Modeling Cognitive Dissonance Using a Recurrent Neural Network Model with Learning

Stephen J. Read (read@usc.edu)
Department of Psychology, University of Southern California
Los Angeles, CA 90089-1061

Brian M. Monroe (monroe@usc.edu)
Department of Psychology, University of Southern California
Los Angeles, CA 90089-1061

Abstract

This paper presents a recurrent neural network model of long term attitude change resulting from the reduction of cognitive dissonance. The model uses Contrastive Hebbian Learning (CHL) to capture changes in weight strength among cognitions resulting from dissonance reduction. Several authors have presented recurrent network models of dissonance reduction that capture the constraint satisfaction nature of dissonance. But as these models did not learn they could only model short-term attitude change represented by changes in activatio. In response, Van Overwalle and Jordens (2002) presented a feedforward model, with delta rule learning, in an attempt to capture long-term attitude change caused by dissonance reduction. However, the feedforward nature of their model created two problems. First, it could not capture the parallel constraint satisfaction mechanisms that underlie dissonance reduction. Second, and perhaps more important, the network was not able to "reason" backwards from its inconsistent behavior to its new attitude, but instead had to be explicitly taught its new attitude. The present model overcomes the weaknesses of the previous approaches to modeling dissonance reduction. Because it has learning it can represent long-term attitude change by weight change and because it is a recurrent model it can propagate changes from inconsistent behavior to the attitudes linked to that behavior.

Keywords: Coherence, Cognitive Consistency, Constraint Satisfaction, Neural networks, Connectionist models.

Introduction

One of the most famous and productive theories in social psychology has been Festinger's (1957) theory of Cognitive Dissonance. The term cognitive dissonance has entered everyday language and is often used to describe inconsistencies among cognitions. According to Cognitive Dissonance Theory, when two cognitions are dissonant with each other, an individual is motivated to reduce that dissonance. Festinger defined two cognitions as being in a dissonant relationship if the obverse of one cognition followed from the other. To date, researchers have used dissonance theory to generate a number of counter intuitive findings, such as finding that people like groups better the more painful is the initiation to join that group (Gerard & Mathewson, 1966) and finding that people like a boring task better if they agree to tell someone that the task is actually interesting for small amounts of money, compared to receiving a large amount (Festinger & Carlsmith, 1959).

Several authors (e.g., Read & Miller, 1994; Read, Vanman, & Miller, 1997; Shultz & Lepper, 1996; Simon & Holyoak, 2002) have argued that cognitive dissonance (Festinger, 1957) and related consistency phenomena (Abelson et al., 1968) can be modeled as a parallel constraint satisfaction process in a neural network, where the relevant cognitions are treated as nodes in the network and the consistent and inconsistent relations between cognitions are treated as excitatory and inhibitory relationships, respectively. Read, Vanman, and Miller (1997) have noted that parallel constraint satisfaction processes provide a computational implementation of the Gestalt processes that provide the theoretical underpinnings of dissonance theory. These authors, and particularly Shultz and Lepper (1996) have shown that a number of different Cognitive Dissonance findings can be successfully modeled in a parallel constraint satisfaction network.

However, recently Van Overwalle and Jordens (2002) have noted that these approaches to modeling cognitive dissonance can only model the immediate attitude and belief change in the specific experimental situation, but that they are incapable of modeling any long-term attitude or belief change that might result from resolving the dissonance. They rightly note that this is due to the lack of any kind of learning mechanism in the proposed models. All of these models use recurrent or feedback neural networks in which attitude and belief change is captured by changes in the activation of relevant nodes representing the key cognitions. However, none of these models have a learning mechanism that can modify the associations or links between cognitions as a result of the changes in pattern of activations. Such weight change would allow for the representation of long-term attitude change.

To attempt to remedy this lack, Van Overwalle and Jordens (2002) proposed a feedforward neural network model, with delta rule learning, that would change the links among cognitions in response to changes in the pattern of activations of the nodes. They argued that this model could successfully capture the long-term attitude change that would result from cognitive dissonance processes.

