	5	write	⊖
Experimental Conditions			
Choice			
Pleasant Side Effects: T + D ⁺	1	write	⊕ (4.13/5.80 ^d)
No Side Effects: T	1	write	⊖ (3.69)
Unpleasant Side Effects: T + D ⁻	1	write	⊖ (1.81/1.80 ^d)
No choice			
No Side Effects: T + F	1	write	⊖ (2.26)

Note: Behavior is denoted by *no* when absent; Affect is denoted by ⊕ for pleasant, ⊖ for neutral, and ⊗ for unpleasant. Between parentheses are the means from the survey ($n = 60$); the ratings range from 1 to 7 and higher ratings indicate more pleasantness. Each cause or outcome was represented by 5 nodes with a random activation drawn from a Normal distribution with mean +1 when present and 0 when absent (or +1 when pleasant, 0 when neutral and -1 when unpleasant) and standard deviation of 0.20. During the pre-experimental phase, random noise was added at each trial drawn from a Normal distribution with Mean 0 and Standard Deviation 0.20

^a Playing with an attractive toy; ^b Participating in an initiation; ^c Writing a counter-attitudinal essay; ^d Different posters were chosen or rejected as indicated by different subscripts. ^eThe second mean is reversed from the 7-point "aversiveness" scale in Higgins et al. (1979, table 1).

that illustrate the presumed role of affect in the classical paradigms of forced compliance and free choice (Jordens & Van Overwalle, 2001, 2002, see Simulations 4 & 6). Our model incorporates affect and therefore can explain these results to which the consonance model was not applied.

Before turning to the simulations, it is instructive to discuss briefly two major types of adjustments in the connections, which produce dissonance reduction and permanent attitude change.

Compensatory Adjustments

The first type of dissonance reduction involves compensatory adjustments. These adjustments implement the attributional perspective on cognitive dissonance. When a participant's behavior and affect during an experiment disconfirms pre-existing beliefs and attitudes, this creates dissonance because the connections up to that point provide too little or too much weight to anticipate and justify the behavioral or affective outcomes. If the novel outcome is underestimated (i.e., insufficiently justified), then the connections weights are increased to compensate for the discrepancy. The process is akin to the augmentation principle in social explanation (Kelley, 1971; see also Van Overwalle & Van Rooy, 2001b). For instance, when there are no sufficient grounds for justifying the lie that the task was attractive, then participants enhance their liking for the task (Festinger & Carlsmith, 1959).

In contrast, if the magnitude of the outcome is overestimated (i.e., overjustified), then the connection weights are decreased. This process is akin to Kelley's (1971) discounting principle (see also Van Overwalle & Van Rooy, 2001b). For instance, as we have seen in the example, when little threat provides insufficient explanation why children refrain from playing with a desirable toy, they tend to derogate the toy (Freedman, 1965). Thus, these two compensatory adjustments are responsible for the typical dissonance reduction effects.

Reinforcement Adjustments

The second major type of adjustments involves reinforcement adjustments. These adjustments implement the reinforcement effect that is often opposite to the classical dissonance effect. Reinforcement adjustments are driven by the fact that when a person feels strong negative emotions induced by multiple unpleasant circumstantial constraints, then the amount of dissonance caused by the undesirable behavior is lowered. For instance, when a person is forced to engage in a discrepant behavior in exchange for a minimal financial reward (Linder et al., 1967), this situation will be appraised as extremely uncomfortable, which we presume will generate strong feelings of unpleasantness. This negative affect provides justification for the discrepant behavior (e.g., "I feel so guilty about my behavior that I don't deserve further blame"). This leads to minimal discrepancy, resulting in negligible adjustments of the connec-

tions and little attitude change. However, any increase of reward may generate more positive affect, resulting in the usual amount of discrepancy (e.g., “Why do I feel so little guilt after lying?”) and more attitude change.

These reinforcement adjustments produce a reverse effect, opposite to the typical dissonance effect. For instance, when engaging in a counterattitudinal behavior by force rather than by free will, participants liked their discrepant behavior more given a large rather than a small monetary reward (Linder et al., 1967). We assume that when given a small reward under forced conditions, these two constraints render the whole experimental situation very unpleasant (i.e., negative affect) so that it effectively lowers the total amount of dissonance, resulting in less attitude change than when reward was high. To avoid misunderstanding, it is important to realize that negative affect is not assumed to arise from the attitude object, but from the combination of multiple unpleasant constraints in the experimental situation. The reinforcement reversal was not anticipated by the original (Festinger, 1957) or attributional (Cooper & Fazio, 1984) theory and has been explained previously in terms of a direct reinforcement by external incentives when dissonance is minimal (Linder et al., 1967) that is akin to the present position, or in terms of a mood generalization effect (Shultz & Lepper, 1996) that underscores its affective origin. Later in this article, we provide some empirical support for our idea of increased negative feelings given multiple situational constraints.

Method

A major characteristic of the present simulations is that they were run in two phases. The first phase was a pre-experimental phase during which the connection weights were developed to simulate the assumption that participants begin the experiment with certain beliefs and evaluations. These were acquired earlier during direct experiences or observations, or by indirect experiences through persuasive communication or observation of other’s experiences. The second was an experimental phase during which the experimental manipulations were closely replicated. We first describe how often the attitude object and external factors in the simulations occurred and under which experimental conditions, next the nature and direction of the behavioral and an affective outcomes, how all these cognitions were coded in a distributed manner, and we end with some general features of the simulations.

Although some of the specifications detailed next may seem arbitrary, they are in fact irrelevant with respect to the basic mechanisms at work, and many of them can be relaxed without affecting the simulation results much (see Robustness section at the end of the Simulations). Our aim is to demonstrate that some plausible

assumptions about learning histories can explain human dissonance data, not that the specifications are necessarily correct nor that they are the only possible ones that make the simulations work. A distinct advantage of these learning histories is that they are in principle testable, either by tracking people’s histories or by giving learning trials experimentally. For instance, some authors demonstrated that causal attributions (Van Overwalle & Van Rooy, 2001a, 2001b) and attitudes (Betsch, Plessner, Schwieren, & Gütig, 2001) are developed online by summative processes akin to the delta algorithm. As a kind of cross-validation, we compared the behavioral connections generated by the learning histories with the connection assumptions of Shultz and Lepper (1996) that were based on a review of the relevant literature. All our behavior connections conformed to their conclusions (see Table 3).

Frequencies

Based on logical considerations, it was assumed that during the pre-experimental phase, single factors (e.g., toy) occurred with greater frequency than joint occurrences of factors (e.g., toy and threat). Specifically, the occurrences of the attitude object alone was simulated

Table 3. Connection Weights with Behavioral and Affective Outcome after the Pre-experimental Phase

Study & Target Behavior / Causal Factor	Outcome		
	Behavior	Affect	Mean
1. Prohibition: Playing With Toy			
Attractive Toy	.87 ^a	.88	.88
Surveillance	.01	-.74	-.36
Threat	-1.02 ^a	-.99	-1.00
2. Initiation: Participating to Join Group			
Attractive Group	.96 ^a	.73	.85
Initiation Procedure	.01	-.22	-.10
Shock	-1.08 ^a	-1.17	-1.13
3-4. Forced Compliance: Writing Essay on Topic			
Counterattitudinal Topic	.30 ^a	-.10	.10
Payment	.57 ^a	.35	.46
Force	.48	-.75	-.13
5-6. Free Choice: Choosing Poster			
Attractive Poster	.99 ^a	1.01	1.00
Unattractive Poster	.00 ^a	.00	.00
7. Misattribution: Writing Essay on Topic			
Counterattitudinal Topic	.00	-.00	-.00
Pleasant Drug	-.00	.66	.33
Unpleasant Drug	-.00	-.64	-.32
Force	.98	-.98	.00

Note: Cell entries reflect the weights averaged across all 5 nodes representing each cause or outcome.

^aDirection of the connection is identical to specifications by Shultz & Lepper (1996, averaged across all conditions in tables 3-6); connections without superscript were not specified by Shultz & Lepper.

20 times, the joint occurrence of the attitude object with one external factor 10 times, and the joint occurrence of the attitude object with two external factors 5 times (see Table 2). The external factors chosen in the pre-experimental phase were the same as those in the experimental phase. This was guided by the plausible assumption that external pressures or factors with similar properties had been experienced before the experiment.

During the experiment itself, given that the critical manipulation usually lasted between 1 and 5 min or more, it seemed reasonable to assume that participants would think at least once about the causes for their behavior and emotions. This was implemented by using one trial frequency in each experimental condition for all experiments, unless noted otherwise. It is important to note that the frequencies in the experimental phase were intentionally kept low to avoid a complete destruction of the connection weights learned in the pre-experimental phase, a result that is known as catastrophic interference (French, 1999). It is implausible that a single dissonance experience would totally reverse long-term background knowledge, and this indeed never occurred in the simulations (see also the Concluding Comments section).

Varying Levels of External Constraints

In most experiments, depending on the condition, one critical external constraint was administered in a weak or in a strong amount (e.g., mild or severe threat; mild or severe shock, low or high payment). The presence of a weaker level of an external constraint was simulated by activating its corresponding input node for only 20%, 50%, or 70% of the default activation level. In the Linder et al. (1967) experiment, 20% was chosen to reflect the relative amount of low payment (i.e., \$0.5 versus \$2.5 respectively). In the other simulations, it was difficult to gauge the exact degree of the weak treatment level because there is no comparative numerical data to compare with the strong treatment level. Therefore, we chose the percentage that provided the best fit with the observed data (see Table 2). However, as we will see, this percentage appeared to be quite critical in simulating some conditions of two paradigms (prohibition and initiation).

There are two possible ways to interpret this critical dependence on the activation values for weaker treatment. Either it can indicate that our simulations work merely by fitting the data. Alternatively, they may indicate some lack of robustness in the empirical data itself, in the sense that the obtained effects may easily disappear with other levels of external constraints. This latter possibility is in principle testable by manipulating different treatment levels and seeing how they affect attitude change. We will return to this issue when we come to the specific paradigms.

Behavioral Outcomes

As noted earlier, a central assumption of the feedforward model is that participants seek an explanation for their discrepant behavior and for their feelings. The coding of the behavioral outcomes during the experimental phase is straightforward, as the original reports involved only participants that complied with the experimenters' request. For the pre-experimental outcomes, it was assumed that a person would typically engage in proattitudinal behavior, except when the same external pressures as in the experiment were present, in which case he or she would engage in counterattitudinal behavior. The behavioral output node in each paradigm was chosen to reflect approach behavior toward the attitude object (i.e., playing with the toy, joining the group, writing the essay, choosing the poster were coded +1; see Table 2), regardless of participants' initial attitude.

