A Constraint Satisfaction Model of Cognitive Dissonance Phenomena

Thomas R. Shultz
Department of Psychology
McGill University
1205 Penfield Avenue
Montréal, Québec, Canada H3A 1B1
shultz@psych.mcgill.ca

Mark R. Lepper
Department of Psychology
Stanford University
Jordan Hall, Building 420
Stanford, CA 94305-2130
lepper@psych.stanford.edu

Abstract

A constraint satisfaction network model simulated cognitive dissonance data from the insufficient justification and free choice paradigms. The networks captured the psychological regularities in both paradigms. In the case of free choice, the model fit the human data better than did cognitive dissonance theory.

Cognitive Dissonance

Cognitive dissonance theory (Festinger, 1957) has been a pillar of social psychology for some 30 years. The theory holds that dissonance is a psychological state of tension which people are motivated to reduce. Two cognitions are dissonant when, considered by themselves, one of them follows from the obverse of the other. The amount of dissonance is a function of the ratio of dissonant to consonant relations, with each relation weighted by its importance. Dissonance can be reduced by decreasing the number and/or the importance of the dissonant relations, or by increasing the number and/or the importance of consonant relations. How dissonance gets reduced depends on the resistance to change of the relevant cognitions, with less resistant cognitions being more likely to change. Resistance derives from the extent to which change would produce new dissonance, the degree to which the cognition is anchored in reality, and the difficulty of changing those aspects of reality.

Festinger (1957) used dissonance theory to account for a number of existing psychological phenomena, including the evaluation of choices, attitude change following attitude-relevant actions, and responses to the disconfirmation of beliefs. It has since been successfully applied in a wide variety of both predictive and postdictive contexts.

Consonance Model

In this paper, we present a computational model of cognitive dissonance. The model is based on the idea that dissonance reduction is a constraint satisfaction problem. Such problems are solved by the simultaneous satisfaction of many soft constraints which can vary in their relative importance. In this framework, beliefs are represented as units in a network and implications among the beliefs are represented as connections among the units. The units can be variously active and the connections (weights) can vary in strength. Hopfield (1982, 1984) has worked out the mathematics for solving such constraint satisfaction problems in parallel networks.

Hopfield networks are capable of simulating a variety of psychological phenomena, including belief revision, explanation, schema completion, analogical reasoning, and content-addressable memories (Holyoak & Thagard, 1989; Rumelhart, Smolensky, McClelland, & Hinton, 1986; Thagard, 1989). Unless used to model memory, these networks are generally considered ephemeral in the sense that they are created on line to deal with some particular task, although the creative process is not usually modeled. Hopfield networks function by reducing energy (equivalently, maximizing goodness) subject to the constraints supplied by the connections and any external input. Our Consonance Model for reducing cognitive dissonance is a Hopfield network lacking some of the parameters of other Hopfield networks and introducing some special parameters of its own.

Maximizing the consonance (goodness) of any pair of connected units depends on the sign of the connection between them. Assume an activation range of 0 to 1. If connected by a positive weight, both units should be active in order to maximize consonance. With a negative weight, consonance is maximized when both units are not active, that is, when both are inactive or only one is active. Activations change over time cycles so as to satisfy weight constraints and maximize consonance.

More formally, the consonance contributed by a particular unit i is

$$\text{consonance}_i = \sum_j w_{ij} a_i a_j \qquad (1)$$

where w_{ij} is the weight between units i and j, a_i is the activation of unit i, and a_j is the activation of unit j.

