

# **Learning about an Absent Cause: Discounting and Augmentation of Positively and Independently Related Causes**

Frank Van Overwalle & Bert Timmermans

## **Abstract**

Standard connectionist models of pattern completion like an auto-associator, typically fill in the activation of a missing feature with internal input from nodes that are connected to it. However, associative studies on competition between alternative causes, demonstrate that people do not always complete the activation of a missing feature, but rather actively encode it as missing whenever its presence was highly expected. Dickinson and Burke's revaluation hypothesis [4] predicts that there is always forward competition of a novel cause, but that backward competition of a known cause depends on a consistent (positive) relation with the alternative cause. This hypothesis was confirmed in several experiments. These effects cannot be explained by standard auto-associative networks, but can be accounted for by a modified auto-associative network that is able to recognize absent information as missing and provides it with negative, rather than positive activation from related nodes.

## **1. Introduction**

In connectionist models that produce pattern completion like an auto-associator, whenever a feature is missing, its activation is filled up with internal input from nodes that are connected to it [9]. However, in human induction, we so often do not complete the activation of a missing feature, but rather actively encode it as missing whenever its presence was highly expected on the bases of related features at input. This phenomenon of missing features has recently been studied in associative learning research, in the context of backward competition between alternative causes [4, 7, 15, 16]. Because standard connectionist networks [9] cannot explain backward competition, our aim is to provide an alternative connectionist account of this effect by modifying the standard auto-associative network.

## 1.1 Forward and Backward Competition

Competition refers to the tendency to alter the perceived strength of a cause or cue in the face of alternative or competing causal explanations. Perhaps the most well known illustration of competition is discounting (or blocking), which refers to the tendency to disregard or underestimate potential causes when the facilitatory influence of an alternative cause is already established. For instance, a person's success is attributed less to internal abilities given evidence of additional external aid. The opposite effect, augmentation (or superconditioning) refers to the tendency to overestimate the strength of a focal cause when it overcomes the inhibitory influence of an alternative cause. For instance, success is more strongly attributed to internal capacities when the task was hard.

Competition can have an effect on a cause that is either known or novel. When the alternative cause is already known and exerts its influence on a novel focal cause, this is called forward competition. For instance, when we know that someone is a hell of a good tennis player (A+, see Table 1) and when that player wins a doubles game with a novel partner (AT+), we tend to discount the contribution of the novel partner in the win. Conversely, when we know that someone is a poor player (A-), we tend to augment the contribution of the novel partner in the win (AT+).

Conversely, when a novel alternative cause exerts its influence afterwards, that is, on a known focal cause, this is called backward competition [4]. For instance, when we only recently learn that one of two partners of a well-known successful doubles tennis team (AT+, see Table 1) is now winning all his or her single games (A+), we tend to discount our initial high estimation of the other partner. Conversely, when one partner of a successful doubles team (AT+) is losing all his or her single games (A-), we are now likely to augment our initially evaluation of the other partner.

	Phase 1	Phase 2
Forward		
Discounting of T	A+	AT+
Augmentation of T	A-	AT+
Backward		
Discounting of T	AT+	A+
Augmentation of T	AT+	A-

Table 1: Schematic illustration of forward and backward discounting and augmentation. Note. T = target cause that is discounted or augmented, A = alternative cause that produces discounting or augmentation, + = focal outcome (e.g., winning a game), - = opposite outcome (e.g., losing a game).

## 1.2 Backward Competition and Missing Features

Previous associative theories of causal induction such as the popular model of Rescorla and Wagner [12] — which is identical to the delta learning algorithm embedded in a feedforward network [9] — are unable to account for backward competition. The reason is that in the original Rescorla-Wagner model, the absent cause on which competition (i.e., discounting or augmentation) should be exerted is assumed to have zero activation. Hence, its causal weight cannot be adjusted.

Van Hamme and Wasserman [13] therefore proposed to give absent causes negative activation that would result in backward competition. Dickinson and Burke [4] further hypothesized that such a negative activation of an absent cause would occur only if this absence was unexpected, for instance, after the cause had formed a consistent relation with the alternative cause. Thus, it was predicted that backward discounting and augmentation are more likely when "target and competing cues were consistently paired during compound training" [4, p. 73]. Note that this differs from forward competition, which occurs regardless of the relation between causes in line with the original Rescorla-Wagner model [12].

