

# Learning Causal Structure from Reasoning

Aron K. Barbey (abarbey@emory.edu) and Phillip Wolff (pwolff@emory.edu)

Department of Psychology, Emory University  
532 Kilgo Circle, Atlanta, GA 30322 USA

## Abstract

According to the transitive dynamics model, people can construct causal structures by linking together configurations of force. The predictions of the model were tested in two experiments in which participants generated new causal relationships by chaining together two (Experiment 1) or three (Experiment 2) causal relations. The predictions of the transitive dynamics model were compared against those of Goldvarg and Johnson-Laird's model theory (Goldvarg & Johnson-Laird, 2001). The transitive dynamics model consistently predicted the overall causal relationship drawn by participants for both types of causal chains, and, when compared to the model theory, provided a better fit to the data. The results suggest that certain kinds of causal reasoning may depend on force dynamic—rather than on purely logical or statistical—representations.

**Keywords:** Causal reasoning; causal structures; lexical semantics; knowledge representation.

## Introduction

People sometimes acquire new knowledge from what they already know (Genter & Wolff, 2000). They transform their knowledge to represent the world in new ways, or rearrange what they know to arrive at new conclusions. Sometimes learning involves taking conceptual structures apart so that new structures can be formed in their place. Acquiring knowledge in this manner can be viewed as *learning from reasoning*.

Learning from reasoning is extremely common, but we know relatively little about how it occurs. Most of the research on learning has examined how people discover the statistical properties of the input (e.g., Redington, Chater & Saffran, 1998; Newport & Aslin, 2004). Statistical approaches to learning offer accounts of many important phenomena, but they also face several difficult challenges. In particular, it is not clear how statistical approaches might give rise to deeper knowledge about a domain, specifically, the causal structure of a domain (e.g., Marcus, 1998).

There have been several proposals in the Bayes' net literature suggesting how people might acquire networks of causal relationships (Pearl, 2000; Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005; Tenenbaum & Griffiths, 2001; Griffiths & Tenenbaum, 2005). However, as noted by Gopnick et al. (2004), some of these techniques are psychologically unrealistic due to the high memory and processing requirements they place on the learner. Other discovery procedures

are more psychologically realistic but limit learning to situations where the learner has direct experience with sets of events from which the probability distributions of the variables can be inferred. Such situations do not seem consonant with the way in which causal relations are often learned, that is, through simple verbal or written descriptions of causal relations, as in a classroom. Once individual causal relations are learned, it may be possible to combine them into larger structures. It is in this process of combining causal relations that learning from reasoning may play a critical role in the acquisition of the deeper knowledge of a domain.

Imagine, for example, a person with the knowledge that *vegetation prevents erosion* and that *erosion causes landslides*. These two beliefs can be represented as nodes and links (see Figure 1). At some point, this person might connect these two assertions to form a new conceptual structure and a new causal relationship, namely that *vegetation prevents landslides*.

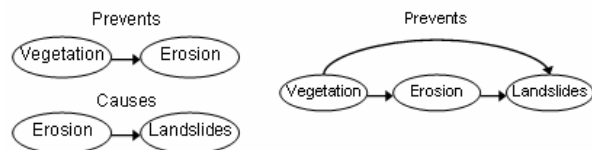


Fig. 1. Relationships can combine to form a new structure.

In this paper we examine how the process of joining causal relationships might occur. Two models of structure building will be investigated. One of these, the *transitive dynamics model*, extends previous work on the representation of causal relations using vector semantics (Wolff, 2007). The other, the *model theory* (Goldvarg & Johnson-Laird, 2001), extends Johnson-Laird's (2001) theory of mental models.

## Vector semantics

Everyday language suggests that we think about causal relations in terms of force. We say, for example, *The force of his argument changed my mind*, or *Peer pressure led my son to skip class*, or *The moral force of his argument persuaded me to make a contribution*. Such statements suggest that people might even reason with representations that reflect forces. Recent work in vector semantics suggests how these intuitions might be realized computationally. The transitive dynamics model is a natural extension of the dynamics model

(Wolff, 2007), which in turn is based on Talmy's (1988) theory of force dynamics.