The overall consonance in the network is the sum of the values given by (1) over all units in the network

$$\text{consonance}_O = \sum_i \sum_j w_{ij}\, a_i\, a_j \qquad (2)$$

Activation spreads over time cycles by two simple update rules:

$$a_i(t+1) = a_i(t) + \text{net}_i\,(\text{ceiling} - a_i(t)) \quad \text{when } \text{net}_i >= 0 \qquad (3)$$

$$a_i(t+1) = a_i(t) + \text{net}_i\,(a_i(t) - \text{floor}) \quad \text{when } \text{net}_i < 0 \qquad (4)$$

where $a_i(t+1)$ is the activation of unit i at time t + 1, $a_i(t)$ is the activation of unit i at time t, ceiling is the maximal level of activation, floor is the minimal activation, and net_i is the net input to unit i, defined as

$$\text{net}_i = \text{resist}_i \left(\sum_j w_{ij}\, a_j\right) \qquad (5)$$

The parameter resist_i is a measure of the resistance of unit i to having its activation changed. The larger the value of the resistance multiplier, the less the resistance to change. The default values for floor and ceiling are 0 and 1, respectively.

At each time cycle, n units are randomly selected and updated according to rules (3) and (4). By default, n is the number of units in the network.

A few additional parameters concerning the construction of the networks are described later in the context of particular simulations.

Simulations

With more than 1000 published entries in the cognitive dissonance literature, there is considerable choice in deciding what to simulate. Here we present two of our current simulations, one representing each of two of the major paradigms in dissonance theory: insufficient justification and free choice.

Insufficient Justification

The insufficient justification paradigm deals with situations in which subjects engage in some counter-attitudinal action with rather little justification. Dissonance theory predicts that the less the justification for the behavior, the greater the dissonance and, at least when it is difficult to retract one's action, the more people will be motivated to change their attitudes so as to provide additional justification for their action.

Several different types of experiments have been developed to test these insufficient justification predictions (e.g., Aronson & Carlsmith, 1963; Aronson & Mills, 1959; Festinger & Carlsmith, 1959). In the present paper we simulate one of the best studied and most robust of these.

In one of the seminal studies within this paradigm, nursery school children were forbidden to play with a desirable toy under either mild or severe threat (Aronson & Carlsmith, 1963). Both of these threats were sufficient to prevent the children from playing with the desirable toy during a play period in which the experimenter was absent from the room. In subsequent ratings, the children derogated the forbidden toy more under mild threat than severe threat. The theoretical explanation is that the children committed themselves to the dissonant behavior of not playing with the desirable toy. Since dissonance increases with the fewer cognitions that support the behavior, there was more dissonance in the mild threat condition than in the severe threat condition. Because the counter-attitudinal behavior could not be retracted, dissonance was reduced by derogating the forbidden toy. The greater the dissonance, the greater the derogation.

Alternative explanations of these findings included the notion that severe threat focused more attention on the toy or made it seem more desirable and the idea that the experimenter was more likeable or more credible in the mild threat condition. To rule out such alternatives, Freedman (1965) added surveillance conditions to the experiment in which the experimenter stayed in the room while the child played. In the surveillance conditions, the same threats were used but temptation, and thus dissonance, was lowered by the experimenter's continued presence. Actual play with the previously forbidden toy five weeks later indicated greater derogation in the mild than in the severe conditions only when there was no surveillance, thus supporting the dissonance explanation against the alternatives.

Our simulation focused on the Freedman (1965) experiment. The constraint satisfaction network for the non-surveillance conditions of this simulation is presented in Figure 1. Because unit activations have a floor of 0, two units are used to encode each dimension of interest: *toy evaluation*, *threat*, and *play with toy*. In each pair of units, the unit coded + represents the positive end of the dimension and the unit coded - represents the negative end of the dimension. The units in each pair are connected by a negative weight so that only one of them is active at a time. In these network diagrams, negative weights are symbolized by dashed lines and positive weights by solid lines. Each pair of units is surrounded by an

ellipse to convey idea that they refer to opposite ends of the same dimension.