To elucidate people's reaction to missing features, in research on causal competition, the relation between causes was manipulated [4, 7, 15, 16]. This relation was either consistent (always paired together) or varied (paired with many other causes), resulting in a positive or independent statistical relation respectively between the causes. Once this relationship was established, researchers investigated the effect of deleting one of the causes from the input (i.e., backward competition). A schematic example of such a procedure used by Van Overwalle and Timmermans [15] is given in Table 2.

The results revealed either discounting when the effect occurred even in the absence of a focal cause, or augmentation when the effect reversed in the absence of a focal cause. More importantly, this only occurred when the relation between the causes was positive (consistent pairing) as predicted by Dickinson and Burke's revaluation hypothesis. These findings have been replicated in several experiments [4, 7, 15, 16].

## 2. A Modified Auto-Associative Network

In this section, we develop a connectionist approach to the backward revaluation hypothesis of Dickinson and Burke [4] by introducing a modification to the standard auto-associative network [9]. The key idea is that this modification recognizes absent but expected input as missing and provides it with negative, rather than positive activation from related nodes.

To account for backward revaluation, Van Hamme and Wasserman [13] proposed that absent causes could take on negative activation when their absence is unexpected. Dickinson and Burke [4] further refined this proposal and asserted that the expectation that a cause should be present depends on the relationship with other causes: "Only the omission of an expected cue should generate a ... negative

activation ... it is the formation of within-compound associations during the first stage of training that provides the basis for this expectation" [4, p. 63]. To allow such within-compound connections (i.e., between all nodes including nodes representing a causal input), we used a recurrent network and more in particular, an auto-associator [9].

Block & Relation	Trials		
Block 1: Target & Alternative Causes ("compound" trials)			
Consistent	<i>A &amp; B</i>	<i>A &amp; B</i>	<i>A &amp; B</i>
	<i>R &amp; S</i>	<i>R &amp; S</i>	<i>R &amp; S</i>
	<i>X &amp; Y</i>	<i>X &amp; Y</i>	<i>X &amp; Y</i>
Varied	<i>A &amp; B</i>	<i>A &amp; S</i>	<i>A &amp; Y</i>
	<i>R &amp; B</i>	<i>R &amp; S</i>	<i>R &amp; Y</i>
	<i>X &amp; B</i>	<i>X &amp; S</i>	<i>X &amp; Y</i>
Block 2: Alternative Cause Only			
	B	B	B
	S	S	S
	Y	Y	Y

Table 2: Illustration of a Backward Competition Design. Note. Each row represents three trials. Target causes A, R, X are in italic. In forward competition, the order of Blocks 1 and 2 was reversed. In the control conditions, Block 2 was omitted. To illustrate, a trial from Block 1 reads as follows: "[Target and alternative names] won their doubles tennis game", and from Block 2 reads as follows: "[Alternative name] won her singles tennis game".

In an auto-associative network, causes and outcomes are represented in nodes that are all interconnected. The perceived influence of a cause on an outcome is reflected in the weight connecting the cause with the outcome. Processing information in this model takes place in two phases. In the first phase, the activation of the nodes is computed, and in the second phase, the weights of the connections are updated. We describe the auto-associative model in some more detail below together with a discussion of our modifications.

## 2.1 Activating Present and Absent Nodes

During the first phase of information processing, each node in the network receives activation from external sources. Because the nodes are all interconnected, this activation is then spread throughout the network where it influences all other nodes.

The activation coming from the other nodes is called the internal input. Together with the external input, this internal input determines the final pattern of activation of the nodes, which reflects the short-term memory of the network.

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

$$int_i = \Sigma (activation_j * weight_{ij}), \quad (1)$$

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

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

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

$$\Delta activation_i = net_i - D * activation_i, \quad (3)$$

where  $D$  reflects a memory decay term. To simplify the recurrent model and to demonstrate its relationship with the original Rescorla-Wagner model of associative learning, in the present simulations, we used only one internal updating cycle and the parameter values  $D = I = E = 1$ . Given these simplifying assumptions, the final activation of node  $i$  reduces simply to the sum of the external and internal input, or:

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

A typical emergent feature of an auto-associative model is that an absent input stimulus  $m$  will be assimilated. That is, missing input will generally be 'filled up' by similar activation from the internal input. This is so, because although an absent cue does not receive any external input ( $ext_m = 0$ ), it still receives internal input from all other nodes with which it is connected.