We agree that having a learning mechanism is critical for any adequate model of cognitive dissonance that wishes to address the issue of long-term attitude change resulting from dissonance reduction. However, there are two fundamental and interrelated problems with their proposed solution. First, the characterization of cognitive dissonance suggested by their model is a radical departure from the consensual understanding of cognitive dissonance processes. Rather than treating dissonance reduction processes as a Gestalt like seeking for good form and coherence (the historical view) or as a constraint satisfaction process (the modern rendition of Gestalt ideas of coherence (see Read, Vanman, & Miller, 1997; Simon & Holyoak, 2002), their model treats dissonance reduction purely as an error correcting learning process. Second, a perhaps more fundamental problem with their model is that the network does not model how a network might infer changes in the evaluation of an attitude object. Because the model is a feedforward network, it cannot "reason" backward from the inconsistent behavior to the underlying attitude. Instead during learning, the authors directly tell the network what their evaluative response is.

In a typical dissonance study, the subject is subtly induced to perform a behavior that is inconsistent with the behavior that they would expect, given their current attitude: for example, for a minimal and apparently insufficient inducement, they will agree to convince someone that a truly boring task is interesting (Festginger & Carlsmith, 1959) or they will write an essay espousing a position that contradicts their true attitude (Linder, Cooper, & Jones, 1967). Because the inducement is subtle the subject is typically not aware that they were actually induced by the experimenter. Lacking an obvious justification for their counter-attitudinal behavior, they then change their attitude to become more consistent with their previously "counter-attitudinal" behavior, thereby justifying the unexpected behavior. In modeling this experimental situation, Van Overwalle and Jordens do not have the network infer the attitude change as a function of dissonance reduction mechanisms (or constraint satisfaction processes) in the individual. Rather, they directly instruct the network that their attitude or evaluation has changed in ways consistent with typical experimental findings. Thus, what they have essentially done is show that their network is capable of using delta rule learning to modify an association in response to their direct training. The necessity of having to directly instruct their network is a result of their choosing to use a feedforward network. Because of this, they have no way to model how performing an unexpected behavior can modify the attitudes that might drive that behavior. (An example of the general structure of their network can be seen in Figure 1.)

In contrast, in our model, we use a recurrent network that does not receive any direct information about the network's new evaluation or attitude. Instead, we simply tell it that it performed a behavior counter to what would be expected, given it's previous experience. This results in a change in the patterns of activations, which the learning process then transforms into appropriate weight changes. The model changes the evaluations of the object and the relevant weights so that the previously unexpected behavior becomes consistent with the network's new evaluation and relations.

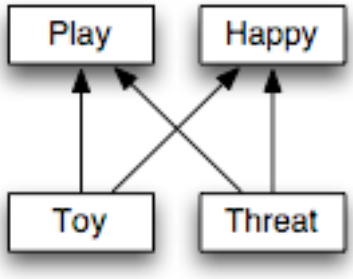

Figure 1: Example of Van Overwalle and Jordens (2002) network

Although we believe that Van Overwalle and Jordens model fails to capture fundamental aspects of dissonance processes, no other model of dissonance reduction attempts to model long-term attitude change. The current model seeks to address the weaknesses in both existing approaches to modeling cognitive dissonance. The proposed model is a recurrent (feedback) network with Contrastive Hebbian Learning. It can capture the constraint satisfaction processes that we believe are central to dissonance processes as well as capturing long-term attitude change represented by changes in the weights in the network.

Simulations

Network Overview

The simulations were done with the constraint satisfaction module (cs++) in the PDP++ neural network software (Dawson, O'Reilly, & McClelland, 2003; O'Reilly & Munakata, 2000). The cs++ module allows for the use of bidirectional weights and thus can function as a constraint satisfaction system. The models used a Contrastive Hebbian Learning (CHL) algorithm developed for the Boltzmann machine and then generalized by O'Reilly (1996). This algorithm compares the activation of the network in a plus phase (when both inputs and desired outputs are presented to the network) to its activation in a minus phase (when only the inputs are presented). CHL then adjusts weights to reduce the difference in activation between the two phases. One strength of Contrastive Hebbian Learning is that it allows for the use of hidden units in a network and adjusts weight to those units in a more biologically plausible way than does backpropagation. Backpropagation requires the use of a nonlocal error term and assumes that the error can be propagated back through multiple layers. CHL uses a local error term and thus does not require the propagation of an error signal.

