Clinging to Beliefs: A Constraint-satisfaction Model

Thomas R. Shultz (shultz@psych.mcgill.ca)
Department of Psychology; McGill University
Montreal, QC H3C 1B1 Canada

Jacques A. Katz (jakatz@cnbc.cmu.edu)
Department of Psychology; Carnegie Mellon University
Pittsburgh, PA 15213 USA

Mark R. Lepper (lepper@psych.stanford.edu)
Department of Psychology; Stanford University
Stanford, CA 94305-2130 USA

Abstract

Beliefs tend to persevere even after evidence for their initial formulation has been invalidated by new evidence. If people are assumed to rationally base their beliefs on evidence, then this belief perseverance is somewhat counterintuitive. We constructed a constraint-satisfaction neural network model to simulate key belief perseverance phenomena and to test the hypothesis that explanation plays a central role in preserving evidentially challenged beliefs. The model provides a good fit to important psychological data and supports the hypothesis that explanations preserve beliefs.

Introduction

It is perhaps surprising that people are so often reluctant to abandon personal beliefs that are directly contradicted by new evidence. This tendency to cling to beliefs in the face of subsequent counterevidence has been well demonstrated for opinions (Abelson, 1959), decisions (Janis, 1968), impressions of people (Jones & Goethals, 1971), social stereotypes (Katz, 1960), scientific hypotheses (T. S. Kuhn, 1962), and commonsense ideas (Gilovich, 1991).

Belief perseverance is puzzling because it is commonly assumed that beliefs are based on evidence. If it is rational for people to form a belief based on evidence, then why is it not equally rational for them to modify the belief when confronted with evidence that invalidates the original evidence?

Debriefing Experiments

Some of the clearest cases of apparently irrational belief perseverance come from debriefing experiments. In these experiments, subjects learn that the initial evidential basis for a belief is invalid. For example, Ross, Lepper, and Hubbard (1975) first provided subjects with false feedback concerning their ability to perform a novel task. Their subject's task was to distinguish authentic from fake suicide notes by reading a number of examples. False feedback from the experimenter led subjects to believe that they had performed at a level that was much better than average or much worse than average. Then, in a second phase, subjects were debriefed about the random and predetermined nature of the feedback that they had received in the first phase. There were three debriefing conditions. In the outcome debriefing condition, subjects were told that the evidence on which their initial beliefs were based had been completely fabricated by the experimenter. Subjects in the process debriefing condition were additionally told about the procedures of outcome debriefing along with explanations about possible mechanisms and results of belief perseverance. Subjects in this condition were also told that belief perseverance was the focus of the experiment. Finally, subjects in a no-debriefing control condition were not debriefed at all after the feedback phase. Subsequently, subjects in all three conditions rated their own ability at the suicide-note verification task. This was to assess the perseverance of their beliefs about their abilities on this task that were formed in the feedback phase.

The mean reported beliefs for the three debriefing conditions are shown in Figure 1. There is an interaction between debriefing condition and the nature of feedback (success or failure at the note-verification task). The largest difference between success and failure feedback occurs in the no-debriefing condition. In this control condition, subjects who were initially led to believe that they had succeeded continue to believe that they would do better than subjects initially led to believe that they had failed. After outcome debriefing, there is still a significant difference between the success and failure conditions, but at about one-half of the strength of the control condition. The difference between success and failure feedback effectively disappears after process debriefing. This sort of belief perseverance after debriefing has been convincingly demonstrated for a variety of different beliefs and debriefing techniques (Jennings, Lepper, & Ross, 1981; Lepper, Ross & Lau, 1986).

One explanation for such belief perseverance is that people frequently explain events, including their own beliefs, and such explanations later sustain these beliefs in the face of subsequent evidential challenges (Ross et al., 1975). For example, a person who concludes from initial feedback that she is very poor at authenticating suicide notes might attribute this inability to something about her experience or personality. Perhaps she has had too little contact with severely depressed people, or maybe she is too optimistic to empathize deeply with a suicidal person. Then in the second

phase, when told that the feedback was entirely bogus, these previously constructed explanations may still suggest that she lacks the ability to authenticate suicide notes. Analogously, a subject who is initially told that he did extremely well at this task may explain his success by noting his familiarity with some depressed friends or his sensitivity to other people's emotions. Once in place, such explanations could inoculate the subject against subsequent evidence that the initial feedback was entirely bogus.