However, to simulate Dickinson and Burke's backward revaluation hypothesis [4], some modifications were introduced to this typical assimilation effect of a standard auto-associative network. Essentially, this modification prescribes that during learning by the network, an absent stimulus  $m$  that is clearly anticipated through a connection with another stimulus  $j$ , will not be 'filled up' by the activation coming from node  $j$ , but rather will receive an opposite activation resulting in a contrast or correction. Specifically, when the activation from one of the other nodes  $j$  exceeds some 'missing threshold' (denoted by  $\mu$ ), then the missing stimulus  $m$  accrues *missing input* rather than internal input.

In mathematical terms, for any node  $m$  where  $ext_m = 0$  and  $|activation_j * weight_{mj}| > \mu$ , Equation 1 is replaced by

$$miss_m = \Sigma (activation_j * weight_{mj}). \quad (4)$$

Note that the missing input is summed only for those nodes  $j$  where the missing threshold  $\mu$  is exceeded. For the final activation pattern, the missing input is then subtracted from the external input (instead of being added as in Equation 2), effectively resulting in a contrast or correction effect. Hence, once the missing threshold of an absent input  $m$  is exceeded, Equation 2 is replaced by

$$net_m = E ext_m - I miss_m. \quad (5)$$

After making the simplifying assumptions of one internal updating cycle and parameter values  $D = I = E = 1$ , this becomes

$$activation_m = ext_m - miss_m = - miss_m. \quad (5')$$

## 2.2 Weight Updating

After this first phase, the auto-associative model then enters in its second learning phase, where the short-term activation is consolidated in long term weight changes to better represent and anticipate future external input. Here we follow the standard assumption of the auto-associator [9, p. 166]. Basically, weight changes are driven by the discrepancy between the internal input from the last but one updating cycle of the network and the external input received from outside sources, formally expressed in the delta algorithm:

$$\Delta weight_{ij} = \varepsilon * (ext_i - int_j) * activation_j, \quad (6)$$

where  $\varepsilon$  is a learning rate that determines how fast the network learns.

In summary, the proposed recurrent network can reproduce not only the typical assimilation effect of standard auto-associative networks but also contrastive reevaluation as proposed by Dickinson and Burke [4]. When the presence of a missing stimulus is not greatly anticipated (because of weak interconnections with other nodes), a typical assimilation effect will ensue. In contrast, when the absence of a missing stimulus is very much unexpected (because of strong connections with other nodes), a contrast effect will result.

## 3. Simulations

The predictions of the modified recurrent network were tested with the data from Van Overwalle and Timmermans [15]. The results of these experiments, and of the simulation (to be explained shortly) are shown in Figure 1.

As predicted by Dickinson and Burke's reevaluation hypothesis [4], there was significant discounting in comparison with the control conditions (see top panel), except with the backward varied condition. Likewise, there was also significant augmentation in comparison with the control conditions (see bottom panel) but less so with the backward varied condition. Although the predicted attenuation of augmentation was less clear-cut (but see [4, 7] for stronger effects), with a backward order, augmentation was still significantly stronger in a consistent than varied relation, while this difference failed to reach significance with a forward order as predicted.