Affective Outcomes

The coding of affective outcomes is more problematic because there exists little data on participants' feelings, except in some forced compliance studies. These studies explored emotional reactions by giving participants the opportunity to misattribute dissonance arousal to a pill (Cooper, Fazio, & Rhodewalt, 1978; Higgins et al., 1979; Zanna, Higgins, & Taves, 1976), by experimentally inducing positive or negative emotions (Kidd & Berkowitz, 1976; Rhodewalt & Comer, 1979), or by asking participants to rate their emotions (Elliot & Devine, 1994; Rhodewalt & Comer, 1979; Shaffer, 1975). Although it is typically assumed that dissonance leads to negative affect (Cooper & Fazio, 1984), the data from emotional ratings seem to suggest that freely engaging in a discrepant behavior is not overwhelmingly negative. Rather, the situation is most often experienced as mildly uncomfortable.

Hence, the coding of the affective outcomes was guided by the hypothesis that participants would experience their proattitudinal behaviors as pleasant, and their unwillingness to engage in counterattitudinal behaviors as affectively neutral. Likewise, when participants were constrained to engage in counterattitudinal behavior, this would generally lead to mild negative affect, except in a combination of two unpleasant external constraints (e.g., high threat and surveillance, severe shock and noninitiation, low payment and lack of choice) that was hypothesized to lead to strong negative affect.

To provide some empirical validation for these affective outcome assumptions, a small survey was conducted in which the paradigms of interest were described and participants indicated the emotions they would most likely feel. Although this is only a rough equivalent of participants' true feelings in the original experiments, it provides some empirical con-

straints to the simulations and avoids making too many arbitrary assumptions.

Sixty sophomore students were given a brief description of each experiment, including the specific procedures of each condition. The students were asked to empathize with the participants in each condition, and to indicate how they would feel in that situation on two 7-point scales measuring pleasantness (*very unpleasant* to *very pleasant*) and discomfort (*not at all uncomfortable* to *very uncomfortable*; Elliot & Devine, 1994). Note that we requested to report on feelings toward the whole situation, not to the attitude object. Given the high correlation between the means of these two measures ($r = -.98$), the discomfort scale was reversed, and both scales were averaged into a single pleasantness score ranging from 1 to 7. Consistent with earlier forced compliance research, the means indicated that the conditions that caused most dissonance reduction were generally experienced as only moderately negative or positive (i.e., the means never exceeded 1 point off the scale midpoint; except in the initiation paradigm, see Table 2). More important, as hypothesized, the combination of two unpleasant external constraints was generally experienced as most negative.³

Two cut-off points were most appropriate in capturing our emotional outcome assumptions. All means below 2.50 were coded as unpleasant (activation level -1), all means above 5.50 as pleasant (level +1), and all means in-between (i.e., reflecting mild levels of unpleasantness or pleasantness) were coded as neutral (level 0). These cut-off points assume a sort of threshold activation function in which affective reactions have an effect on attitude adjustments only when they are extremely positive or negative. As we see, this coding scheme is critical to the success of the simulations, although it remains possible that other cut-off levels may be equally successful.

The affective outcomes in the pre-experimental phases were chosen in accordance with the data from the survey as well. That is, pre-experimental situations with the same conjunction of factors and behavioral outcomes as the experimental conditions were given the same affective outcome. If such combination did not exist, an affective outcome was chosen for the pre-experimental phase that was logically most compatible with the other conditions (see Table 2).

Distributed Coding

A major disadvantage of the coding scheme used in the example was that each cognition (cause or out-

come) was represented by a single node. This implies the assumption that the cognitions before and during the experiment were identical. To relax this limitation and add more realism to the simulations, rather than using a single node that was either present or not (i.e., localist coding with on-off activation), each concept was represented by a pattern of activation across a range of nodes, each of which represented some underlying microfeature that varied in the extent to which it was present or not (i.e., distributed coding with a varying pattern of activation). More important, during the pre-experimental phase, at each trial random noise was added to the activation pattern of each node. Hence, causes and outcomes prior to the experiment differed somewhat from the causes and outcomes during the experiment itself. This allows for making inferences about related cognitions stored earlier in memory by virtue of their similar activation pattern. Nothing else was changed in the model as described earlier.

Technically, each relevant cause or outcome was represented by five nodes. When the cause or outcome was present, a random activation pattern normally distributed with mean +1 (or -1 for negative affect) and standard deviation 0.20 was applied on these five nodes. When the cause or outcome was absent (or neutral for affect), the activation was set to zero. Noise was added during the pre-experimental phase by increasing or decreasing this distributed activation pattern at each trial with a random activation drawn from a normal distribution with zero mean and standard deviation 0.20 (for a similar procedure, see Smith & DeCoster, 1998). Although, technically, a distributed coding seems more complex, this does not change the learning mechanism in any way. Thus, the interpretation of the simulations is the same as if a localist encoding was used.

General Characteristics

Given that the exact course of the learning histories of the participants during the pre-experimental phase was unknown, for each experiment, we ran 50 simulations (or "participants") in each condition with different random orders in the pre-experimental phase. That is, the pre-experimental phase and one condition in the experimental phase were run until completion by going once through all trials. The number of trials (see Table 2) was assumed to be loosely similar to the human participants, and the expected attitude change was typically produced after this run (as was also the case in our illustrative simulation of Figure 2.) This process was repeated 50 times for each condition to mirror 50 "participants" in each condition of an actual experiment. Because of the random ordering of trials as well as the different noise in the pre-experimental phase, the results for each run (or "participant") were slightly different. This reflects the variable and imperfect conditions of human perception in the actual experiments.

³ A follow-up survey using minor rewordings of the original questionnaire conducted 2 years later with another 119 sophomore students yielded almost identical results, as the mean ratings between the conditions in the original and follow-up survey correlated very highly ($r = .96$).

For the same reason, because the (distributed) activation pattern of each cause or outcome was chosen randomly, we ran 20 “replications” of each experiment with a different distributed activation pattern for each cognition. This guarantees that it was not the particular activation pattern that produced our results. The learning rate parameter was set to a fixed default value of 0.10 in all simulations reported. Note that the high amount of runs ($50 \times 20 = 1000$) was performed to safeguard against the potential arbitrariness of trial order and activation levels, as 50 runs (“participants”) produced very similar results.

Results and Discussion

The results over the 50 simulated random runs and the 20 replications in each experiment were averaged and subjected to an analysis of variance (ANOVA) using the same between-subject factors as in the original experiments. Because of the high number of data points ($20 \times 50 = 1000$), all significance levels for the simulations were set a very stringent level of $\alpha = 0.0001$. The simulation results revealed that the main or interaction F tests of interest were significant in all experiments, $F_s > 625.51$, $p_s < .0001$. Comparisons of interest were then tested with unpaired t tests. Because the α level of all tests and the degrees of freedom of the t tests (all 1998 df) were identical throughout all simulations, they will not be reported.

Simulation 1: Prohibition

Experiment. The first insufficient justification paradigm explores the effects of prohibiting a desired action (Freedman, 1965). School children were forbidden to play with an attractive toy (a robot) under either mild or severe threat of punishment, and the experimenter either stayed in the room while the child played (surveillance condition), or went away (this surveillance variable was not included in the introductory example). Actual play with the previously forbidden toy about 40 days later in the absence of the experimenter or any threat, revealed greater derogation of the forbidden toy in the mild than in the severe threat condition when there had been no surveillance. When there had been surveillance, the effect of severity of threat was negligible. These results are depicted in Figure 3 (top panel).

The attributional explanation for these results was that mild threat alone provided insufficient justification for the counterattitudinal behavior of not playing with the attractive toy and thus created high dissonance that was reduced by lowering the attraction for the toy. In contrast, either the high threat or the exper-

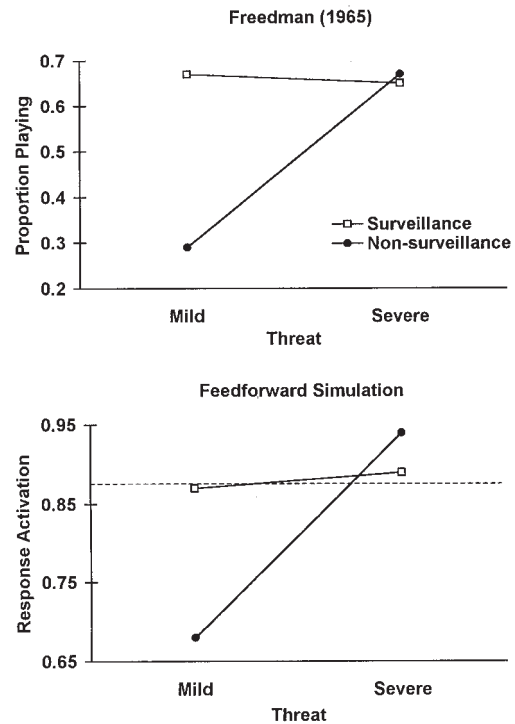


Figure 3. Human data (top) and feedforward simulation (bottom). The broken line shows the attitude prior to the experiment. The human data are from Table 1 in “Long-term behavioral effects of cognitive dissonance,” by J. L. Freedman, 1965, *Journal of Experimental Social Psychology*, 1, 145–155. Copyright 1965 by Academic Press. Adapted with permission.

imenter’s surveillance provided sufficient justification for not playing with the toy and thus created little dissonance and little attitude change.

Simulation. In the simulation, three factors were of interest—toy, threat, and surveillance (see Table 2). For the pre-experimental phase, we assumed that the most natural and most often occurring situation for the child would be one in which it played with an attractive toy, a pleasant experience. If surveillance was present also, we assumed that the child would still play but feel less happy as unfamiliar adults often makes children wary and uncomfortable. When threat was present alone or in combination with surveillance, we assumed that children would not play with the forbidden toy. Our survey data further suggested that children would be only mildly unhappy (neutral) given external constraints, except when surveillance was combined with severe threat in which case they would be very unhappy.

After running through the pre-experimental specifications, the model ends up with behavioral and affective connections that are positive for the attractive toy and negative for threat. The behavioral effect of surveillance was negligible, whereas its emotional effect was negative (see Table 3). These connections are intuitive plausible.

Simulation Results. The results in Figure 3 (bottom panel) depict children's attitude toward the toy (i.e., average connection between toy and outcomes) after the experimental phase. It can be seen that the feedforward network replicated the prohibition effect. That is, liking for the toy was high except when threat and surveillance were combined. The interaction between threat and surveillance was significant, $F(1,3996) = 1338.52$. As predicted, direct comparisons with t tests revealed that under surveillance, weak and severe threat did not differ from one another, $t = 3.74$, *ns*. In contrast, without surveillance, there was more derogation for the toy when threat was weak as opposed to severe, $t = 66.03$.