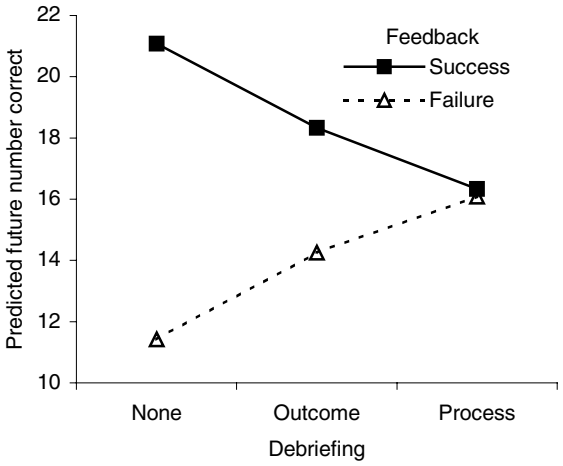

Figure 1: Mean predicted ability in the Ross et al. (1975) experiment after debriefing.

The assumption is that even though contradictory evidence may weaken a belief, it is unlikely to alter every cognition that may have derived from that belief, such as explanations for the belief's existence. The well-known frame problem emphasizes the computational intractability of tracking down every implication of an altered belief (Charniak & McDermott, 1985). People generally do not have the time, energy, knowledge, or inclination to decide which other beliefs to change whenever a belief is changed.

In contrast to the view that people have difficulty distinguishing explanations from evidence (D. Kuhn, 1991), recent research suggests that people can distinguish explanations from evidence and that they tend to use explanations as a substitute for missing evidence (Brem & Rips, 2000).

In this paper, we report on our attempt to simulate the belief perseverance phenomena reported by Ross et al. (1975). Our basic theoretical premise in designing these simulations is that belief perseverance is a special case of a more general tendency for people to seek cognitive consistency. Striving for consistency has long been considered to cause a wide variety of phenomena in social psychology (Abelson, Aronson, McGuire, Newcomb, Rosenberg, & Tannenbaum, 1968). In the case of belief perseverance, we assume that people form percepts that are consistent with external evidence, then acquire beliefs that are consistent with these percepts, and finally construct explanations that are consistent with these beliefs. We view resistance to new evidence that contradicts existing percepts, beliefs, or explanations as part of an attempt to achieve overall consistency among current cognitions, given that not all implications of contradictory evidence are actively pursued.

There was a simulation using non-monotonic logic of how belief can be preserved despite ordinary debriefing, but it did not cover the quantitative differences between conditions in the Ross et al. experiment (Hoenkamp, 1987).

Neural Constraint Satisfaction

Our simulations use a technique called constraint satisfaction, which attempts to satisfy as many constraints as well as possible within artificial neural networks. The present model is closely related to models used in the simulation of schema completion (Rumelhart, Smolensky, McClelland, & Hinton, 1986), person perception (Kunda & Thagard, 1996), attitude change (Spellman, Ullman, & Holyoak, 1993), and dissonance reduction (Shultz & Lepper, 1996).

Constraint satisfaction neural networks are comprised of units connected by weighted links. Units can represent cognitions by taking on activation values from 0 to 1, representing the strength or truth of the cognition. Connection weights can represent relations between cognitions and are assigned positive or negative values representing the sign and strength of the relations. Connection weights are bidirectional to permit cognitions to mutually influence each other. External inputs to units represent influences from the environment. Biases are represented by internal inputs to a given unit that do not vary across different network inputs.

Networks attempt to satisfy the soft constraints imposed by fixed inputs, biases, and weights by changing activation values of the units. Unit activations are updated according to these rules:

$$a_i(t+1) = a_i(t) + net_i\left(ceiling - a_i(t)\right), \text{ when } net_i \geq 0 \quad (1)$$

$$a_i(t+1) = a_i(t) + net_i\left(a_i(t) - floor\right), \text{ when } net_i < 0 \quad (2)$$

where $a_i(t + 1)$ is the updated activation value of unit i, a_i is the current activation of unit i, ceiling is the maximum activation value for a unit, floor is the minimum activation value for a unit, and net_i is the net input to unit i, as computed by:

$$net_i = in\left(\sum_j w_{ij} a_j + bias_i\right) + ex(input_i) \quad (3)$$

where in and ex are parameters that modulate the impact of the internal and external inputs, respectively, with default values of 0.1, w_{ij} is the connection weight between units i and j, a_j is the activation of sending unit j, $bias_i$ is the bias value of unit i, and $input_i$ is the external input to unit i.