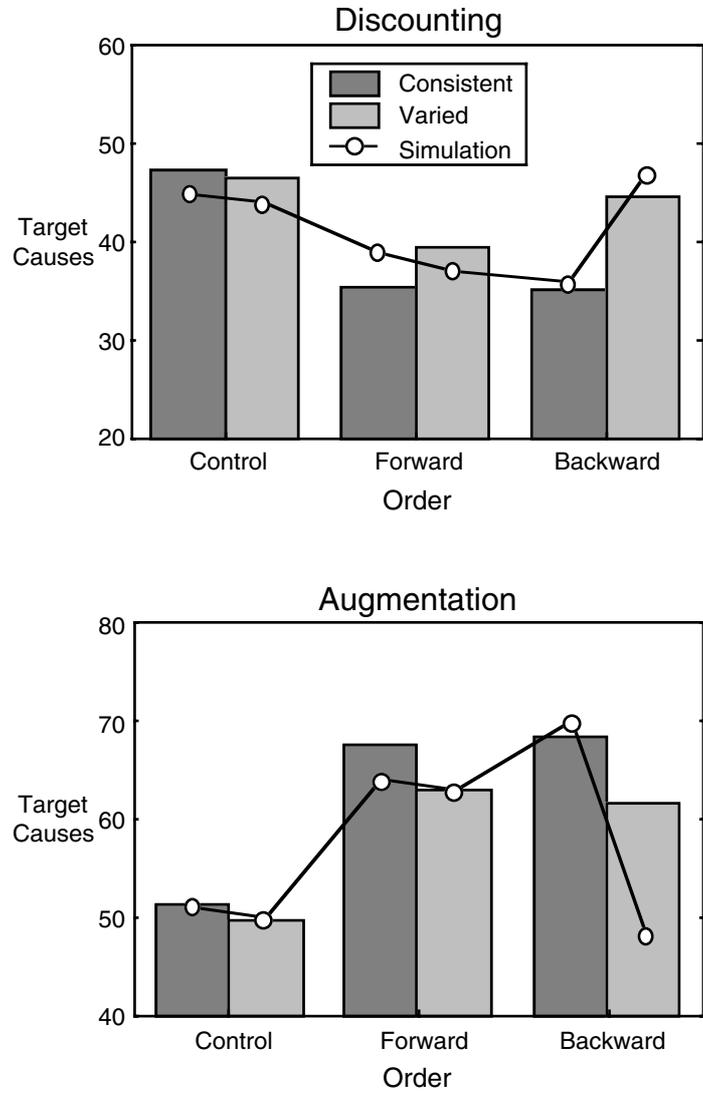


Figure 1: Discounting and Augmentation: Causal ratings in function of Order and Relation (based on Van Overwalle & Timmermans [15]).

In our simulations, we used an auto-associative network architecture with a single node for each cause or outcome, and the modified activation updating algorithm as explained above. The model was run using exactly the same order of trials and blocks as in the experiments. Because trial order was randomized within

blocks, we ran 50 simulations with a random trial order for each block, and averaged the results.

The presence of a cause at a trial was encoded by setting the external input to +1; otherwise the external activation remained at resting level 0. Likewise, a target outcome was encoded by an external input of +1, and the opposite outcome by an external input of -1. The weights of the connections were updated after each trial. In the simulations that we report, we explored different values for the missing threshold parameter  $\mu$  until one was found that resulted in the best visual fit with the observed data. The other model parameters were kept constant ( $\epsilon = .10$ ,  $E = I = D = 1$ ,  $cycles = 2$ ). These parameters were chosen so as to increase the similarity with the original Rescorla-Wagner associative model on which Dickinson and Burke's [4] backward revaluation hypothesis rests.

At the end of each simulated trial history, the causal influence of the target and alternative stimuli was tested by turning the external input of the corresponding nodes to +1 and reading off the resulting activation of the output node. During this testing phase, the revaluation process was not active (i.e., standard Equation 2 was executed instead of the 'missing' correction of Equation 5). The obtained simulation values were projected onto the observed ratings using linear regression (with slope > 0), to demonstrate visually the fit of the simulations.

The full lines in Figure 1 show the results. As can be seen, consistent with the data and Dickinson and Burke's backward revaluation hypothesis, the simulations predict an attenuation of discounting and augmentation for the target ratings in the backward varied conditions only. The simulations matched the discounting data (top panel) very well, as can be verified also from the correlation between simulated and observed means,  $r = .981$  ( $\mu = .15$ ). The observed mean ratings included not only the target causes as shown in Figure 1, but also the alternative causes not shown. However, the fit was less adequate for the augmentation data (bottom panel, as noted earlier, these data showed less of the predicted pattern), although the correlation was still high,  $r = .938$  ( $\mu = .20$ ). These simulation results were not obtained with a standard recurrent model without our revaluation modification.