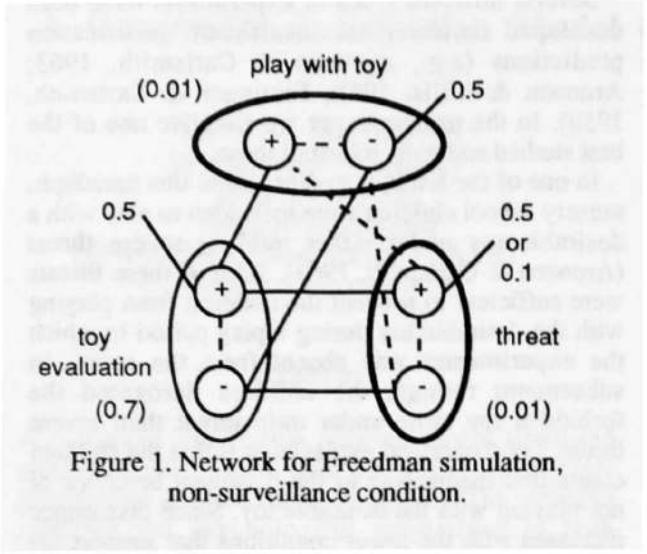

Figure 1. Network for Freedman simulation, non-surveillance condition.

Connections across different dimensions (ellipses) reflect assumed psychological implications among the beliefs. For simplification, we connect positive units only to other positive units and negative units only to other negative units across dimensions. For the Freedman simulation, there were positive connections between toy evaluation and play (the better liked the toy, the more it would be played with), positive connections between toy evaluation and threat (the better liked the toy, the more threat would be required to prevent play), and negative connections between play and threat (the bigger the threat, the less the toy would be played with).

Resistance of units to activation change is portrayed by the thickness of the ellipse. Resistance values for a particular boundary thickness are presented in parentheses in Figure 1: 0.70 for toy evaluation (low resistance) vs. 0.01 for the other two beliefs (high resistance). These resistance values are based on the assumption that, whereas play and threat are relatively fixed, evaluation of the toy should be allowed to vary. In a more complete model, resistance might be implemented by constraining connections to many other beliefs. For simplification, this can be accomplished with an explicit resistance parameter.

Initial activations provided to units are indicated in Figure 1 by pointers coming from outside the units. The toy is given a moderately positive evaluation (0.5) to reflect its desirability, play is given a moderately negative (-0.5) evaluation because it was not done, and the amount of threat is either 0.5 or 0.1 to represent the two severity conditions.

A cap parameter, when set to a high negative proportion, prevents activations from growing to the ceiling of 1.0. Our default setting for cap is -0.8. Mathematically, cap is the value of the connection between each unit and itself, w_{ii}.[1] Hopfield (1982, 1984) had assumed that such self-connections are 0. Allowing self-connections to be other than 0 produces additional spurious states in the neighborhood of a desired attractor, thus increasing the variability of solutions (Hertz, Krogh, & Palmer, 1991). We use cap to enforce the psychologically realistic assumption that the events in most dissonance experiments are not of major importance to the subjects. Therefore, activations should not reach maximal values.

The wrange parameter represents the range of positive weights below 1 and negative weights above -1. We employ a default value of 0.2 for wrange. The weights are not identical across networks, but rather are mainly positive or mainly negative within this specified range. Again, the purpose is to introduce some degree of psychological realism. Such variation is not necessary to qualitatively capture the predicted dissonance phenomena. This randomization of weights violates the symmetry assumed by Hopfield (1982, 1984), in that $w_{ij} <> w_{ji}$. He reported that violations of the symmetry assumption increased memory errors and instability in network solutions. Such results may correspond to psychological variation.

Use of the cap and wrange parameters effectively nullifies the mathematical guarantee that these nets will maximize consonance. It is our view that psychological plausibility should outweigh guaranteed maxima in the context of simulating human data.

The rand% parameter was defined by default as wrange/2. It represents a random percentage added to or subtracted from the initial values of activations, resistances, and caps. This too was for psychological realism; presumably not everyone has precisely the same parameter values.

For the surveillance condition, there was no connection between toy evaluation and play, represented by weights of 0. No matter how much you like toy, you won't be tempted to play with it as long as the experimenter is present. The impact of both threats was scaled up by a multiplier in the spirit of update rule (3): new_threat = old_threat + (0.5 * (1 - old_threat)). This made the value of threat 0.75 in the severe/surveillance condition and 0.55 in the mild/surveillance condition. This reflects the idea that surveillance enhances the value of both threats, but in accordance with the way that activations change.