The feedforward mechanism responsible for this strong derogation is a compensatory adjustment. The mild threat activation without surveillance was insufficient to justify and anticipate prohibition of play. (As can be seen in Table 3, the sum of the mean weights of mild threat [50% of $-1.00 = -.50$] and toy [$+0.88$] is insufficient [> 0] to predict that the child will not play [= 0]). Thus, the network overestimated the possibility that the child would play happily with the toy, resulting in compensatory downward adjustments of the connections and a lower attraction for the toy.

It should be noted that using a different activation value for mild threat (now 50%) changed the nonsignificant effect under surveillance. When threat was very weak (20%), it led to more derogation of the toy although to a much lesser extent than without surveillance. When threat was stronger (70%), it led to less derogation. These results are consistent with the attributional perspective, which predicts that less (vs. more) justification for not playing should result in less (vs. more) liking for the toy. Future research can establish whether these predictions concerning the influence of different levels of threat under surveillance are correct.

Simulation 2: Initiation

Experiment. In the initiation experiment by Gerard and Mathewson (1966), participants were administered mild or severe electric shocks, either as part of an initiation to join an attractive discussion group, or as part of a psychological experiment. After this, all participants heard a boring discussion ostensibly by the discussion group, or as part of the experiment respectively. Ratings of the participants revealed that participants who received a severe shock liked the group better than participants who received a mild shock, but only in the initiation group (see top panel in Figure 4). The attributional perspective predicted this effect because one's willingness to join the attractive

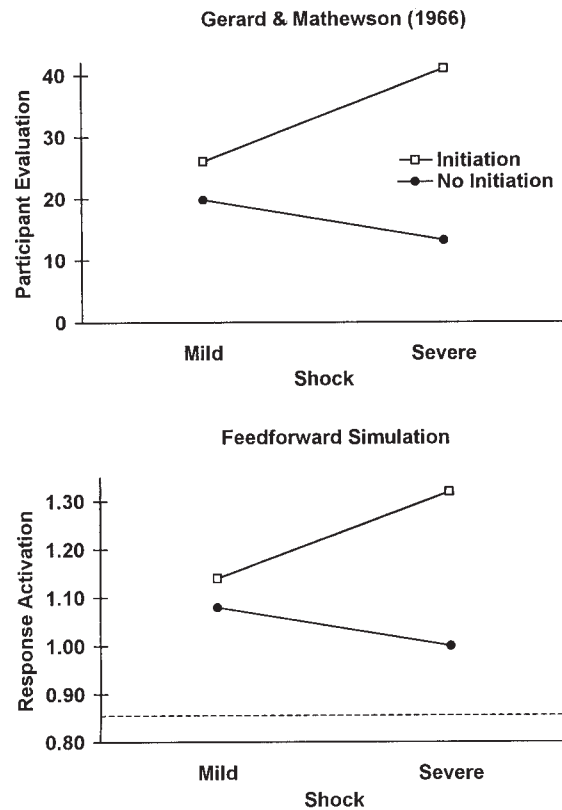


Figure 4. Human data (top) and feedforward simulation (bottom). The broken line shows the attitude prior to the experiment. The human data are from Table 1 in "The effects of severity of initiation on liking for a group: A replication," by H. B. Gerard & G. C. Mathewson, 1966, *Journal of Experimental Social Psychology*, 2, 278-287. Copyright 1966 by Academic Press. Adapted with permission.

"discussion" group was less justified after a severe initiation than after a mild one.

In addition, a reverse trend was found in the noninitiation condition, that is, the "experiment" group was liked more after receiving a mild shock. Although dissonance theory or the original attributional perspective cannot explain this finding, Shultz and Lepper (1996) were able to reproduce it in their consonance model by reversing, rather arbitrarily, the (experimenter-imposed) connection between shock and evaluation of the group from positive in the initiation condition to negative in the noninitiation condition.

Simulation. In the simulation, we manipulated three factors—group, shock, and initiation procedure (see Table 2). For the pre-experimental phase, when an open admission policy is used without additional constraints, we assumed that a person would be willing to join an attractive group and experience this as pleasant. In addition, we assumed that a typical initiation procedure would generally lead to the same

choice, although people might be less pleased by such a procedure. (A typical initiation or admission procedure ranges from easy tasks like filling in a subscription form to more demanding requirements like paying a high fee or doing entrance tests). In Gerard and Mathewson's (1966) experiment, the admission procedure (without shock) consisted simply of telling the participants that they would join an attractive group. We further assumed that undergoing an unpleasant shock would withhold people to join the group. Our survey data suggested that a situation in which a severe electric shock was combined with no perspective of joining an attractive group would be experienced as very unpleasant, whereas the same shock as part of an initiation procedure to join an attractive group would be experienced as only mildly unpleasant (neutral). This is consistent with our initial assumption that a single unpleasant constraint (undergoing an electric shock) would be typically experienced as mildly negative at most, whereas the combination of two unpleasant constraints (undergoing a shock and doing an experiment) would be experienced as much more negative. Without this specification, the simulation is not able to reproduce the major results of Gerard and Mathewson.

After running the feedforward model through these pre-experimental specifications, the behavioral and affective connections were positive for the group and negative for the shock (see Table 3). The initiation procedure had negligible behavioral consequences, but a negative emotional impact.

Simulation Results. The simulation results in Figure 4 (bottom panel) indicate that the feedforward model replicated the human dissonance data. As expected, the interaction between initiation and shock was significant, $F(1,3996) = 736.06$. Direct comparison with t tests revealed that in the initiating condition, the group was liked more after a severe shock as opposed to a weak shock, $t = 23.41$.

The feedforward mechanism producing this result is a compensatory adjustment. The person's willingness to undergo aversive treatment comes as a surprise to the model because the negative connection weights prior to the experiment did not predict this response. (As Table 3 reveals, the sum of the mean weights of the group [+0.85], initiation [-0.10] and shock [-1.13] is insufficient [$< +1$] to predict that the person would join the group [= 1]). This underestimation of the actual behavior created compensatory upward adjustments, leading to stronger positive attitude connections with the group, especially after severe initiation treatment.

The reverse, reinforcement trend in the noninitiation condition was also replicated. The simulation replicated the finding that the group in this

condition is liked less after a severe shock, $t = 14.07$. The feedforward mechanism underlying this effect is a reinforcement adjustment. The extremely unpleasant experience of receiving a severe shock as part of an experiment annihilated the dissonance created by engaging by the counterattitudinal behavior. (That is, the negative external activation reflecting unpleasant affect [-1] completely absorbed the positive external activation reflecting participants' willingness to join the group [+1]). This lack of overall discrepancy produced no adjustments. In contrast, the mild shock was felt as only moderately aversive (neutral), and this resulted in some dissonance. (As shown in Table 3, the sum of the mean weights of the group [+0.85] and the mild shock [70% of $-1.13 = -.79$] is insufficient [$< +1$] to predict that the person will join the group [= 1]). This underestimation was adjusted in the typical compensatory manner by upward adjustments and increased liking of the group.

It should be noted that these results were obtained with a relatively high (70%) level of activation for the mild shock. A much reduced level (20%), however, reversed the obtained reinforcement effect in the no-initiation condition. That is, a very mild shock led to less liking for the group than a somewhat stronger shock, although to a much lesser degree than in the initiation condition. This effect is consistent with an attributional explanation of insufficient justification for participating in a nonattractive group. Again, it is possible to test this prediction concerning varying levels of shock in future research.

Simulation 3: Forced Compliance

Experiment. The third insufficient justification paradigm is the forced compliance experiment by Linder et al. (1967; see also Calder, Ross, & Insko, 1973; Collins & Hoyt, 1972; Sherman, 1970). Participants were asked to write a forceful counterattitudinal essay (supporting a ban on communist speakers on campus) under choice or no-choice conditions and were paid either \$0.5 or \$2.5 for it. After writing the essay in the choice condition, banning communist speakers was favored more after low rather than high payment, whereas the reverse pattern was observed in the no-choice condition (see Figure 5, top panel).

The original attributional interpretation for the results in the choice condition was that low payment provided insufficient justification for writing a counterattitudinal essay, creating high dissonance that was reduced by changing one's attitude in favor of the position taken in the essay. However, the reverse effect in the no-choice condition was not predicted by dissonance or attribution theory. The consonance model (Shultz & Lepper, 1996) could simulate this reverse pattern only by turning the connection be-

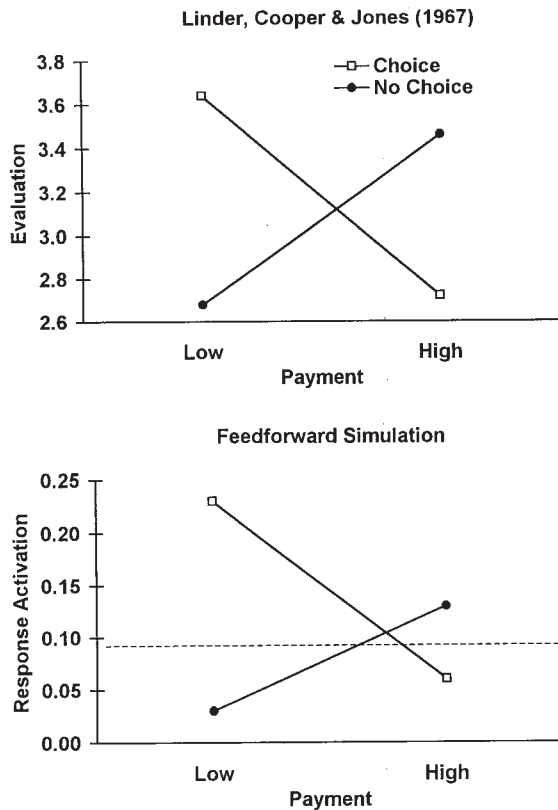


Figure 5. Human data (top) and feedforward simulation (bottom). The broken line shows the attitude prior to the experiment. The human data are from Table 3 in “Decision freedom as a determinant of the role of incentive magnitude in attitude change,” by D. E. Linder, J. Cooper, & E. E. Jones, 1967, *Journal of Personality and Social Psychology*, 6, 245–254. Copyright 1967 by the American Psychological Association. Adapted with permission.

tween payment and one’s counterattitudinal position from negative in the choice condition to positive in the no-choice condition.

Simulation. In the simulation, three factors were simulated—essay topic, payment, and enforcement. For the pre-experimental phase, we assumed that a counterattitudinal topic would never result in writing a favorable essay about it, except when pressed to do so by payment or force. The survey data (see Table 2) further suggested that writing a counterattitudinal essay is experienced as mildly unpleasant only, as long as the person was well paid or not forced to do this. However, the combination of low payment and enforcement is experienced as very unpleasant.

After running these pre-experimental specifications through the feedforward model, this led to positive connections from the essay topic, payment, and enforcement to the behavioral outcomes, whereas the topic and enforcement had a negative emotional impact and payment a positive emotional impact (see Table 3).