We used the default sigmoidal activation function for the units in the networks, with activations limited to the range -1 to 1. Bias weights were set to 0. The learning rate was .20.

Simulation 1: Festinger and Carlsmith (1959) Counter-attitudinal Advocacy

In one of the first published studies of cognitive dissonance, participants first spent an hour doing an excruciatingly boring task. Then the experimenter entered and told them that the study was investigating how to motivate people to perform routine tasks. They were told that some of the previous participants had been told that the task was quite interesting. Suddenly, the lab supervisor rushed in and said that the assistant who typically told participants that the task was interesting had not shown up and that they needed someone to tell the next subject that the task was interesting. The experimenter asked the participants if they would take the assistant's place and tell the next participant that the task was interesting. Some of the participants were offered $20 to do this, but others only $1. All agreed to help. Then, in a separate context, all participants were asked to evaluate the study. Surprisingly, participants paid $1 liked the experiment more than participants paid $20. Festinger and Carlsmith argued that this happened because participants paid $1 experienced strong dissonance between their feeling that the task was boring and the fact that for a trivial amount of money they had tried to convince someone else that the task was actually interesting. So, to reduce their dissonance they changed their feelings about the task. In contrast, subjects who were given $20 could use the large payment to justify their behavior.

Network structure. The structure of the network is shown in Figure 2. One input node represents the boring task and the other input node represents the level of payment. The task node is linked by feedforward links to two paired evaluation nodes, a positive evaluation node and a negative evaluation node with a weight of -.5 between them, indicating that the two evaluations inhibit each other.

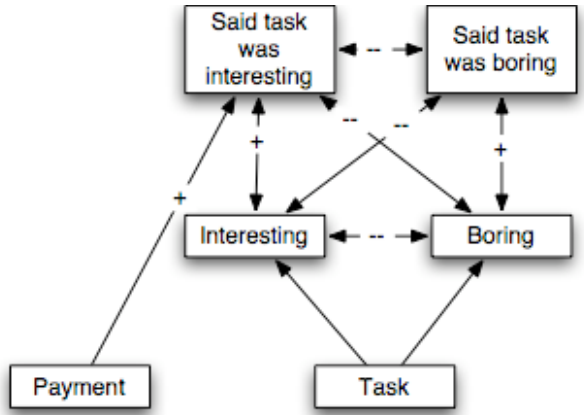

Figure 2: Network for Festinger and Carlsmith(1959)

This use of paired evaluations nodes is similar to the model of Shultz and Lepper (1996). We had two evaluation nodes that could be activated somewhat independently, because research indicates that evaluation is not strictly a bi-polar dimension, but that one can have somewhat separate positive and negative evaluations of an object.

These positive and negative evaluation nodes are then linked to the two behavioral alternatives: "said task was interesting" or "said task was boring." These two nodes have an inhibitory link of -1 between them so that under most circumstances only one of the nodes will have a positive activation. We have separate nodes for "said task was interesting" and "said task was boring", because we wanted to be able to separately and explicitly represent what the participant expected to do and what the participant actually did. So in this simulation, we wanted to see what would happen if we had a network that initially predicted that the participant would say that the task was boring (represented by a positive activation of the "said task was boring" node as a result of the spread of activation throughout the network), followed by the observation that the participant actually said the task was interesting (represented by a teaching activation to the "said task was interesting" node).

Further, the positive evaluation node has a positive link to "said task was interesting" and an inhibitory link to "said task was boring." The negative evaluation node has the reverse pattern of weights, with the negative evaluation node having a positive link to "said boring" and an inhibitory link to "said interesting." Finally, the payment node has a direct link to the "said interesting" node.