The dynamics model holds that people represent causal relations in a manner that partially copies or reproduces the way in which causal relationships are instantiated in the world. Specifically, the dynamics model holds that people represent causal relationships in terms of configurations of forces. The sentences in 1 provide some intuitive support for this view.

- 1 a. Pressure will cause the water to remain liquid below 0°C.
- b. Small ridges cause water to stand on the concrete.
- c. The rubber bottom will cause the cup to stay in place.
- d. The pole will prevent the tent from collapsing.

In each situation described in sentences 1a-d, nothing happens. Because nothing happens, there is no regular sequence of events or transfer or exchange of energy, at least at the macro-level. What is present in each of these situations is a configuration of forces. According to the dynamics model, it is this configuration of forces that makes these situations causal (sentences 1a-c) or preventative (sentences 1d).

Table 1. Representations of several causal concepts.

	Patient tendency for endstate	Affector-patient concordance	Endstate approached
CAUSE	N	N	Y
ALLOW	Y	Y	Y
PREVENT	Y	N	N

The dynamics model holds that the concept of CAUSE and related concepts involve interactions between two main entities: an *affector* and a *patient* (the entity acted on by the affector). It also holds that different causal relationships can be specified in terms of three dimensions: a) the *tendency* of the patient for an endstate, b) the presence or absence of *concordance* between the affector and the patient, and c) *progress toward the endstate* (essentially, whether the result occurs). Table 1 summarizes how these dimensions differentiate the concepts of CAUSE, ALLOW, and PREVENT. According to the dynamics model, when we say *Wind caused the boat to heel*, we mean that the patient (the boat) had no tendency to heel (Tendency = No), the affector (the wind) acted against the patient (Concordance = No) and the result (heeling) occurred (Endstate approached = Yes). When we say *Rain prevented the tar from bonding*, we imply that the patient (the tar) had a tendency to bond (Tend. = Y), the affector (rain) opposed this tendency (Con. = N), and the result did not occur (Endstate = N).

The dynamics model specifies how these three dimensions can be captured in terms of configurations of force vectors. Real-world vectors have precise origins, directions, and magnitudes. Vectors in people's minds, in contrast, are assumed to be more qualitative. Specifically, mental vectors are predicted to be relatively accurate with respect to direction but often imprecise with respect to—although not completely

insensitive to—magnitude. It is hypothesized that people's notions of force vectors are not limited to physical forces but also include social and psychological forces which, like physical forces, can be understood as quantities that influence people in a certain direction.

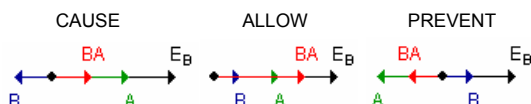


Fig. 2. Configurations of forces associated with CAUSE, ALLOW, and PREVENT. **A** = the affector force, **B** = the patient force. **BA** = the resultant of **A** and **B**. **E** = endstate

In the language of vectors, a patient has a tendency for the endstate when the vector associated with the patient, **B**, is in the same direction as the vector that specifies the endstate, **E**. Given the features specified in Table 1, this means that the patient vector will be in the same direction as the endstate vector in the cases of ALLOW and PREVENT, but not CAUSE, as shown in Fig. 2. The patient and the affector will be in concordance when the vectors associated with the affector and the patient are in the same direction. This is expected to occur in the case of ALLOW, but not CAUSE or PREVENT. Finally, the result is expected to occur when the resultant vector is in the same direction as the endstate vector. This is held to occur in the case of CAUSE and ALLOW, but not PREVENT.