These update rules ensure that network consistency either increases or stays the same, where consistency is computed as:

$$consistency = \sum_{ij} w_{ij} a_i a_j + \sum_i input_i a_i + \sum_i bias_i a_i \quad (4)$$

When a network reaches a high level of consistency, this means that it has settled into a stable pattern of activation and that the various constraints are well satisfied. In such stable solutions, any two units connected by positive weights tend to both be active, units connected by negative weights tend not to be simultaneously active, units with high inputs tend to be more active than units with low inputs, and units with high biases tend to be more active than units with low biases.

Increases in consistency and constraint satisfaction occur gradually over time. At each time cycle, *n* units are randomly selected for updating, where *n* is typically the number of units in the network. Thus, not every unit is necessarily updated on every cycle and some units may be updated more than once on a given cycle.

Unusual Simulation Features

The foregoing characteristics of neural constraint satisfaction are quite common. In addition, the present modeling has a few somewhat unusual features. Perhaps the most important of these is a two-phase structure that accommodates the two main phases of belief perseverance experiments. It is more typical for neural constraint satisfaction models to operate in a single phase in which networks are designed and updated until they settle into a stable state. Our two phases correspond to the feedback and debriefing phases of these experiments. After a network settles in the initial feedback phase, new units can be introduced, and inputs, connection weights, and biases may be changed in a second, debriefing phase. To implement continuity between the two phases, a simple type of memory was introduced such that activation values from the feedback phase would be partially retained as unit biases in the debriefing phase. Final activations in the feedback phase were multiplied by 0.05 to transform them into biases for the debriefing phase. This is not a detailed implementation of a memory model, but is rather a convenient shorthand implementation of the idea that there is a faded memory for whatever conclusions were reached in the previous, feedback phase.

Two other unusual features derived from our earlier simulations of cognitive dissonance reduction (Shultz & Lepper, 1996): a cap parameter and randomization of network parameters. The cap parameter is a negative self-connection weight for every unit that limits unit activations to less than extreme values. The purpose of this activation cap is to increase psychological realism for experiments about beliefs that reach no more than moderate strength.

Robustness of simulation results was assessed by simultaneously randomizing all network parameters (i.e., biases, inputs, and connection weights) by up to 0%, 10%, 50%, or 100% of their initial values according the formula:

$$y = x \pm \{rand \ (abs \lfloor x * rand\% \rfloor) \} \quad (5)$$

The initial parameter value x is multiplied by the proportion of randomization being used (0, .1, .5, or 1) and converted to an absolute value. Then a random number is selected between 0 and the absolute value under a uniform distribution. This random number is then randomly either added to or subtracted from the initial value. This parameter randomization allows efficient assessment of the robustness of the simulation under systematic variations of parameter values. If the simulations succeed in matching the psychological data, even under high levels of parameter randomization, then they do not depend on precise parameter settings. This randomization process also enhances psychological realism because not every subject can be expected to have precisely the same parameter values.

Network Design

Units

Units represent external input and the three types of cognitions that are critical to belief perseverance experiments, i.e., percepts, beliefs, and explanations. Percept units represent a subject's perception of external input, in this case feedback provided by the experimenter. Belief units represent a subject's beliefs, and explanation units represent a subject's explanations of particular beliefs. In each case, the larger the activation value of a given unit, the stronger the associated cognition. Activation values range from 0 to 1, with 0 representing no cognition, and 1 representing the strongest cognition. All unit activations start at 0 as a network begins to run.

Unit names include a sign of +, -, or 0 to represent the direction of a given cognition. For example, in these simulations, *+percept* refers to a perception of doing well on a task, *-percept* to a perception of doing poorly on the task, and *0percept* to not knowing about performance on the task. Percept units sometimes have an external input, to reflect the feedback on which the percept is based. A *0percept* unit is required for simulating debriefing experiments, where information is encountered that explicitly conveys a lack of knowledge about performance. Analogously, *+belief* represents a belief that one is performing well at a task, *–belief* represents a belief that one is performing poorly at a task, *+explanation* represents an explanation for a *+belief,* and *-explanation* represents an explanation for a *–belief*.

Connections

Units are joined by connection weights that have a size and a sign. The sign of a weight represents a positive or negative relation between connected units. A positive weight signals that a cognition follows from, leads to, is in accordance with, or derives support from another cognition. A negative weight indicates that a cognition is inconsistent with or interferes with another cognition. Decisions about signs are based on descriptions of psychological procedures. Initial nonzero connection weights are + or - 0.5 in our simulations. Connection weights of 0 indicate the absence of relations between cognitions. All connection weights are bi-directional to allow mutual influences between cognitions.