The network is set up so that in the dissonance simulation feedback from the environment or a "teaching" signal will only be applied to the two behavior nodes. In contrast to Van Overwalle and Jordens' (2002) model, in our model there is no direct feedback to the evaluation / affect nodes. This was done so that we could see if expectations about behavior and feedback about actual behavior could lead to indirect changes in evaluations, without the need to give direct feedback about affect. Another way to frame this difference between the two models is that in the Van Overwalle and Jordens' model, the network is explicitly told what it did and how it "felt" in response to that behavior, whereas in our model, the network is only explicitly told what it did and it must "infer" how it felt and presumably change it's evaluations to make them more consistent with its behavior. We would argue that this latter process is much more similar to the original process of dissonance reduction. Perhaps more important, we believe that our model is a much better representation of the actual experimental conditions in dissonance experiments.

In the typical dissonance experiment, the participant is not given any explicit feedback about how they feel, only about how they behaved. Any changes in affect and evaluation

would have to be driven by that difference in behavior. This is the logic of how we set up our model.

Initial learning and expectations. To insure sure that the model generated the correct set of expectations, we developed a set of training events that would insure that the network had a plausible set of weights that would lead to the expected outcome. After training with the events, the network behaved as expected. When the task was present and there was low payment (indicated by low activation of the payment node), the "said task was boring" node was more highly activated than the "said task was interesting" node. However, when the task was activated and the payment node was highly activated (indicating a payment of $20), then the "said task was interesting" node was more highly activated than the "said task was boring" node.

Results and discussion. Ten separate simulations were done with different random initializations of the weights before learning. Weights were initialized from a uniform distribution with a mean of 0 and a variance of .1.

Dissonance reduction was tested by first testing how the model would predict the participants' behavior and attitudes (before dissonance columns in Table 1) when the task was present and activation of the payment node was low. We then creating dissonance by having the model predict that it would produce the expected behavior (say the task was boring) and then receive feedback (activation of the appropriate node) that the participant actually said that the task was interesting. To simulate rumination and repeated consideration of the dissonance, we presented the dissonant behavior five times. That is, we turned on the task node and the "said task was interesting" node, let the network settle, and then updated the weights. This sequence was repeated a total of five times.

Table 1: Node activations before and after dissonance in Festinger & Carsmith (1959)

	Positive Eval.	Negative Eval.	Said Interesting	Said Boring
Before Dissonance	.26	.72	.40	.51
After Dissonance	.34	.66	.47	.42

Averaged across the 10 simulations with different random weight initializations, it is clear that after the "dissonance induction" (after dissonance columns in Table 1) the activation of the "said task was interesting" node was now higher than the activation of the "said task was boring" node and the activation of the positive evaluation node had increased and the evaluation of the negative evaluation node had decreased. Thus, the network did successfully "infer" a change in attitude evaluation from its dissonant or unexpected behavior. Unlike the model by Van Overwalle and Jordens (2002) we did not have to explicitly tell the network its new attitude. Instead the network inferred its new attitude from its behavior.

Simulation 2: Insufficient Justification: The Forbidden Toy Paradigm (Freedman, 1965)

In this paradigm a young child is brought into the experimental room and shown an attractive toy: an interesting robot. The experimenter than administers either a mild or a severe threat to the child to not play with the robot. The experimenter then either leaves the room or stays. So the child is in one of four conditions: (1) mild threat, no surveillance, (2) mild threat, surveillance, (3) severe threat, no surveillance, or (4) severe threat, surveillance. The child is then observed through a one-way mirror. In this initial observation, none of the children play with the robot.