As a simulation begins, activations of units are updated in a random, asynchronous fashion. On each time cycle, n units are randomly selected and updated using rules (3) and (4). By default, n is the number of units in the network, 6 in this simulation. Updating

[1]Thanks to Denis Mareschal for this suggestion.

continued for 20 cycles because asymptotes were reached well within that period. We ran 10 networks in each condition.

Mean evaluation of the toy after cycle 20 is shown in Figure 2. This was computed as the difference between activation of the positive unit and the negative unit. As in Freedman (1965), there was an interaction between surveillance and severity of threat, $F(1, 36) = 169.02$, $p < .001$. There was more derogation in the mild than in the severe condition, but this effect was much larger without surveillance.

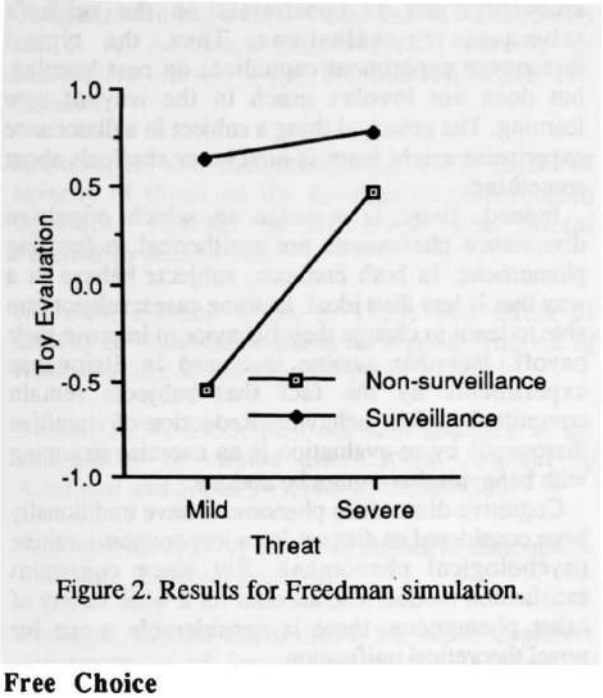

Figure 2. Results for Freedman simulation.

Free Choice

Another major paradigm in cognitive dissonance concerns free choice. Choosing between alternatives creates cognitive dissonance due to the fact that the chosen alternative is never perfect and the rejected alternative often has desirable aspects which are foregone when a final choice is made. Dissonance can be reduced either by making the chosen object more desirable or by making the rejected object less desirable. Thus, dissonance reduction further separates the alternative choices in desirability. The magnitude of dissonance is greater the closer the alternatives are in desirability, and hence the more difficult the choice between them is, before the choice is made.

The classic free choice experiment asked female university students to rate eight small appliances (Brehm, 1956). They were then given a difficult choice, between two objects that they had rated high, or an easy choice, between one object they had rated high and one they had rated low. Then the objects were rated again. Degree of separation was measured by subtracting the second rating from the first rating for each object. Although the dissonance theory prediction was for greater separation in the difficult choice condition than in the easy choice condition, most of the actual separation obtained was due to a relatively large decrease in the value of the rejected alternative in the difficult choice condition.

The network for simulating the Brehm experiment is portrayed in Figure 3. There were pairs of units to represent each of the three critical dimensions: chosen alternative, rejected alternative, and decision. There were positive weights between chosen and decision, and negative weights between rejected and decision. The initial activations were 0.5 for chosen, 0.4 for rejected difficult, 0.1 for rejected easy, and 0.7 for decision. There was high resistance for the decision and low resistance for evaluation of the two alternatives, with the default values of 0.01 and 0.7, respectively. Other parameter settings were the same as in the Freedman simulation.