The general connection scheme in our simulations of belief perseverance has external inputs feeding percepts, which are in turn connected to beliefs, which are in turn connected to explanations. For failure conditions, a -percept unit receives external input and is connected to a -belief unit, which is in turn connected to a -explanation unit. For success conditions, a +percept unit receives external input and is connected to a +belief unit, which is in turn connected to a +explanation unit. Connection weights between incompatible cognitions, such as between +belief and -belief or between -percept and 0percept, are negative.

The principal dependent measure in many belief perseverance studies is a subject's self-rated ability on a task. This is represented as net belief, computed as activation on the +belief unit minus activation on the -belief unit, after the network settles in the debriefing phase. This technique of using two negatively connected units to represent the different poles of a single cognition was used by Shultz and Lepper (1996) in their simulation of cognitive dissonance phenomena.

Networks for Feedback Phase

Figure 2 shows specifications for the negative feedback condition. Negative feedback, in the form of external input, with a value of 1.0, is positively connected to the –percept unit. This same network design is used for the no-debriefing condition of the debriefing phase.

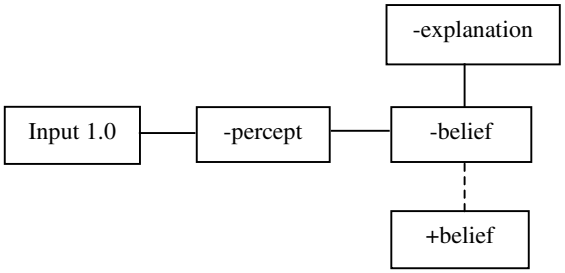

Figure 2: Network for negative feedback. Positive connection weights are indicated by solid lines; negative connection weights by dashed lines.

A feedback phase represents the presentation of information on how a subject is doing on a task. It is assumed that this information forms the basis for a belief about ability and to explanations of that ability. Because of the connection scheme and the fact that all unit activations start at 0, percept units reach activation asymptotes first, followed by belief units, and finally by explanation units.

Networks for Debriefing Phase

Figure 3 shows network specifications for the debriefing phase. This network was used for both outcome debriefing and process debriefing. The particular network shown in Figure 3 shows a debriefing phase that follows negative feedback. As noted earlier, an unusual feature here is the inclusion of biases for percept, belief, and explanation units from the earlier, feedback phase. These biased units are represented by bolded rectangles around unit names, and implement a faded memory of the feedback phase. There is also a new unit, the 0percept unit, with an input of 1.0, to represent that nothing valid is known about task performance. This unit has no bias because it was not present in the previous phase. It is negatively connected to the – or + percept unit to represent the idea that the feedback data from the previous phase are false, and thus convey no information about task ability.

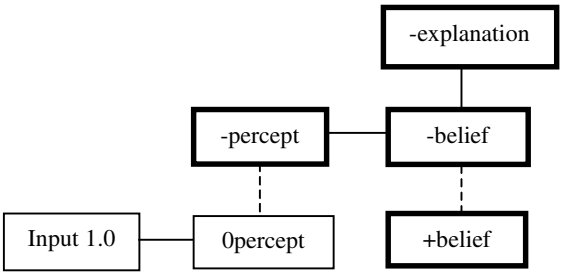

Figure 3: Network for outcome and process debriefing following negative feedback. Units that have biases from the feedback phase are indicated by bolded rectangles.

We implemented the stronger, process debriefing by multiplying bias values by a factor of 0.1. This reflects the idea that process debriefing is so thorough that it severely degrades all cognitions that were created in the preceding feedback phase. Networks in the no-debriefing condition were identical to those described in Figure 2, with no topology changes after the feedback phase. However, as in all debriefing conditions, biases of .05 of final activations were used for any units being carried over from the feedback phase. Networks were run for 120 update cycles in each of the two phases; by this time they had typically settled into stable states.

Principles of Network Design

In summary, network design can be summarized by 13 principles:

1. Units represent cognitions.
2. The principal cognitions in belief perseverance experiments are input feedback, percepts, beliefs, and explanations.
3. The sign of unit names represent the positive or negative poles of cognitions.
4. Unit activation represents strength of a cognition (or a pole of a cognition).
5. The difference between positive and negative poles of a cognition represents the net strength of the cognition.
6. Connection weights represent constant implications between cognitions.
7. Connection weights are bi-directional, allowing possible mutual influence between cognitions or poles of cognitions.