Simulation Results. Figure 5 (bottom panel) depicts the simulation results, which replicated the dissonance effect in the choice condition and the (reverse) reinforcement effect in the no-choice condition. The interaction between choice and payment was significant, $F(1,3996) = 3321.35$. As expected, t tests revealed that, in the choice condition, attitude change in favor of the counterattitudinal topic was greater when payment was low rather than high, $t = 59.56$.

The feedforward mechanism producing the dissonance effect in the choice condition is a compensatory adjustment. Small payment was insufficient to anticipate the person’s choice to engage in the counterattitudinal behavior. (As can be seen in Table 3, the weight of low payment [20% of $+0.46 = +0.09$] was insufficient [$< +1$] to predict that the person would write the essay [$= 1$]). This underestimation generated compensatory upward adjustments, leading to stronger attitude change.

The reverse, reinforcement trend in the no-choice condition was also reproduced. The counterattitudinal position was endorsed less after low than high payment, $t = 25.72$. The feedforward mechanism underlying this trend is a reinforcement adjustment, as the negative emotion given low payment almost entirely absorbed the dissonance created by the discrepant behavior. (That is, the positive external activation generated by the counterattitudinal behavior [$+1$] was canceled by the negative external activation of the unpleasant affect [-1]). This *small* dissonance led to weak downward adjustments. In contrast, high payment led to more moderate feelings (neutral), and this resulted in greater dissonance, which was reduced by the typical upward compensatory adjustments and more support for the counterattitudinal essay. This explanation of the reinforcement effect in terms of cancellation of dissonance by extreme negative affect was also given in the preceding paradigm. It is interesting to note that higher activation levels (50% & 70%) for the small payment condition only attenuated the dissonance effect with free choice but did not alter them substantially.

Simulation 4: Forced Compliance with Mood Induction

Experiment. Perhaps one of the most innovative aspects of the present connectionist model is the important role of affect in a dissonant situation. To provide independent empirical support for the affective coding used in our simulations, Jordens and Van Overwalle (2001, 2002) first replicated the original Linder et al. (1967) forced compliance experiment and then induced positive and negative mood.

In the replication (no-mood) part of the experiment, participants were randomly assigned to one of the Linder et al. (1967) conditions with varying levels of choice and payment. Dissonance was produced by writing a counterattitudinal essay on abolishing the university credit system. Choice was manipulated by telling the participants that the decision to write the counterattitudinal essay was entirely their own (high choice) or that they had been randomly assigned to write the essay (no choice). Payment was manipulated by giving the participants either the equivalent of \$10 for writing the essay (high payment) or \$0.25 (low payment). They were all paid before starting to write the essay. After finishing the essay, participants' attitude toward abolishing the credit system was measured.

To verify the role of affect, Jordens and Van Overwalle (2001) then attempted to eliminate the reinforcement effect in the no-choice replication conditions by inducing the opposite affect in two additional no-choice mood conditions. Recall that our model assumes that negative affect is experienced in the low payment condition whereas relatively neutral affect is experienced in the high payment condition, and that the difference in affect drives the reinforcement effect. Hence, opposite affect was created by inducing positive affect in the low payment condition and negative affect in the high payment condition. The affect was induced by providing false performance feedback about an ostensibly unrelated intelligence test completed before the essay-writing task. The mood feedback itself was given right before the final attitude measure to avoid that affect would dissipate while writing the essay and to have a maximal impact on the dissonance experienced before and during assessment of one's attitude.

The results revealed that the interaction between choice and payment in the attitude toward the essay obtained in the Linder et al. (1967) study was successfully replicated. Figure 6 (top panel) shows the no-choice conditions. As can be seen for the no-choice replication, the predicted reinforcement effect was again obtained in that the attitude after a large payment changed more than after a low payment. In contrast, as expected, in the no-choice mood condition, this reinforcement effect was eliminated. The attitude changed slightly more after a low payment with positive mood than after a high payment with negative mood. This trend in the direction of a reversed reinforcement effect was marginally significant ($p = .07$). Although the decrease of attitude might be explained in traditional terms by a misattribution of dissonance to the negative test feedback, it is unclear how misattribution can explain an increase in attitude after positive feedback. Hence, these results provide the first empirical support for the hypothesis advanced in our model that strong

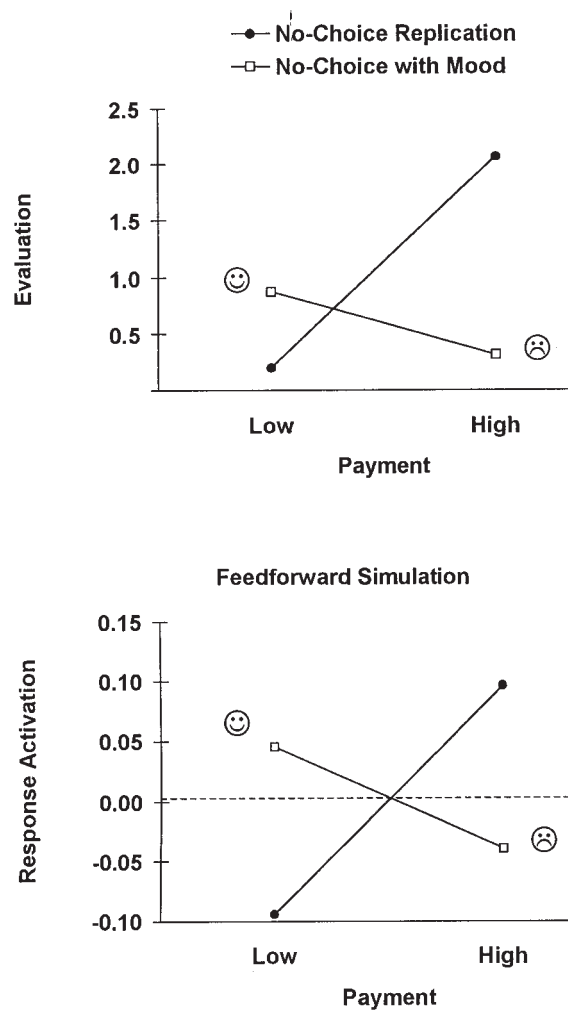


Figure 6. Human data (top) of the No-Choice condition of Jordens & Van Overwalle (2001) with or without mood induction and feedforward simulation (bottom). (☺ and ☹ refer to negative and positive induced mood respectively). The broken line shows the attitude prior to the experiment. The human data are from Figure 2 in "Een empirische toetsing van een feedforward connectionistisch model van cognitieve dissonantie: De rol van affect in het geïnduceerd-inwillingsparadigma," by K. Jordens & F. Van Overwalle, 2001, In D. A. Stapel, C. Martijn, E. van Dijk, & A. Dijksterhuis (Eds.), *Fundamentele sociale psychologie* (Vol. 15, pp. 91–102), Delft, The Netherlands: Eburon. Copyright 2001 by the authors. Adapted with permission.

negative affect plays a crucial role in the reinforcement effect given forced compliance.

Simulation. Can the feedforward model replicate the mood manipulation by Jordens and Van Overwalle (2001)? To verify this, we simulated the no-choice conditions with or without mood induction. All the model specifications were identical to the previous simulation except that in the mood induction conditions, affective outcome was additionally determined by the mood manipulation. Because Jordens and

Van Overwalle (2001) reported that positive mood induction was less effective than negative mood induction, affective outcome activation was increased by +0.5 given a positive induction, and decreased by -1 given a negative induction, resulting in moderate (-0.5) and extreme (-1) negative affect (see Table 2). In addition, two rather than one experimental trials were provided in the mood induction conditions to reflect the fact that unexpected feedback triggers additional cognitive activity.

Simulation Results. As can be seen in Figure 6 (bottom panel) the reinforcement effect in the no-choice condition was reversed after inducing the opposite mood. The interaction between payment and mood was significant, $F(1,3996) = 3260.64$. As expected, t tests revealed that, in the original no-choice condition, attitude change in favor of the counterattitudinal topic was greater when payment was high rather than low, $t = 52.26$, whereas this effect was reversed after inducing mood, $t = 26.79$.

Simulation 5: Free Choice

Experiment. The next paradigm involves a study by Shultz et al. (1999), in which participants were provided with various posters with different levels of attractiveness. After an initial evaluation of the posters, the participants had to make a choice between two very attractive posters (difficult-high condition), a very attractive and a less attractive poster (easy condition), or two less attractive posters (difficult-low condition). Then the posters were evaluated again. It was predicted that making a choice between two alternatives creates cognitive dissonance because the chosen alternative is never perfect and the rejected alternative often has desirable aspects that have to be foregone.

Figure 7 (top panel) depicts the change between final and initial evaluation. Most of the negative change for the rejected alternative was found in the difficult-high condition. According to attribution theory, the insufficient justification for the rejected, but attractive alternative created cognitive dissonance that was reduced by decreasing its attractiveness. A similar result was reported in an earlier study by Brehm (1956). In contrast, most of the positive change for the chosen alternative was found in the difficult-low condition. The attributional interpretation is that there was insufficient justification for the chosen alternative when both alternatives were unattractive, creating cognitive dissonance that was reduced by increasing the attractiveness of the chosen poster.

Simulation. In the simulation, there were two factors of interest—an attractive and an unattractive

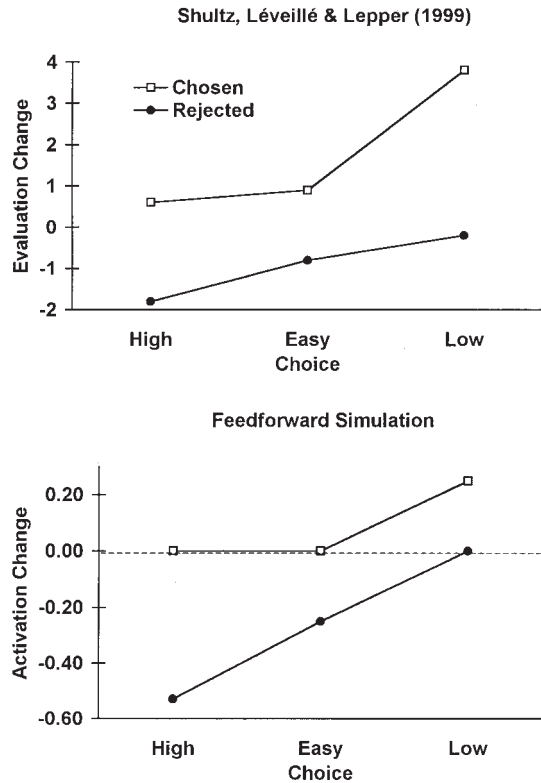


Figure 7. Human data (top) and feedforward simulation (bottom). The broken line shows the attitude level prior to the experiment. The human data are from Figure 4 in “Free choice and cognitive dissonance revisited: choosing ‘lesser evils’ versus ‘greater goods’,” by T. R. Schultz, E. Léveillé, & M. R. Lepper, 1999, *Personality and Social Psychology Bulletin*, 25, 40–48. Copyright 1999 by Society for Personality and Social Psychology. Adapted by permission of Sage Publications, Inc.

poster. In the pre-experimental phase, we made the straightforward assumption that the attractive poster was always chosen with pleasure, whereas the unattractive poster was always rejected with neutral feelings (see Table 2). After running through these pre-experimental specifications, the connections were all positive for the attractive poster and all zero for the unattractive poster (see Table 3).