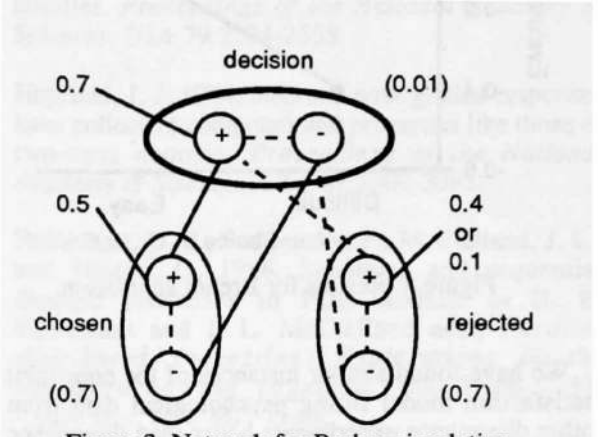

Figure 3. Network for Brehm simulation.

The mean difference scores (re-evaluation - initial evaluation) are plotted in Figure 4. Each evaluation was computed as the difference in activation between the positive and negative units. Evaluation of the chosen object increased and evaluation of the rejected object decreased in both conditions. The amount of change was greater in the difficult condition, as predicted by dissonance theory, $F(1, 18) = 57.70$, $p < .001$. Notice that most of the change in the difficult condition is due to a decrease in evaluation of the rejected alternative. This outcome fits Brehm's (1956) results more precisely than does dissonance theory, which predicts only a larger separation of the alternatives following a difficult choice than following an easy choice.

Discussion

The simulation results matched the psychological findings and, in the case of free choice, provided even

better coverage of the psychological data than did dissonance theory. In the free choice simulation, the locus of most of the action was in the re-evaluation of the rejected alternative in the difficult condition. This was indeed what Brehm (1956) found, although he did not comment on the discrepancy from strict dissonance theory predictions.

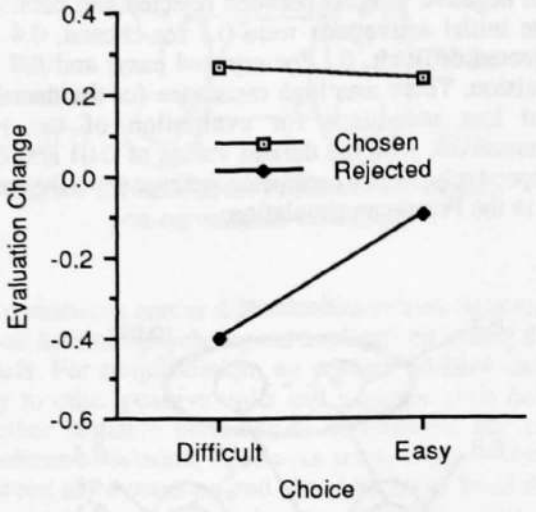

Figure 4. Results for Brehm simulation.

We have found similar instances of the constraint satisfaction model fitting psychological data from other dissonance experiments better than dissonance theory does. These superior fits derive from the capacity of constraint satisfaction models to deal with variables other than those unique to dissonance theory and the increased precision that is inherent to computational formulations.

The present simulations were conducted with a minimum of parameter adjustment. Network weights were positive, negative, or zero; resistance was high or low; and initial levels of activation were either high or low. Additional experimentation revealed that these effects were robust against parameter variation, and that the default parameter settings were applicable to a variety of other dissonance simulations.

The present simulations began with some units having initial, non-zero activation values. More conventionally, constraint satisfaction programs start all units at zero activation and provide some units with external inputs. Activations then gradually build up from zero as a function of both external input and internal network input. This conventional scheme did not seem appropriate for cognitive dissonance phenomena because it yielded results indicating a gradual increase in consonance, but no dissonance. To ensure that the networks modeled dissonance, we initialized some unit activations in conformity with procedures in the psychological experiments.