8. Cognitions whose poles are mutually exclusive have negative connections between the positive and negative poles.
9. Size of external input represents strength of environmental influence, such as evidence or feedback.
10. External inputs are connected to percepts, percepts to beliefs, and beliefs to explanations, representing the assumed chain of causation in belief perseverance experiments. That is, environmental feedback creates percepts, which in turn create beliefs, which eventually lead to explanations for the beliefs.
11. Networks settling into stable states represent a person's tendency to achieve consistency among cognitions.
12. Final unit activations from the feedback phase are converted to unit biases for the start of the debriefing phase of belief perseverance experiments, representing the participant's memory of the feedback phase.
13. Multiplying activation bias values by 0.1 represents thorough, process debriefing.

Results

We focus on the final net belief about one's ability after the debriefing phase. This is computed as activation on the +belief unit minus activation on the -belief unit. Here, we report only on the 10% randomization level, but similar results are found at each level of parameter randomization.

Net belief scores were subjected to a factorial ANOVA in which debriefing condition (none, outcome, and process) and feedback condition (success, failure) served as factors. There was a main effect of feedback, $F(1, 114) = 29619$, $p < .001$, and an interaction between debriefing and feedback, $F(2, 114) = 9102$, $p < .001$. Mean net ability scores are shown in Figure 4. For success feedback, net belief scores were higher after no debriefing than scores obtained after outcome debriefing, which were in turn higher than scores obtained after process debriefing. The opposite holds for failure feedback.

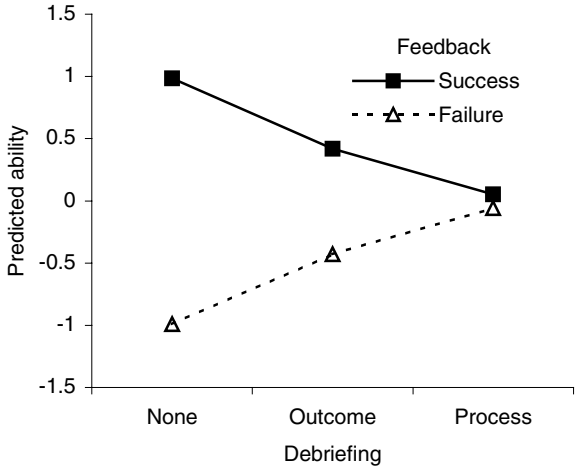

Figure 4: Mean predicted ability in the simulation after debriefing.

To assess the fit to human data, we computed a regression F with regression weights based on the pattern of the Ross et al. (1975) results. The regression weights were 2, -2, 1, -1, 0, and 0 for the no debriefing/success, no debriefing/failure, outcome debriefing/success, outcome debriefing/failure, process debriefing/success, and process debriefing/failure conditions, respectively. This produced a highly significant regression $F(1, 114) = 47558$, $p < .001$, with a much smaller residual $F(4, 114) = 67$, $p < .001$. The regression F accounts for 99% of the total variance in net belief. As with human subjects, there is a large difference between success and failure with no debriefing, a smaller but still substantial difference after outcome debriefing, and very little difference after process debriefing.

To assess the role of explanation in the simulation, we subjected activations on the explanation unit after the debriefing phase to the same ANOVA. There is a main effect of debriefing, $F(2, 114) = 3787$, $p < .001$, a much smaller main effect for feedback $F(1, 114) = 15.37$, $p < .001$, and a small interaction between them, $F(2, 114) = 6.76$, $p < .005$. The mean explanation scores are presented in Figure 5. Explanations are strong under no-debriefing, moderately strong under outcome debriefing, and weak under process debriefing. But because explanations had been strongly active in all three conditions at the end of the feedback phase, these post-debriefing results reflect relative differences in maintenance of explanations. Explanations are maintained under no debriefing, partially maintained under outcome debriefing, and eliminated in process debriefing.