Simulation Results. The simulation results depicted in Figure 7 (bottom panel) show that the feedforward network replicates the major aspects of the human data. The predicted interaction between difficulty and alternative (chosen vs. rejected) was significant, $F(2,5994) = 29922.96$, indicating that the chosen poster was rated more favorably than the rejected poster. For each of the chosen and rejected posters, the different levels of difficulty were also significant, $F_s(2,2997) = 9889.34–13990.43$. As predicted, t tests showed that the attitude for the chosen poster in the difficult-low condition was more positive than in the difficult-high and easy choice conditions, $t_s =$

140.50–144.02, whereas the two latter conditions did not differ, $t = 0.65$, *ns*. Likewise, as predicted, the attitude for the rejected poster in the difficult-high condition was more negative than in the easy and difficult-low conditions, $t_s = 60.07$ – 122.77 , although these latter two conditions also differed from another, $t = 154.30$, unlike the human data.

The feedforward mechanism underlying these changes is compensatory adjustments. The positive change in the difficult-low condition for the chosen unattractive poster is because its choice came as a surprise to the network because its pre-experimental mean weight is zero (see Table 3). This underestimation led to a compensatory upward adjustment. Similarly, the negative change in the difficult-high condition for the rejected attractive poster is because rejection came as a surprise to the network because the pre-experimental mean weight was very positive (+1.00). This overestimation by the network produced a compensatory downward adjustment.

Simulation 6: Free Choice with Mood Induction

Experiment. To provide independent empirical support for the role of affect in our model, Jordens and Van Overwalle (2002) replicated the high-difficulty condition of Shultz et al.'s (1999) poster experiment and, more important, attempted to attenuate attitude change by inducing negative affect. In the replication of the high-difficulty condition of Shultz et al. (1999), participants rated the likeability of eight different posters and were then offered a choice between two highly evaluated posters. After their choice, participants' rated again the likeability of the posters. More important, in an additional mood condition, negative affect was induced by providing false feedback on an intelligence test in the same manner as described previously for the forced compliance experiment (Simulation 4). The affect feedback was given before the likeability of the posters was rated again.

As can be seen in Figure 8 (top panel), in the replicated high-difficulty condition (without mood), dissonance was reduced by increased liking for the chosen poster (whereas in Schultz et al.'s study dissonance was reduced mainly by derogating the rejected alternative, as also the previous simulation would predict). More important, after inducing negative mood, the attitude change in favor of the chosen poster disappeared, while the rejected poster was now significantly derogated ($p < .05$). Thus, as predicted, negative mood led to a significant decrease in attitude change for both the chosen and rejected poster. An interpretation of these findings in misattribution terms would predict less attitude change for both the

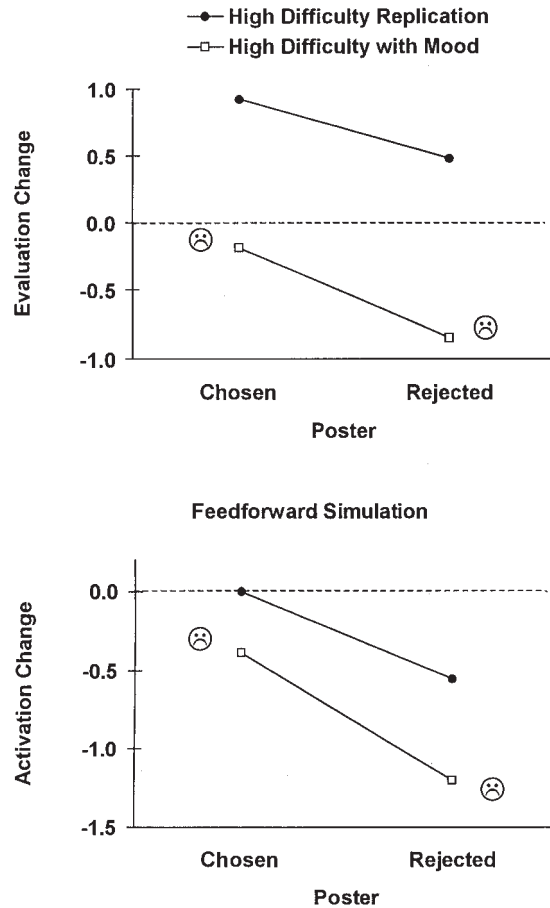


Figure 8. Human data (top) of the high-difficulty condition of Jordens & Van Overwalle (2002) with or without mood induction and feedforward simulation (bottom) (⊗ refers to negative induced mood). The broken line shows the attitude prior to the experiment.

chosen and rejected poster after attributing the dissonance to the negative feedback. In contrast, Jordens and Van Overwalle (2002) found more derogation only for the rejected poster. Hence, the obtained findings are consistent only with the affective hypothesis advanced here.

Simulation. To verify whether our connectionist model can replicate the mood manipulation by Jordens and Van Overwalle (2002), we reran the high-difficulty condition with or without mood induction. All the model specifications were identical to the previous simulation except that given a negative mood induction, the affective outcome activation was decreased by -1 , resulting in a neutral (0) and negative (-1) affective outcome for the chosen and rejected poster respectively (see Table 2). In addition, two rather than one experimental trials were provided as in the previous mood induction simulation.

Simulation Results. The simulation results depicted in Figure 8 (bottom panel) show that the network replicates the major aspects of the human data. The predicted main effect of mood was significant, $F(1, 3996) = 52070.78$, indicating that both the chosen and rejected posters were rated less favorably as predicted by our model. Additional t tests showed that the attitude for the chosen poster was decreased, $t = 264.54$, as well as for the rejected poster, $t = 151.08$.

Simulation 7: Misattribution of Forced Compliance

Experiment. To underscore the importance of affect, we finish this series of simulations with a misattribution experiment by Higgins et al. (1979), which can be best understood if it is accepted that participants believed to experience the emotions induced by the alleged side effects of a placebo pill (see later). Participants were given a counterattitudinal essay to write under conditions of high or low choice. In the high choice condition, participants were led to believe that a pill they were taking produced side-effect feelings of pleasantness (“pleasantly excited” or “relaxed”), unpleasantness (“tense” or “unpleasantly sedated”) or produced no side effects. Participants in the no-choice condition received a pill that produced no side effects. The results revealed the typical attitude change in favor of the counterattitudinal essay in the pleasantness and no side-effects conditions and, more important, an attenuation of the attitude change in the unpleasantness side-effects condition. In fact, the attitudes in this latter condition were almost similar to those of the no-choice condition (see top panel of Figure 9). Very similar results were obtained by Losch and Cacioppo (1990) who used prism goggles instead of pills. Together with other findings pointing to negligible effects of arousal, these studies led to the conclusion that not the arousal, but rather negative (as opposed to positive) affect induced dissonance reduction (Higgins et al.; Losch & Cacioppo).

The original attributional interpretation of Higgins et al.’s (1979) misattribution results was that the expectation of unpleasant side effects gave participants the opportunity to mistakenly attribute their negative dissonance feelings to the pill rather than to their discrepant behavior, whereas this was not possible for the pill with pleasant or no side effects. In these latter cases, negative feelings created by the dissonant state could not be explained away by the expectation of pleasant or no side effects; perhaps the pleasant expectation might have further increased the discrepancy with the negative feelings, exacerbating the dissonance.

Simulation. In the present conception, however, we assume that it is participants’ genuine belief in the

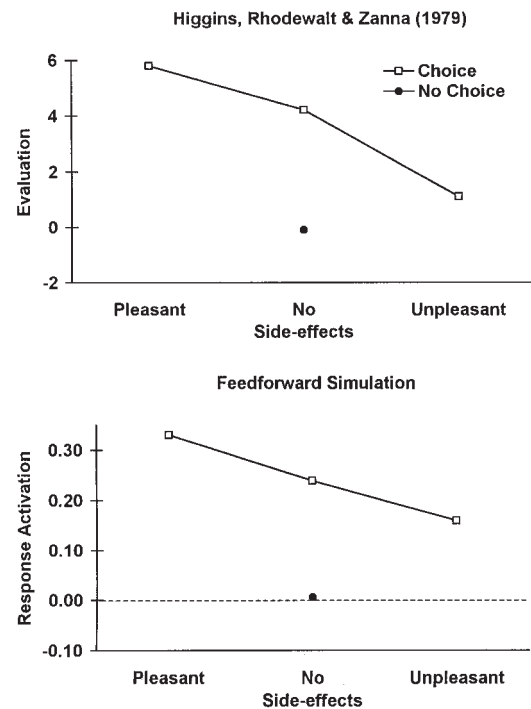


Figure 9. Human data (top) and feedforward simulation (bottom). The broken line shows the attitude prior to the experiment. The human data are from Table 2 and text in “Dissonance motivation: Its nature, persistence, and reinstatement,” by E. T. Higgins, F. Rhodewalt, & M. P. Zanna, 1979, *Journal of Experimental Social Psychology*, 15, 16–34. Copyright 1979 by Academic Press. Adapted with permission.

positive or negative emotion elicited by the alleged side effect that drives most of the dissonance reduction. That participants presumably believed the emotional side effects of the pills is corroborated, among others, by medical research indicating that for a wide range of afflictions, including pain, high blood pressure, asthma, and cough, roughly 30% to 40% of patients experience relief after taking a placebo (Beecher, 1955). In a summary of placebo research, Ross and Olson (1981) concluded that the reported effects of a placebo drug are often similar and typically in the same direction as those of the active drug, although reverse placebo effects have also been reported. Most researchers believe that placebo effects are mainly driven by patient’s expectations, although Totman (1976) presented a cognitive dissonance account that stresses justification for one’s investment in the treatment as a mediating factor.

Our interpretation that participants experienced the reported side effects of the pills is further supported by a manipulation check of the pill labels conducted aside the original study by Higgins et al. (1979). Fourteen additional participants who did not participate in the main study rated the side effects on two 7-point scales ranging from 1 (*positive*) to 7 (*negative*) and from 1 (*pleasant*) to 7 (*unpleasant*). The

mean-reversed ratings are given between parentheses in Table 2 and show that pills with pleasant side effects were experienced as pleasant (unlike our survey that indicated only moderate affect), whereas pills with unpleasant side effects were experienced as unpleasant (like our survey). Because these ratings were obtained from the same population in the same context and period as the original experiment of Higgins et al. (1979), they were taken as basis for coding the affective outcomes.⁴ Our survey data further indicate that pills with no side effects would be experienced as neutral. Note that the proposed coding of the affective outcomes is crucial in obtaining the simulations results reported next. For instance, without the assumed positive feelings given the pleasant side effect, the results in this condition could not be replicated.