Forty days later the child is brought back to the experiment room and then left alone in the room with the toy. The experimenters observed the child through a one-way mirror and recorded whether the child played with the toy. The researchers predicted that all of the children should play with the toy except for those who were initially in the mild threat, no surveillance condition. The rationale is that in all the other conditions, the child believes that they did not play with the toy because of the threat of punishment. Thus, they didn't need to rationalize why they didn't play with the attractive toy. There shouldn't have been any change in the child's liking for the toy and when they are later given an opportunity to play with the toy without any possibility of punishment they will happily play with the toy. In contrast, the argument is that for the children in the mild threat, no surveillance condition, their failure to play with the robot earlier, was perceived as inconsistent with the fact that they had only received a mild threat. To justify this inconsistency they would decrease their liking for the toy. As predicted, the only children who do not play with the toy were those who were given the mild threat and thought they were not watched.

Network structure. The structure of the network is shown in Figure 3. One input node represents the attractive toy, a second input node represents the level of threat, and the third input node represents the level of surveillance. The toy node is linked by feedforward links to two paired evaluation nodes, a positive evaluation node and a negative evaluation node with a weight of -.5 between them.

These positive and negative evaluation nodes are linked to the two behavioral alternatives: "play with the toy" or "do something else." These two nodes have an inhibitory link of -1 between them so that under most circumstances only one of the nodes will have a positive activation. We have separate nodes for "play" and "do something else", because we wanted to be able to separately and explicitly represent

what the child expected to do and what the child actually did. So, we wanted to see what would happen if we had a network that initially predicted that the child would play with the toy (represented by a positive activation of the "play with the toy" node), followed by the observation that the child did not play with the toy (represented by a teaching activation to the "do something else" node).

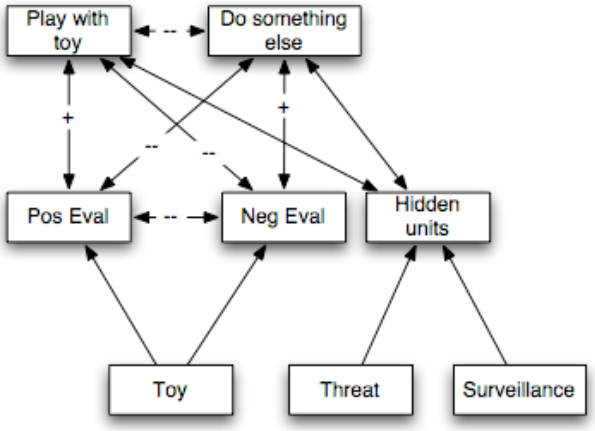

Figure 3: Network for Freedman(1965)

Further, the positive evaluation node has a positive link to "play with the toy" and an inhibitory link to "do something else." The negative evaluation node has the reverse pattern of weights, with the negative evaluation node having a positive link to "don't play with the toy" and an inhibitory link to "do something else."

The surveillance and threat nodes are connected to four hidden units that connect to "play with the toy" and "do something else." The hidden units are used in this model because we need to capture the idea that while a mild threat alone doesn't stop playing and surveillance alone doesn't stop playing, the conjunction of the two will stop playing. As in the first simulation, the network is set up so that in the dissonance simulation the "teaching" signal will only be applied to the two behavior nodes.

Initial learning and expectations. As in Simulation 1, we developed a learning history that would insure that the network had a plausible set of weights that would lead to the expected outcomes. After training with the events, the network behaved as expected. When the toy was present and there was only a mild threat, with no surveillance, the "play with toy" node was highly positively activated and the "do something else" node was negatively activated. However, in the other three combinations of surveillance and threat (mild threat, surveillance; severe threat, no surveillance, and severe threat, surveillance) the "do something else" node was highly positively activated and the "play with toy" node was not. Thus, the nodes were activated in the expected pattern.