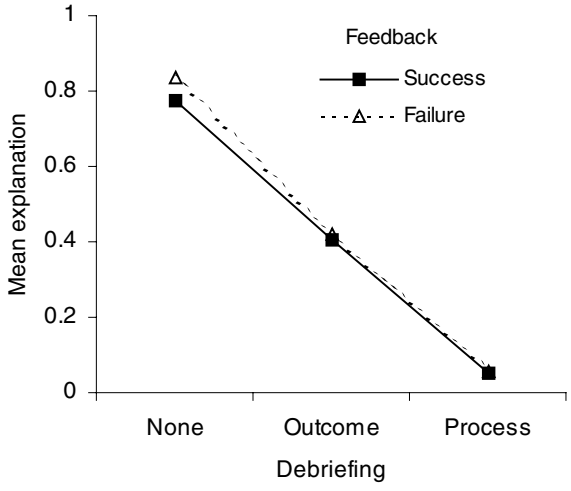

Figure 5: Mean explanation scores in the simulation after debriefing.

Discussion

The tendency for beliefs to persevere even after evidence for them has been fully invalidated challenges some basic assumptions about human rationality. If people reasonably

base their beliefs on evidence, then why is counter-evidence not sufficient to eliminate or change beliefs?

We used constraint-satisfaction neural networks to test the idea that explanation plays a key role in sustaining beliefs in these circumstances. The model provides a good fit to existing psychological data from debriefing experiments in which subjects are informed that that the principal evidence for their beliefs is no longer valid (Ross et al., 1975). Simulated beliefs remain strong without debriefing; belief strength is reduced after standard outcome debriefing, and eliminated after more thorough, process debriefing. This pattern of results matches the psychological data, with about half-strength beliefs under outcome debriefing and elimination of beliefs by process debriefing. As in our earlier simulations of cognitive dissonance phenomena, the neural constraint-satisfaction model is here shown to be robust against parameter variation. Even a high degree of parameter randomization does not change the pattern of results.

The simulations further revealed that belief perseverance is mirrored by strength of explanation. Explanations remain strong with no debriefing, and decrease progressively with more effective debriefing. Although it is obvious that debriefing reduces the strength of erroneous beliefs, the finding that it also reduces explanations is perhaps less obvious. In our simulations, explanation is reduced by effective debriefing via connections from external evidence to percepts, percepts to beliefs, and beliefs to explanations.

People spontaneously generate explanations for events as a way of understanding events, including their own beliefs (Kelley, 1967). If an explanation is generated, this explanation becomes a reason for holding an explained belief, even if the belief is eventually undercut by new evidence.

Future work in our group will extend this model to other belief perseverance phenomena and attempt to generate predictions to guide additional psychological research.

Acknowledgments

This research was supported by a grant to the first author from the Social Sciences and Humanities Research Council of Canada and by grant MH-44321 to the third author from the U.S. National Institute of Mental Health.

References

Abelson, R. P. (1959). Modes of resolution of belief dilemmas. *Conflict Resolution, 3*, 343-352.

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.

Brem, S. K., & Rips, L. J. (2000). Explanation and evidence in informal argument. *Cognitive Science, 24*, 573-604.

Charniak, E., & McDermott, D. (1985). *Introduction to artificial intelligence*. Reading, MA: Addison-Wesley.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: Free Press.

Hoenkamp, E. (1987). An analysis of psychological experiments on non-monotonic reasoning. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (Vol. 1, pp. 115-117). Los Altos, CA: Morgan Kaufmann.

Janis, I. (1968). Stages in the decision-making process. In R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, & P. H. Tannenbaum (Eds.) (1968). *Theories of cognitive consistency: A sourcebook*. Chicago: Rand McNally.

Jennings, D. L., Lepper, M. R., & Ross, L. (1981). Persistence of impressions of personal persuasiveness: Perseverance of erroneous self-assessments outside the debriefing paradigm. *Personality and Social Psychology Bulletin, 7*, 257-263.

Jones, E. E., & Goethals, G. R. (1971). Order effects in impression formation: Attribution context and the nature of the entity. In E. E. Jones et al. (Eds.), *Attribution: Perceiving the causes of behavior*. Morristown, NJ: General Learning Press.

Katz, D. (1960). The functional approach to the study of attitudes. *Public Opinion Quarterly, 24*, 163-204.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation.* Vol. 15. Lincoln: University of Nebraska Press.

Kuhn, D. (1991). *The skills of argument*. Cambridge: Cambridge University Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

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

Lepper, M. R., Ross, L., & Lau, R. R. (1986). Persistence of inaccurate beliefs about self: Perseverance effects in the classroom. *Journal of Personality and Social Psychology, 50*, 482-491.

Ross, L, Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32*, 880-892.

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

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

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.