Support for the model was provided in a series of experiments in which participants categorized 3-D animations of realistically rendered objects with trajectories that were wholly determined by the force vectors entered into a physics simulator (Wolff, 2007). (The animations can be viewed at <http://userwww.service.emory.edu/~pwolff/CLSAnimations.htm>.) In these experiments, the very same physical forces used to generate physical scenes were used as inputs into a computer model to predict how those scenes would be described. The top of Table 2 shows the directions and magnitudes of the force vectors

Table 2. Results from Experiment 1 (Wolff, 2007)

Config. #	1	2	3	4
Affector (→)				
Patient (→)	E ← ← ← ← ←	E ← ← ← ← ←	E ← ← ← ← ←	E ← ← ← ← ←
Result. (→)				
Predicted	CAUSE	ENABLE	ENABLE	PREVENT
"Cause"	94%	11%	6%	-
"Help"	6%	89%	94%	-
"Prevent"	-	-	-	100%
"No verb"	-	-	-	-
	5	6	7	8
Affector (→)				
Patient (→)	E ← ← ← ← ←	E → → → → →	E ← ← ← ← ←	E → → → → →
Result. (→)				
Predicted	Unspecified	Unspecified	Unspecified	Unspecified
	-	-	6%	-
	11%	-	-	-
	6%	-	-	6%
	83%	100%	94%	94%

associated with the affector and patient that were entered into the physics simulator and their resultant. Below these configurations are the causal categories predicted by the model and participants' (N=18) responses. The fit to the model was excellent.

Additional experiments in Wolff (2007) show that the predictions of the dynamics model extend to two-dimensional interactions and to the identification of social causation. In sum, the results provide support for the hypothesis that people think about causal relationships in terms of configurations of forces.

*Representing negation.* In addition to explaining the core meaning of the concepts of CAUSE, ALLOW, and PREVENT, the dynamics model offers an explanation of how negation transforms the meaning of a causal statement. When the consequent is negated, the configuration of forces is interpreted with respect to the inverse of the endstate vector,  $\sim E$  (see Fig. 3). The dynamics model predicts, then, that CAUSE\_NOT (A causes not-B) has the same meaning as PREVENT (e.g., *Pain causes lack of sleep* means, roughly, *Pain prevents sleep*).

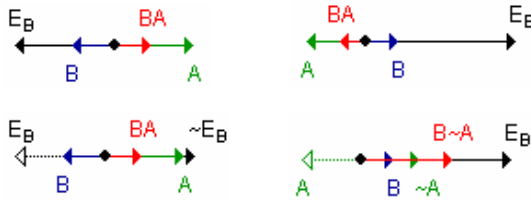


Fig 3. Config.s associated with A prevents B and A causes not-B

In contrast, when the antecedent is negated, the affector vector is reversed,  $\sim A$ , (see Fig. 4). Such a transformation predicts that NOT\_ALLOW is also linked to PREVENT. Once again, this correspondence is supported by intuition (e.g., *Absence of snow allows construction* implies *Snow prevents construction*). Note, however, that the relationship between NOT\_ALLOW and PREVENT is one of implication, not synonymy. In the case of synonymy, the forces are the same but the position vector changes (see Fig. 3). In implication, the forces themselves differ (A is replaced by  $\sim A$ ; see Fig. 4).

Speakers often use negation, though not always obviously. Besides the actual word *not*, people may use phrases like *the absence/lack of* or *fail to*. Sometimes the *not* is included in the meaning of the word (e.g., *stay for not go*). In addition, several types of negation can appear in the same sentence, as when we say *The vegan diet does not cause lack of vitamins* or *Lack of fitness causes lack of attention*. Explaining the effect of negation is an important problem for any theory of causal meaning. With these assumptions in place, we

can now begin to describe how causal relations and networks might be learned through reasoning.