In the simulation, four factors were simulated—essay topic, pleasant drug, unpleasant drug, and enforcement. As noted earlier, we assume that participants are misled in believing that they actually feel the side effects supposedly generated by the drug. All the other pre-experimental specifications are identical to the forced compliance simulation discussed earlier (see Table 2).

After running these pre-experimental specifications through the model, the connections with the essay topic were negligible, but enforcement had a positive connection with behavior and a negative one with affect (see Table 3). More important, although the behavioral connections of both types of drugs were zero, the affective connections were positive for the pleasant drug and negative for the unpleasant drug. Thus, as intended, the drugs had only an emotional impact.

Simulation Results. Figure 9 (bottom panel) depicts the simulation results. As can be seen, the effects of the presumed side effects of the pills in the choice condition differed significantly, $F(2,2997) = 625.51$. As expected, t tests revealed that all conditions differed from each other, $t_s = 18.95\text{--}68.02$. Most positive attitudes in favor of the counterattitudinal essay are found with pleasant side effects, somewhat less with no side effects, and

even less with unpleasant side effects. No attitude changes were found in the no-choice condition. In the no-choice condition, there was less attitude change than in the unpleasant side-effects condition, $t = 37.02$. This latter difference did not reach significance in the human data from Higgins et al. (1979), although it did in the study by Losch and Cacioppo (1990).

As in the earlier forced compliance simulation, the feedforward mechanism responsible for the strong positive attitude change was the fact that the positive feelings provided less justification (as they add up to the dissonance created by the discrepant behavior), whereas the negative feelings provided more justification (as they cancel the dissonance created by the behavior). This generated respectively an upward and downward adjustment, and a positive and negative attitude change. Moreover, in the no-choice condition, the additional external constrained provided sufficient justification for the discrepant behavior, leading to the least attitude change overall.

Robustness and Model Comparisons

Before ending this series of simulations, we first discuss the strength and robustness of our feedforward implementation and point out its potential weaknesses. Next, we compare the feedforward simulations with the earlier consonant model of Shultz and Lepper (1996).

Robustness of the Simulations

To what extent can the feedforward simulations replicate the experimental data, and are these simulations robust, that is, do they stand up against variations in the specified learning histories prior and during the experiment? To address this question, we assessed the overall fit of various simulations by computing correlations between the mean human data reported in the dissonance experiments and the means of the simulations. These correlations are merely indicative, as the number of means (4 or 6) is too few to obtain reliable differences between correlations. Therefore, we also analyzed the interaction or main F test of interest for each experiment (see Table 4 for a list of these effects). Although we also conducted t tests to make direct comparisons, these are not reported as long as the results are identical to those mentioned earlier for the specific paradigms.

To determine a standard of comparison for the robustness analyses, we first assessed the overall fit of the feedforward simulations previously discussed. As can be seen in the “Overall Fit” portion of Table 4, the feedforward simulations resulted in very high correla-

⁴ Pilot testing of a replication of the Higgins et al. (1979) misattribution experiment showed that more than one half of our psychology freshmen did not believe the side-effect manipulation of the pill. This underscores the temporal and cultural character of some of these misattribution manipulations. However, another pilot study revealed that our freshmen did believe the manipulation of new-age-like “brain frequencies” emitted by prism goggles (Losch & Cacioppo, 1990). A manipulation check measuring how they felt after wearing the goggles indicated that they felt significantly more pleasant after wearing goggles supposedly emitting “positive arousal” than after wearing goggles emitting “negative arousal,” $t(29) = 1.78, p < .05$ (one-sided).

Table 4. Overall Fit, Robustness and Model Comparison

Simulation	1	2	3	4	5	6	7	Mean
	Overall Fit							
Feedforward Model	.96	1.00	.94	.92	.91	.88	.95	.94
	Robustness							
Covariation Level								
(20% Zero Output)	.87 ^b	.99	.91	.97	.90	.89	.98	.93
(40% Zero Output)	.58 ^b	.92	.87	.98	.83	.89	.99	.87
Trial Frequency								
(5 Times) ^a	.93	.96	.91	.81	.91	.91	.97	.91
(All Pre-exp. 10)	.93	.97	.95	.94	.91	.89	.99	.94
Noise								
(<i>SD</i> = .40)	.95	.99	.93	.93	.90	.88	.96	.93
Affect								
Linear Function	.78 ^x	.40 ^x	.29 ^x	-.30 ^x	.78	.88	.50 ^x	.47
No Affect	.56 ^x	.16 ^x	.17 ^x	-.84 ^x	.94	.58 ^x	.70 ^x	.32
	Alternative Model							
Consonance Model	.94	.94	.92	—	.97	—	—	.94

Note: Cell entries are correlations between means of 20 replications of 50 random simulation runs and means of experiments as indicated by the same number code in Tables 2–3; the ^x superscript denotes if the effect of interest is not significant; these effects were the interactions (1) surveillance × threat, (2) initiation × shock, (3) choice × payment, (4) payment × mood, (5) alternative × difficulty, and the main effects of (6) mood and (7) side effect.

^aLearning rate = .01. ^bCrossover interaction.

tions ranging from .88 to 1.00, (mean $r = .94$) and, as noted earlier, replicated all F tests of interest.

Next, we ran a number of robustness simulations (each with 20 replications across the 50 random runs) to explore whether some variations in the sometimes arbitrary model specifications potentially reduced the fit. The results of these analyses are listed in the “Robustness” portion of Table 4. Note that when the correlations of these variations differ only slightly (less than 0.05) from those of the original simulations, this typically means that the pattern of the simulation means is very similar.

Degree of Covariation. To what extent does a perfect covariation between causes and outcomes as specified in the simulations matter? To explore this issue, we decreased the degree of covariation between causes and outcomes by randomly reducing 20% or 40% of the outcome activations to zero. This procedure reduces a perfect covariation of 1.00 to weaker covariation levels of 0.80 and 0.60 respectively; whereas lower covariations are reduced by similar proportions. With 20% reduction, the mean correlation across all experiments showed a decrease of -0.01 (mean $r = .93$); with 40% reduction, the mean correlation decreased by -0.07 (mean $r = .87$). As can be seen in Table 4, in most cases the decrease was marginal and substantial only for the prohibition experiment (Simulation 1). The F tests of interest remained significant in all simulations, and their pattern did not change substantially, except in the prohibition simulation where the interaction showed a crossover that did not appear in the

original experiment. That is, although in the original experiment, severity of threat did not differ under surveillance; in the simulations with reduced covariation, there appeared a typical dissonance effect or increased attitude change given weaker threat, $t = 13.50$ – 26.87 for a 20% and 40% reduction respectively.

Trial Frequencies. To what extent do the exact frequencies in the simulations matter? To study this question, we increased the number of all trials prior and during the experimental phase five times (the learning rate parameter was accordingly reduced to 0.01). As an alternative, we set all frequencies prior to the experimental phase to 10 (with original learning rate of 0.10). Both interventions had little effect (mean $r = .91$ – $.94$). All F tests were significant, and the pattern of these effects was similar to the original feedforward simulations.

Noise. To what extent does the degree of noise matter? The higher the noise, the smaller the overlap between causal factors before and during the experimental trials. To address this issue, we increased all noise (and accordingly reduced the overlap) by drawing from a normalized distribution with standard deviation of 0.40 instead of 0.20. As can be seen, the correlations did not change substantially (mean $r = .93$) and all F tests remained significant with a similar pattern of the simulation results.

Affect. One of the most novel assumptions of our model concerns the affective outcomes. To assess

whether the current threshold mapping (that is sensitive to extreme affect only) is crucial, we tested a smoother, linear mapping for the coding of the affective outcomes during the experiments. Specifically, the 1–7 scale ratings of our survey were subtracted by 4 (the scale midpoint) and then divided by 2, which yields a continuous coding between approximately –1 and +1. In addition, we also tested how deleting the affective outcomes altogether would influence the simulation results. As one would expect, these interventions had a detrimental impact and most correlations dropped substantially (mean $r = .32-.47$). Moreover, most of the F tests did not attain significance. This deterioration of the simulations is in line with our emphasis on the crucial role of extreme affect when external constraints are combined (Simulation 1–3) or in mood manipulation studies (Simulations 4, 6, & 7).

In sum, the simulations were immune to substantial variations in the learning history such as a reduced cause–outcome covariation, increased or equalized frequencies, and additional noise prior to the experiment. However, as might be expected, the simulations were very vulnerable to the changes in affective output coding, which therefore appear crucial in our model.

Comparison with the Consonance Model

How does the feedforward model compare against Shultz and Lepper's (1996) consonance model, which was the first full-fledged connectionist model of cognitive dissonance? Shultz and Lepper proposed that the motive to reduce cognitive dissonance and to seek cognitive consistency can be usefully modeled by a constraint satisfaction network (e.g., McClelland & Rumelhart, 1988; Read & Marcus-Newhall, 1993; Spellman & Holyoak, 1992; Thagard, 1992). Basically, their consonance model involves the simultaneous satisfaction of multiple, sometimes conflicting constraints on an individual's cognitions, including the attitude itself, external factors, and the behavior toward the object (but not the attitude object or any emotional reactions). These constraints are represented by relations or connections in the network that include "logical implication, causal relations, psychological implication, expectation and association" (p. 222). The connections impose constraints that are soft rather than hard, so that they are desirable, but not essential to satisfy.

The overall fit of the consonance model was obtained for the simulations conducted by Shultz and Lepper (1996) with the least random noise (rand% = .1) that provided the best fit with the data. As can be seen in Table 4 (bottom panel), Shultz and Lepper's

consonance simulations achieved very high correlations with the human data, which ranged from 0.92 to 0.97. Although the model provided good fits with the empirical data, the consonance model has a number of important shortcomings.

First, perhaps the most important shortcoming mentioned earlier is that the consonance model has no learning mechanism. As acknowledged by Shultz and Lepper (1996), "the process of creating the network is not usually modeled, presumably because it is not sufficiently understood psychologically" (p. 220). Thus, their model is nonadaptive as the connections have to be handset by the experimenter and do not develop automatically from prior learning.

Second, the consonance model commits all major aspects of dissonance and attitude change to temporary changes of activation in the network. Hence, the model reflects only a short-lived mental state of cognitive dissonance that occurs only when all relevant conflicting beliefs and constraints are activated (consciously or subconsciously) in the individual's mind. However, this is contradicted by the data. Dissonance effects persist over time even when attitude change was measured during a second, ostensibly unrelated experiment where the experimental pressures such as threat, payment, and others were absent (Festinger & Carlsmith, 1959), sometimes more than several weeks later (Collins & Hoyt, 1972; Freedman, 1965; Higgins et al., 1979).