Table 2: Node activations before and after dissonance in Freedman (1965)

	Positive Eval.	Negative Eval.	Play	Don't Play
Before Dissonance	.844	.123	.793	.186
After Dissonance	.810	.167	.675	.284

Results and discussion. As in Simulation 1, we ran ten different versions with different random starting weights and then averaged the results across the ten initializations. For each of the ten initializations, after learning we first turned on the toy node and set the threat to mild. As expected, the "play with toy node" was more highly activated than the "do something else node" (the before dissonance column in Table 2). Then for each initialization to test whether the predicted activation and weight changes would occur in the "dissonance" conditions, we ran the following sequence. We turned on the "toy" node and the mild surveillance node, and then turned on the "do something else" node. Since the network with this configuration of inputs should predict "play with toy" this activation of the "do something else" node was dissonant. We presented this sequence five times to the network. That is, we turned on the input nodes, the "do something else" node, let the network settle, and then updated the weights. At the end of this sequence, we then turned on only the "toy" node to see what the network would predict. The average pattern of activation for the behavior and evaluation nodes can be seen in Table 2. As predicted, after the dissonance experience (after dissonance column) the activation of the positive evaluation node decreased and the activation of the negative evaluation node increased, indicating decreased liking for the toy. Further, the activation of the "play with toy" node is now lower than the activation of the "do something else" node, predicting that, as was shown in the experiment, the child would not play with the toy after the mild threat, no surveillance experience. Thus, as in simulation 1 we successfully simulated changes in liking for the toy after dissonance without having to explicitly instruct the network about the current evaluation of the toy.

Discussion

This recurrent neural network model, with Contrastive Hebbian Learning, successfully modeled both the immediate and long-term attitude change that results from reduction of cognitive dissonance. Immediate attitude change is represented by changes in the temporary activation of concepts and evaluations and long-term attitude change is represented by changes in the weights among cognitions and evaluations.

Van Overwalle and Jordens (2002) had modeled long-term attitude change with delta rule learning in a feedforward network, but at the expense of not being able to capture the parallel constraint satisfaction nature of cognitive dissonance processes. Further, the feedforward nature of their network required that they explicitly tell the network what its new attitude was. Their network could not "infer" its new attitude from its behavior. In contrast, our recurrent network can both capture the parallel constraint satisfaction nature of dissonance as well as being able to model how the network could "infer" a new attitude from its behavior. Thus, we can combine the strengths of the two previous approaches to modeling cognitive dissonance, while avoiding their major weaknesses.

References

Abelson, R. P., Aronson, E., McGuire, W. J., Newcomb, T. M., Rosenberg, M. J., & Tannenbaum, P. H. (Eds.). (1968). *Theories of cognitive consistency: A sourcebook*. Chicago: Rand McNally.

Dawson, C. K., O'Reilly, R. C., & McClelland, J. L. (2003). *The PDP++ Software User's Manual, version 3.0*. Carnegie-Mellon University: Pittsburgh, PA.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58*, 203–210.

Freedman, J. L. (1965). Long-term behavioral effects of cognitive dissonance. *Journal of Experimental Social Psychology, 1*, 145–155.

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.

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.

O'Reilly, R.C. (1996). Biologically Plausible Error-driven Learning using Local Activation Differences: The Generalized Recirculation Algorithm. *Neural Computation, 8*, 895-938.

O'Reilly, R. C. & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. A Bradford Book: The MIT Press; Cambridge, MA.

Read, S. J., & Miller, L. (1994). Dissonance and balance in belief systems: The promise of parallel constraint satisfaction processes and connectionist modeling approaches. In R. C. Schank & E. J. Langer (Eds.), *Belief, reasoning, and decision-making: Psycho-logic in honor of Bob Abelson* (pp. 209–235). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review, 103*, 219–240.

Simon, D., & Holyoak, K, J. (2002). Structural Dynamics of Cognition: From Consistency Theories to Constraint Satisfaction. *Personality and Social Psychology Review,* 6, 283–294.

Spellman, B. A., Ullman, J. B., & Holyoak, K. J. (1993). A coherence model of cognitive consistency: Dynamics of attitude change during the Persian Gulf War. *Journal of Social Issues, 49*, 147–165.

Van Overwalle, F., & Jordens, K. (2002). An Adaptive Connectionist Model of Cognitive Dissonance. *Personality and Social Psychology Review, 6*, 204–231.