### Transitive Dynamics Model

Whereas the dynamics model is a theory of how individual causal relations are represented, the *transitive dynamics model* is a theory of how those relations are combined to form new relations. Consider, for example, the causal relations *Cell phones cause accidents* and *Accidents causes delays*. They can be combined to form a new relation: *Cell phones cause delays*. In the language of syllogistic reasoning, the two causal relations are *premises* that form an *argument* that leads to a *conclusion*. In the transitive dynamics model, the conclusion is generated by connecting the premises. According to the model, the first and second premises are connected by using the resultant vector in the first premise (BA) as the affector vector in the second premise (B<sub>BA</sub>) (see Fig. 5).

The direction of the affector in the second premise is the same as the resultant in the first premise, unless the B terms in the two premises conflict (i.e., one is negated), in which case the direction of the affector in the second premise is reversed. A conclusion (i.e., new relation) is drawn by forming a new configuration of forces based on the two premises. The affector in the conclusion is the affector from the first premise; the endstate vector in the conclusion is the endstate vector from the last premise; and the patient in the conclusion is the resultant of the patient vectors in the premises. With all of these vectors in place, the conclusion can be interpreted like any other configuration of forces.

Findings from Wolff's (2007) Experiment 4 indicate that people's representations of force are often underspecified with respect to magnitude. Not knowing the exact magnitude of the forces adds indeterminacy to people's representations of causation. The effects of this indeterminacy can have consequences when configurations of force are combined: variations in the magnitudes of the forces can lead to more than one possible conclusion. Thus, the transitive dynamics model offers an explanation of how causal interpretations might shift from deterministic to probabilistic. Consider, for example, the left panel in Figure 5. In an argument in which both relations are

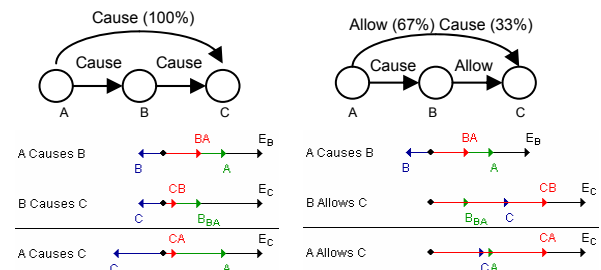


Fig 5. Transitive arguments and configurations of force

CAUSE, the conclusion will also always be CAUSE. According to the model, this occurs because the patient vectors in the premises (**B** and **C**) point in the same direction, so their resultant, which forms the patient vector in the conclusion, will also be in the same direction. In contrast, when the premises of the argument are CAUSE and ALLOW relations, more than one conclusion is possible, depending on the relative magnitudes of the patient vectors.

A program has been written that implements the combination procedure above and that allows users to conduct simulations in which the magnitudes of the vectors are systematically varied or randomly specified (<http://userwww.service.emory.edu/~pwofff/Transitivedynamics.htm>). The only constraint on the magnitudes is that they preserve the relations in the premises. The program tallies the conclusions that follow from each set of magnitudes. For the ALLOW / CAUSE argument (A/C), for example, the program finds that if the magnitudes of the vectors are systematically varied, the premises will lead to ALLOW and CAUSE conclusions 67% and 33% of the time. The program can generate conclusions for chains of up to 25 relations. In generating conclusions, the program may examine many thousands of vector magnitudes. However, it is not assumed that people consider thousands of magnitudes. Indeed, the simulation program shows that the predictions of the model will emerge within an individual if they consider a small set of possible magnitudes or across a small set of people (e.g., 5-10), who each consider only one set of magnitudes.

*Goldvarg and Johnson-Laird's (2001) model theory.* The transitive dynamics model can be contrasted with another theory of causal reasoning, Goldvarg and Johnson-Laird's (2001) model theory. Goldvarg and Johnson-Laird (2001) associate the notions of CAUSE, ALLOW, and PREVENT with different combinations of possible co-occurrences (see Table 3 for representations of CAUSE and PREVENT). Negation simply involves representing the absence of the antecedent or consequent (e.g.,  $\neg A$ ). One procedure for combining relations is shown in Table 3. Across a wide range of arguments, the conclusion of the model theory is also the most frequent conclusion predicted by the transitive dynamics model. An important difference, however, is that for certain arguments the transitive dynamics predicts more than one possible conclusion. The predictions of the models were tested in the following experiments (see also Barbey & Wolff, 2006).