Third, another important shortcoming of the consonance model refers to the manner in which Shultz and Lepper (1996) hand coded the connections in their network. They claimed that the connections between nodes were specified according to their pairwise relations. However, that is not what they did. Some connections require consideration of much more than the pair of nodes they connect, resting instead on the characteristics of the whole situation. To illustrate, in one of their simulations Shultz and Lepper specified a positive connection between an attractive toy and an adult's threat not to play with it, because "the better liked the toy, the more threat would be required to prevent play" (p. 226). This rationale uses the additional knowledge that play should be or was prevented in some situations, another aspect that goes beyond the two cognitions: toy and threat. By using more information than that of the two cognitions, Shultz and Lepper violated an important principle held in many connectionist models that local information on relations should be encoded in the system at a lower level only and that higher level observable characteristics should emerge from this local information (Cleeremans & French, 1996). It is this latter property that made connectionism so powerful and attractive because it obliterates the need for a supervisory homunculus in the brain.

Concluding Comments

Summary and Contribution

This article proposed an adaptive learning approach based on the perspective, radically different from traditional dissonance theories, that cognitive dissonance is mainly driven by a discrepancy between our mental representations of the dissonant situation and the reality of that situation. In order to survive and prosper, this discrepancy must be reduced. One way to do this is by changing the situation through adjustments in our or other's behaviors; another way is by making adjustments in our mental representations of the situation—the topic we focused on. We suggested that changing one's mental representations implies that adjustments are made in the causal explanations for our behavior and emotions with respect to the attitude object. This learning perspective reflects the view of humans as rational adaptive cognizers who seek the reasons and explanations for their thoughts, feelings, and behaviors.

These ideas were implemented in one of the most simple and robust network models using the error-correcting delta learning algorithm, the feedforward architecture (McClelland & Rumelhart, 1988). The delta algorithm in adaptive connectionist models accounts for a wide variety of phenomena, such as animal conditioning (Rescorla & Wagner, 1972), human causal learning and categorization (Allan, 1993; Estes et al., 1989; Shanks, 1991), and a whole series of phenomena in social cognition, including impression formation, assimilation and contrast, causal attribution, and attitude formation and change (Read & Montoya, 1999; Smith & DeCoster, 1998; Van Overwalle, 1998; Van Overwalle et al., 2001; Van Rooy et al., 2002). The present model is thus a member of a growing unifying connectionist theory of cognitive change.

Armed with these basic connectionist learning principles, the proposed network model was able to reproduce the findings of major representative paradigms in cognitive dissonance research, as well as novel findings that highlighted the role of affect in cognitive dissonance (Jordens & Van Overwalle, 2001, 2002). This suggests that an adaptive learning mechanism can underlie many findings of cognitive dissonance. The major achievement of our approach is that the model can reproduce participants' beliefs and attitudes prior to the experiment by simulating how the connections between concepts are created and strengthened after repeated exposures of co-occurrences between the attitude object and the person's behavioral and affective outcomes. In addition, the network reflects long-term connection changes, which is consistent with data showing that attitude

change after dissonant experiences persist over time. These are capacities that are shared by most adaptive network models but that are absent in other models (e.g., Shultz & Lepper, 1996).

The proposed model was inspired by the attributional perspective of cognitive dissonance (Cooper & Fazio, 1984). However, Cooper and Fazio's model was reproduced with some modifications. First, rather than focusing on personal causality, our model stressed attributions of deviant behavior and emotion to the attitude object as a psychological means of dissonance reduction. This notion is more consistent with current activation spreading models of attitudes (Ostrom et al., 1994). Second, the model emphasizes unexpected rather than undesirable outcomes. This proposal is closer to Festinger's (1957) original idea that dissonance arises when information about the environment or the self disconfirms cognitions or expectations (see also, Festinger et al., 1956). Dissonance was implemented in the model as the error that drives the weight adjustments. Third, instead of focusing on negative arousal as the instigator of an attributional analysis, we assumed that participants directly use information on their feelings of (un)pleasantness and discomfort for making attitude judgments. This latter psychological approach is more in line with recent theorizing on the role of emotions in cognitive dissonance (Elliot & Devine, 1994; Higgins et al., 1979; Losch & Cacioppo, 1990), appraisal (Frijda, 1986; Ortony et al., 1988; Roseman, 1991; Smith & Ellsworth, 1985), attribution (Weiner, 1986), and cognitive judgments (Schwarz, 1990). The present simulations suggest that especially the most intense negative or positive emotions determine the attributional analysis and dissonance process.

The inclusion of an affective outcome also enabled the feedforward network to reproduce the so-called reinforcement effect in the initiation and forced compliance paradigms (Simulation 2 & 3), which was not predicted by the original dissonance theory (Festinger, 1957) or the attributional reformulation (Cooper & Fazio, 1984). As anticipated by our adaptive approach, reinforcement effects were found only when dissonance was counterbalanced by strong negative feelings of unpleasantness and discomfort. In addition, the idea of affective outcomes generated novel predictions concerning the role of mood in dissonance reduction. Recent studies by Jordens and Van Overwalle (2001, 2002) demonstrated that mood may attenuate or increase attitude change and even reverse the reinforcement effect, in line with the predictions of the model (Simulations 4 & 6).

Implications for Other Theories

The affective nodes in the present network allow to integrate models that explore more generally the influ-

ence of mood on social judgments: affective priming (Bower, 1981; Isen, 1984) and affect as information (Schwarz & Clore, 1983, Schwarz, 1990). The affect priming model posits that affective states lead to a mood-congruent attention, encoding, and interpretation of social judgments. This is incorporated in the learning phase of our model where the connections between attitude object and affect are developed. The affect-as-information model suggests that people evaluate objects during retrieval by asking themselves how they feel about it. This mechanism is implemented in the test phase of our network model, where the affective output is tested after cuing the attitude object.

The present network can also simulate the role of the self in cognitive dissonance. According to the self-consistency model (Aronson, 1968; Thibodeau & Aronson, 1992), people hold expectancies for competent and moral behavior that lead to dissonance if they perceive a discrepancy with their self-expectations. According to Stone and Cooper (2001), this only happens if the threatened aspect of the self is related to the attitude object. In contrast, the self-affirmation model (Steele, 1988) posits that people attempt to restore the moral and adaptive integrity of the overall self-esteem by focusing on positive aspects of the self. According to Stone and Cooper, this discrepancy can be reduced only by focusing on positive features of the self that are unrelated to the attitude object. Based on the suggestions by Stone and Cooper, the network is able to encompass both self-consistency and self-affirmation. Self-consistency (Glass, 1964; Stone, 1999) is simulated by assuming that related feedback about the self is represented in the same (set of) nodes as chronic self-esteem. In contrast, self-affirmation (Blanton, Cooper, Skurnik, & Aronson, 1997; Steele, Spencer, & Lynch, 1993) is simulated by representing irrelevant feedback on the self in a different (set of) nodes. However, the degree of overlap between self-related feedback and the chronic self-concept to explain these opposing effects is still very much an empirical question (see Stone & Cooper), as well as the influence of trivialization on self-affirmation effects (Simon, Greenberg, & Brehm, 1995).

Limitations and Predictions

The present approach also has important limitations. Perhaps some simplifying assumptions and criteria on which the simulations rest might have been wrong. Because many variables were unknown, they were chosen somewhat arbitrary in the model specifications, so that many degrees of freedom remain. Although the robustness analyses suggest that many of these choices were in fact of little concern; some of them contain some arbitrary quality and yet were critical for an adequate fit.

One of these controversial specifications involves the coding of weaker treatment levels. In three paradigms, we choose 20%, 50% and 70% of the default activation value to simulate a weaker level of some external constraints (e.g., threat, shock, & payment). These specifications resulted in the best fit of the simulations, but we have no independent data to substantiate these choices. However, it seems plausible that all weaker treatments are not necessarily identical. Perhaps, it is not a coincidence that the activation of weaker negative factors (i.e., threat and shock) was higher than that of the weaker positive factor (i.e., payment), because negative stimuli are typically experienced more intensely than positive stimuli. Nevertheless, the impact of different levels of weaker treatment levels is an open question for future research.

Another controversial specification involves the affective coding. This aspect of the model was most novel and was substantiated with survey data as well with some new empirical dissonance research inspired by our model (Jordens & Van Overwalle, 2001, 2002). This latter research demonstrated that by inducing positive or negative emotions, the typical effects found in earlier research can be attenuated or eliminated. We consider this as convincing evidence for our affective hypothesis, although more research is needed with respect to other paradigms that were not yet tested, such as prohibition, initiation, and misattribution.

Another severe limitation dealt with most adaptive connectionist models is known as “catastrophic interference” (French, 1999; McCloskey & Cohen, 1989; Ratcliff, 1990), which is the tendency of neural networks to forget abruptly and completely previously learned information in the presence of new input. In the simulations, this shortcoming was avoided by presenting only a limited number of experimental trials after the pre-experimental trials. However, this limitation is perhaps untenable for a realistic model of cognitive dissonance and attitude change in general where people are sometimes quite resistant to adjust their behavior (e.g., quit smoking). 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 information is processed in the hippocampus and old 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, 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. It is clear that the present network

fits within the rapid hippocampal system and that only the strong connections with the attitude object will survive transference to the slow and long-lasting neocortical system, and that the weak episodic connections will fade out.

There are other limitations in our model as well. One of the things the model has nothing to say about is the role of physiological arousal that may concur with experiences of dissonance and discomfort. However, this is not very problematic as recent developments in cognitive dissonance research suggest that not arousal, but rather negative affect is the factor that stimulates people to seek an explanation for their dissonant state (Elliot & Devine, 1994; Higgins et al., 1979; Losch & Cacioppo, 1990). More important is that the model is unable to explain how aversive emotions for deviant behavior are generated. In addition, although the model can encode and process consonant information resulting from selective search in support of an existing attitude (Frey, 1986; Jonas, Schulz-Hardt, Frey, & Thelen, 2001), the active search itself cannot be modeled, because this involves controlled and strategic processes that go beyond the capacities of most connectionist models that simulate mainly implicit and automatic associative processes (Smith & DeCoster, 2000).

Yet, we believe that the present simulations have great heuristic value. The specifications of the learning histories are directly testable. Moreover, as mentioned earlier, the model makes a number of novel predictions on the role of the level of weaker treatments and affect, some of which have already provided empirical support for the model. Assuming that the feedforward model is sufficiently adequate and rich, testing the model with new data and—when necessary—adapting it, should result in a greater accuracy and a better insight in the processes underlying cognitive dissonance.