Although connection weight values can be learned for constraint satisfaction models (e.g., Anderson & Mozer, 1981), there was no such learning in the present simulations. This reflects the fact that the typical dissonance experiment is not an occasion for learning. Instead, acculturated, experienced subjects enter a situation in which they commit themselves to some behavior under the influence of a few salient, experimentally engineered cognitions. These cognitions, the behavioral commitment, and existing knowledge act as constraints on the subject's subsequent re-evaluations. Thus, the typical dissonance experiment capitalizes on past learning, but does not involve much in the way of new learning. The principal thing a subject in a dissonance experiment might learn is how he or she feels about something.

Indeed, there is a sense in which cognitive dissonance phenomena are antithetical to learning phenomena. In both contexts, subjects behave in a way that is less than ideal. In some cases, subjects are able to learn to change their behavior to improve their payoff. But that avenue is closed in dissonance experiments by the fact that subjects remain committed to their behavior. Reduction of cognitive dissonance by re-evaluation is an exercise in coping with behavior that cannot be undone.

Cognitive dissonance phenomena have traditionally been considered as distinct from less counter-intuitive psychological phenomena. But since constraint satisfaction models also account for a wide variety of other phenomena, there is considerable scope for novel theoretical unification.

Cognitive dissonance theory is but one of a number of theories in social psychology emphasizing that people try to achieve consistency among cognitions (Abelson, Aronson, McGuire, Newcombe, Rosenberg, & Tannenbaum, 1968; Abelson & Rosenberg, 1958; Heider, 1958). Although these consistency theories have enjoyed considerable success as verbal formulations, the underlying reasoning mechanisms for establishing consistency have not been precisely specified. It may be that connectionist constraint satisfaction models could serve as a general modeling technique and explanatory device in these areas (cf. Holyoak & Spellman, 1991).

Acknowledgements

This research was supported by a grant to the first author from the Social Sciences and Humanities Research Council of Canada and a grant to the second author from the U. S. National Institute of Mental Health.

References

Abelson, R. P., Aronson, E., McGuire, W. J., Newcombe, T. M., Rosenberg, M. J., and Tannenbaum, P. H. eds. 1968. *Theories of cognitive consistency: A sourcebook.* Chicago: Rand McNally.

Abelson, R. P., and Rosenberg, M. J. 1958. Symbolic psycho-logic: A model of attitudinal cognition. *Behavioral Science* 3:1-13.

Anderson, J. A., and Mozer, M. C. 1981. Categorization and selective neurons. In G. E. Hinton and J. A. Anderson eds., *Parallel models of associative memory*, pp. 213-236. Hillsdale, NJ: Erlbaum.

Aronson, E., and Carlsmith, J. M. 1963. Effect of severity of threat on the devaluation of forbidden behavior. *Journal of Abnormal and Social Psychology* 66:584-588.

Aronson, E., and Mills, J. 1959. The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology* 59:177-181.

Brehm, J. W. 1956. Post-decision changes in the desirability of choice alternatives. *Journal of Abnormal and Social Psychology* 52:384-389.

Festinger, L. 1957. *A theory of cognitive dissonance.* Evanston, IL: Row, Peterson.

Festinger, L., and 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.

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

Hertz, J., Krogh, A., and Palmer, R. G. 1991. *Introduction to the theory of neural computation.* Reading, MA: Addison-Wesley.

Holyoak, K. J., and Spellman, B. A. 1991. If Saddam is Hitler then who is George Bush? Analogical mapping between systems of social roles. Forthcoming.

Holyoak, K. J., and Thagard, P. 1989. Analogical mapping by constraint satisfaction. *Cognitive Science* 13:295-355.

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

Hopfield, J. J. 1984. Neurons with graded responses have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences, USA* 81:3008-3092.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., and Hinton, G. 1986. Schemata and sequential thought processes in PDP models. In D. E. Rumelhart and J. L. McClelland eds., *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 2, pp. 7-57. Cambridge, MA: MIT Press.

Thagard, P. 1989. Explanatory coherence. *Behavioral and Brain Sciences* 12:435-502.