## 4. Discussion

To account for backward competition, learning theories must take into account the order of presentation at encoding and the interrelations between causes or cues. Because probabilistic and other rule-based models of causality are not sensitive to order or inter-cause relationships [2, 3, 6, 11], they cannot account for the present data. It is not easy to see how these two fundamental flaws can be amended in a reformulation of these theories. Alternative theories that posit that all competition takes place at the time of testing [1, 10] also fail to account for backward competition.

The backward revaluation data also pose difficulties for standard feedforward connectionist models because they fail to incorporate interrelations between causes [12, 14]. Although recurrent auto-associative models include interrelations between

causes [9], we demonstrated that they need an additional corrective mechanism to reflect the revaluation hypothesis, that is, to create a special status for positively related causes that are expected but "missing". Currently, these recurrent models only assume that that absent concepts are "filled up" with activation from related concepts in memory, which leads to less rather than more discounting and augmentation, contrary to the revaluation hypothesis. Graham [5] developed a similar corrective revaluation mechanism into a recurrent network that is, however, not so successful in replicating our data.

Our network model of backward revaluation can also account for related phenomena that are also potentially driven by a revaluation mechanism. An obvious case is conditioned inhibition, which is exactly the opposite of augmentation. Conditioned inhibition refers to the tendency to increase the inhibitory strength of a focal cause when it attenuates the facilitatory influence of an alternative cause. For instance, failure is more strongly attributed to lack of internal capacities when the task was easy. The mechanism that explains backward conditioned inhibition is identical to that for augmentation, the only difference being the negative direction of the outcome. Simulations of our proposed modified auto-associator with the backward conditioned inhibition data of Larkin et al. [7, exp. 4] showed a high fit.

Another illustrative case is backward competition between causes that occur sequentially (and thus are, in statistical terms, negatively related). For instance, as suggested by the proverb that new brooms sweep clean, novel leaders who replace older management often get the sole credit for success, although their success often depends in part on prior managerial decisions. This indicates that initial causes can be discounted by subsequently presented alternative causes. Matute and Pineño [8] recently documented such backward discounting. According to our proposed corrective mechanism for missing nodes, because the initial causes are expected given their prior pairings with the outcome, their omission is very salient and results in downward revaluation. Simulations of our modified auto-associator with the data of Mutate and Pineño [8, exp. 1] again showed a substantial fit.

## References

1. Bouton, M. E. (1993). Context, time, and memory retrieval in the inference paradigms of Pavlovian learning. *Psychological Bulletin*, *114*, 80—99.
2. Busemeyer, J. R. (1991) Intuitive statistical estimation. In N. Anderson (Ed.) *Contributions to Information integration theory, vol. 1: Cognition*. Hillsdale, NJ: Erlbaum.
3. Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*, 545—567.
4. Dickinson, A. & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgments. *Quarterly Journal of Experimental Psychology*, *49B*, 60-80.
5. Graham, S. (1999). Retrospective revaluation and inhibitory associations: Does perceptual learning modulate our perceptions of the contingencies between events? *Quarterly Journal of Experimental Psychology*, *52B*, 159-185.

6. Hogarth, R. M. & Einhorn, H. J. (1992) Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, 1—55.
7. Larkin, M. J. W., Aitken, M. R. F., Dickinson, A. (1998). Retrospective reevaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology, Learning, Memory, and Cognition*, 24, 1331—1352.
8. Matute, H. & Pineño, O. (1998). Stimulus competition in the absence of compound conditioning. *Animal Learning and Behavior*, 26, 3-14.
9. McClelland, J. M. & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. Cambridge, MA: Bradford.
10. Miller, R. R. & Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 51—92). San Diego, CA: Academic Press.
11. Morris, M. W. & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of Discounting in causal attribution. *Psychological Review*, 102, 331—355.
12. 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.
13. Van Hamme, L. J. & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, 25, 127—151.
14. 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.
15. Van Overwalle, F. & Timmermans, B. (2000). *Discounting and Augmentation in Attribution: The Role of the Relationship between Causes*. Manuscript submitted for publication.
16. Wasserman, E. A. & Berglan, L. R. (1998). Backward blocking and recovery from overshadowing in human causal judgment: the role of within-compound associations. *Quarterly Journal of Experimental Psychology*, 51B, 121—138.