Table 3. Procedure for combining relations to draw causal inferences

Represent premises				Conjoin			Reduce		Interpret
A causes B	B prevents C			A	B	$\neg C$			
A	B	B	$\neg C$	$\neg A$	B	$\neg C$	A	$\neg C$	
$\neg A$	B	$\neg B$	C	$\neg A$	$\neg B$	C	$\neg A$	C	
$\neg A$	$\neg B$	$\neg B$	$\neg C$	$\neg A$	$\neg B$	$\neg C$	$\neg A$	$\neg C$	

## Experiment 1

### Method

**Participants** The participants were 30 Emory University undergraduates.

**Materials** The materials were based on real-world causal statements found on the internet. For example, for the argument not-A causes not-B and B causes C, people saw statements like *Leaf loss causes lack of shade* and *Shade causes cooling*. Six examples were found for all 32 argument types shown in Table 5 for a total of 192 arguments (384 causal statements).

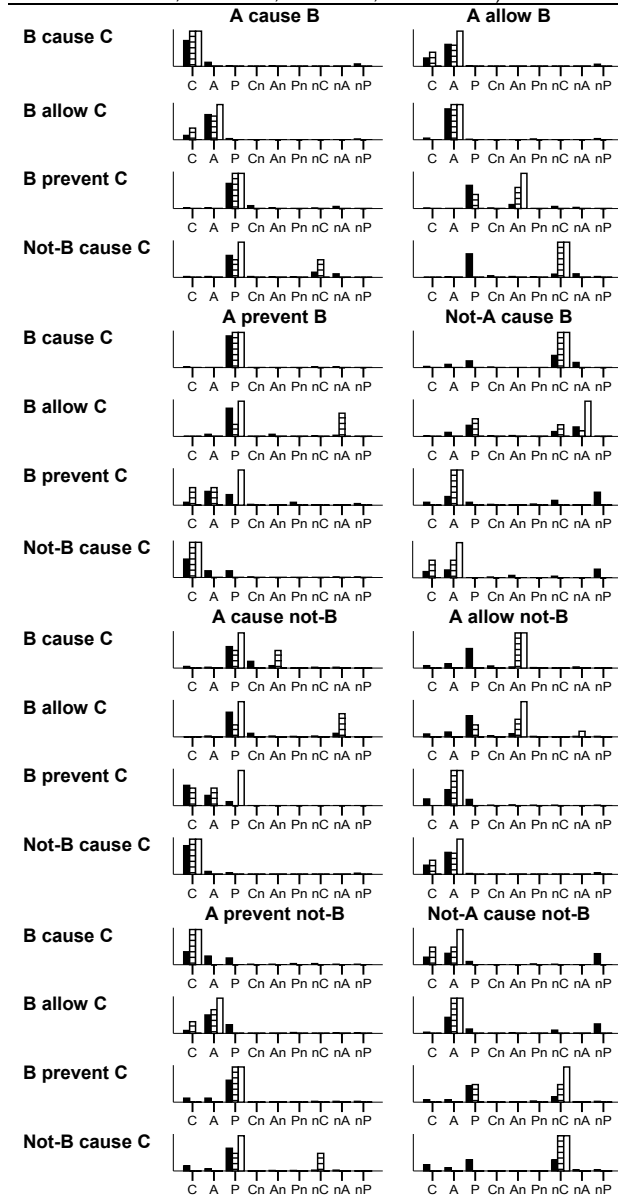
**Procedure** In this experiment, participants saw the 32 arguments (in Table 5) one at a time and then chose a conclusion from a list of ten possible conclusions (*A causes C, A allows C, A prevents C, A causes not-C, A allows not-C, A prevents not-C, not-A causes C, not-A allows C, not-A prevents C, or none of the above*).