References

- Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin*, *114*, 435–448.
- Ans, B., & Rousset, S. (1997). Avoiding catastrophic forgetting by coupling two reverberating neural networks. *Académie des Sciences de la vie*, *320*, 989–997.
- Aronson, E. (1968). Dissonance theory: Progress and problems. In R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, & P. H. Tannebaum (Eds.), *Theories of cognitive consistency: A sourcebook* (pp. 5–27). Chicago: Rand McNally.
- Beecher, H. K. (1955). The powerful placebo. *Journal of the American Medical Association*, *159*, 1602–1606.
- Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 1–62). New York: Academic.
- 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.
- Blanton, H., Cooper, J., Skurnik, I., & Aronson, J. (1997). When bad things happen to good feedback: Exacerbating the need for self-justification with self-affirmation. *Personality and Social Psychology Bulletin*, *23*, 684–692.
- Bower, G. H. (1981). Mood and memory. *American Psychologist*, *36*, 129–148.
- Brehm, J. W. (1956). Post-decisional changes in the desirability of choice alternatives. *Journal of Abnormal and Social Psychology*, *52*, 384–389.
- Calder, B. J., Ross, M., & Insko, C. A. (1973). Attitude change and attitude attribution: effects of incentive, choice, and consequences. *Journal of Personality and Social Psychology*, *25*, 84–99.
- Cleeremans, A., & French, R. M. (1996). From chicken squawking to cognition: Levels of description and the computational approach in psychology. *Psychologica Belgica*, *36*, 5–30.
- Collins, B. E., & Hoyt, M. F. (1972). Personal responsibility-for-consequences: An integration and extension of the “forced compliance” literature. *Journal of Experimental Social Psychology*, *8*, 558–593.
- Cooper, J., & Fazio, R. H. (1984). A new look at dissonance theory. In L. Berkowitz (Ed.) *Advances in experimental social psychology* (Vol. 17, pp. 229–266). New York: Academic.
- Cooper, J., Fazio, R. H., & Rhodewalt, F. (1978). Dissonance and humor: Evidence for the undifferentiated nature of dissonance arousal. *Journal of Personality and Social Psychology*, *36*, 280–285.
- Elliot, A. J., & Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, *67*, 382–394.
- Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology, Learning, Memory and Cognition*, *15*, 556–571.
- Festinger, L. (1957) *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson.
- Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58*, 203–210.
- Festinger, L., Riecken, H. W., & Schachter, S. (1956). *When prophecy fails*. Minneapolis: University of Minnesota Press.
- Freedman, J. L. (1965). Long-term behavioral effects of cognitive dissonance. *Journal of Experimental Social Psychology*, *1*, 145–155.
- French, R. (1997). Pseudo-recurrent connectionist networks: An approach to the “sensitivity-stability” dilemma. *Connection Science*, *9*, 353–379.
- French, R. (1999). Catastrophic forgetting in neural networks. *Trends in Cognitive Sciences*, *3*, 128–135.
- Frey, D. (1986). Recent research on selective exposure to information. In L. Berkowitz (Ed.) *Advances in experimental social psychology* (Vol. 19, pp. 41–80). New York: Academic.
- Frijda, N. (1986). *The emotions*. New York: Cambridge University Press.
- Gerard, H. B., & Mathewson, G. C. (1966). The effects of severity of initiation on liking for a group: A replication. *Journal of Experimental Social Psychology*, *2*, 278–287.
- Glass, D. (1964). Changes in liking as a means of reducing cognitive discrepancies between self-esteem and aggression. *Journal of Personality*, *32*, 531–549.
- Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227–247.
- Harmon-Jones, E., & Mills, J. (1999). *Cognitive dissonance: Progress on a pivotal theory in social psychology*. Washington, DC: American Psychological Association.

- Higgins, E. T., Rhodewalt, F., & Zanna, M. P. (1979). Dissonance motivation: Its nature, persistence, and reinstatement. *Journal of Experimental Social Psychology, 15*, 16–34.
- Isen, A. M. (1984). Towards understanding the role of affect in cognition. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 3, pp. 179–236). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology, 80*, 557–571.
- Jones, E. E. (1985). Major developments in social psychology during the past five decades. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (3rd ed., pp. 47–108). New York: Random House.
- Jordens, K., & Van Overwalle, F. (2001). Een empirische toetsing van een feedforward connectionistisch model van cognitieve dissonantie: De rol van affect in het geïnduceerd-inwillingsparadigma [An empirical test of a feedforward connectionist model of cognitive dissonance: The role of affect in the induced compliance paradigm]. In D. A. Stapel, C. Martijn, E. van Dijk, & A. Dijksterhuis (Eds.), *Fundamentele sociale psychologie* (Vol. 15, pp. 91–102). Delft, The Netherlands: Eburon.
- Jordens, K., & Van Overwalle, F. (2002). *Cognitive dissonance and affect: An empirical test of a connectionist account*. Manuscript in preparation.
- Kelley, H. H. (1971). Attribution in social interaction. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 1–26). Morristown, NJ: General Learning Press.
- Kidd, R. F., & Berkowitz, L. (1976). Effect of dissonance arousal on helpfulness. *Journal of Personality and Social Psychology, 33*, 613–622.
- Linder, D. E., Cooper, J., & Jones, E. E. (1967). Decision freedom as a determinant of the role of incentive magnitude in attitude change. *Journal of Personality and Social Psychology, 6*, 245–254.
- Lord, C. G. (1992). Was cognitive dissonance theory a mistake? *Psychological Inquiry, 3*, 339–341.
- Losch, M. E., & Cacioppo, J. T. (1990). Cognitive dissonance may enhance sympathetic tonus, but attitudes are changes to reduce negative affect rather than arousal. *Journal of Experimental Social Psychology, 26*, 289–304.
- 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. 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.
- Mutate, H., Arcediano, F., & Miller, R. R. (1996). Test question modulates cue competition between causes and between effects. *Journal of Experimental Psychology, Learning, Memory and Cognition, 22*, 182–196.
- Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. New York: Cambridge University Press.
- Ostrom, T. M., Skowronski, J. J., & Nowak, A. (1994). In P. G. Devine (Ed.), *Social cognition: Impact on social psychology* (pp. 195–251). San Diego, CA: Academic.
- 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.
- 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.
- Rhodewalt, F., & Comer, R. (1979). Induced-compliance attitude change: Once more with feeling. *Journal of Experimental Social Psychology, 15*, 35–47.
- Roseman, I. J. (1991). Appraisal determinants of discrete emotions. *Cognition and Emotion, 5*, 161–200.
- Rosenberg, M. J., & Hovland, C. I. (1960). Cognitive, affective and behavioral components of attitudes. In C. I. Hovland & M. J. Rosenberg (Eds.), *Attitude organization and change: An analysis of consistency among attitude components* (pp. 1–14). New Haven, CT: Yale University Press.
- Ross, M., & Olson, J. M. (1981). An expectancy attribution model of the effects of placebos. *Psychological Review, 88*, 408–437.
- Sakai, H. (1999). A multiplicative power-function model of cognitive dissonance: Toward an integrated theory of cognition, emotion, and behavior after Leon Festinger. In E. Harmon-Jones & J. Mills (Eds.), *Cognitive dissonance: Progress on a pivotal theory in social psychology*. Washington, DC: American Psychological Association.
- Schwarz, N. (1990). Feelings as information: Informational and motivational functions of affective states. In E. T. Higgins & R. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behaviour* (Vol. 2, pp. 527–561). New York: Guilford.
- Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*, 513–523.
- Shaffer, D. R. (1975). Some effects of consonant and dissonant attitudinal advocacy on initial salience and attitude change. *Journal of Personality and Social Psychology, 32*, 160–168.
- Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17*, 433–443.
- Shanks, D. R. (1993). Human instrumental learning: A critical review of data and theory. *British Journal of Psychology, 84*, 319–354.
- Sherman, S. J. (1970). Attitudinal effects of unforeseen consequences. *Journal of Personality and Social Psychology, 16*, 510–520.
- Shultz, T., & Lepper, M. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review, 2*, 219–240.
- Shultz, T. R., Léveillé, E., & Lepper, M. R. (1999). Free choice and cognitive dissonance revisited: Choosing “lesser evils” versus “greater goods.” *Personality and Social Psychology Bulletin, 25*, 40–48.
- Simon, L., Greenberg, J., & Brehm, J. (1995). Trivialization: The forgotten mode of dissonance reduction. *Journal of Personality and Social Psychology, 68*, 247–260.
- Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology, 48*, 813–838.
- 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: Guilford.
- Spellman, B. A., & Holyoak, K. J. (1992). If Saddam is Hitler who is George Bush? Analogical mapping between systems of social roles. *Journal of Personality and Social Psychology*, *62*, 913–933.
- Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261–302). San Diego, CA: Academic.
- Steele, C. M., Spencer, S. J., & Lynch, M. (1993). Self-image resilience and dissonance: The role of affirmational resources. *Journal of Personality and Social Psychology*, *64*, 885–896.
- Stone, J. (1999). What exactly have I done? The role of the self-attribute accessibility in dissonance. In E. Harmon-Jones & J. Mills (Eds.), *Cognitive dissonance: Progress on a pivotal theory in social psychology* (pp. 175–200). Washington, DC: American Psychological Association.
- Stone, J., & Cooper, J. (2001). A self-standards model of cognitive dissonance. *Journal of Experimental Social Psychology*, *37*, 228–243.
- Thagard, P. (1992). *Conceptual revolutions*. Princeton, NJ: Princeton University Press.
- Thibodeau, R., & Aronson, E. (1992). Taking a closer look: Reasserting the role of the self-concept in dissonance theory. *Personality and Social Psychology Bulletin*, *18*, 591–602.
- Totman, R. G. (1976). Cognitive dissonance and the placebo response. *European Journal of Social Psychology*, *5*, 119–125.
- 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., Labiouse, C., & French, R. (2001). *Connectionist exploration in social cognition*. Unpublished Manuscript.
- 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. 142–171). New York: Lawrence Erlbaum Associates, Inc.
- Van Overwalle, F., & Van Rooy, D. (2001a). 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 Overwalle, F., & Van Rooy, D. (2001b). How one cause discounts or augments another: A connectionist account of causal competition. *Personality and Social Psychology Bulletin*, *27*, 1613–1626.
- Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C., & French, R. (2002). *A recurrent connectionist model of group biases*. Manuscript submitted for publication.
- Weiner, B. (1986). *An attributional theory of achievement motivation and emotion*. New York: Springer-Verlag.
- Zanna, M. P., & Cooper, J. (1974). Dissonance and the pill: An attributional approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology*, *29*, 703–709.
- Zanna, M. P., Higgins, E. T., & Taves, P. A. (1976). Is dissonance phenomenologically aversive? *Journal of Experimental Social Psychology*, *12*, 530–538.