### Results and Discussion

As Table 5 shows, participants' conclusions were well explained by the transitive dynamics model. To test the transitive dynamics model against the model theory, the predictions of each theory were correlated with people's responses to each of the arguments listed in Table 5. A paired *t*-test indicated that the average Spearman correlation between people's responses and the transitive dynamics model ( $M = .59$ ) was higher than the average Spearman correlation between people's responses and the model theory ( $M = .49$ ),  $t(31) = 3.80$ ,  $p = .001$ . The findings were the same when analyzed nonparametrically using the Wilcoxon Signed Ranks Test,  $Z = 3.39$ ,  $p = 0.001$ . The transitive dynamics model predicted participants' first or second response for all of the arguments. The model theory was also often able to predict participants' most frequent response. Nevertheless, the transitive dynamics model out-performed the model theory because it could identify not only the primary response but also secondary responses.

It should be noted that the predictions of each theory were based on a very conservative criterion of support. For example, according to both models, NOT-PREVENT implies ALLOW. However, in evaluating the models, we only counted the simplest possible expressions (e.g., ALLOW) as evidence in support of the models. We did this because neither model (currently) provides an account of how people choose between synonymous expressions of causation. The one exception to this rule was for arguments that could be interpreted as implying either NOT-CAUSE or ALLOW-NOT (e.g., A/P). For these arguments we predicted the conclusion that was most consistent with the atmosphere of the argument.

Table 5. Exp. 1 percentages (Black = observed; Striped = TD model; White = MT model; C=CAUSE, A=ALLOW, P=PREVENT)



The conclusion that results from a PREVENT/PREVENT (or C\_N/P) argument is especially interesting. The dynamics model predicts that the conclusion will either be CAUSE or ALLOW, which is how people responded. The model theory, in contrast, predicts that the conclusion will be PREVENT. The relation that arises from “to prevent a prevention” seems to capture the core meaning of the verbs allow, permit, and enable. It may be that the meaning of these verbs is based on a chain of prevent relations, rather than on an individual configuration.

## Experiment 2

In Experiment 2 we contrasted the transitive dynamics model and the model theory with respect to three-premise arguments.

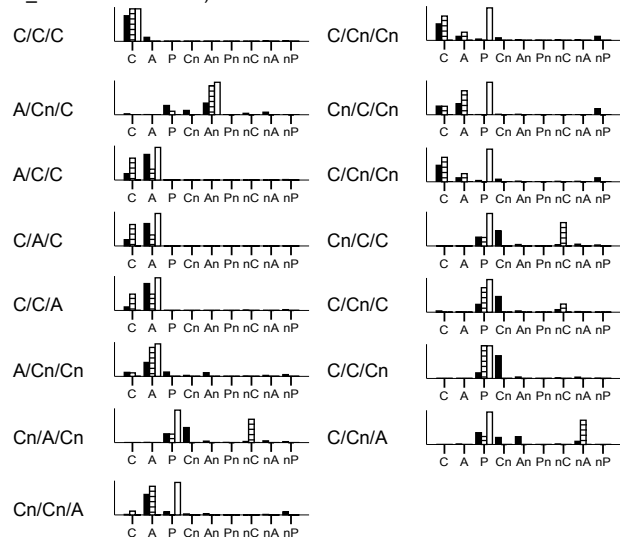
### Method

**Participants** The participants were 16 Emory University undergraduates.

**Materials** The materials were based on real-world causal statements found on the internet. For example, for the argument A allows B, B prevents C, and C causes D (i.e., A/P/C), people saw statements like *Vacations allow rest. Rest prevents exhaustion. Exhaustion causes sickness.* Participants saw four examples of fifteen kinds of 3-premise arguments (see Table 6) for a total of 60 arguments.

**Procedure** Participants chose conclusions for each argument from a list of 10 possible conclusions (*A causes D, A allows D, A prevents C, A causes not-D, A allows not-D, A prevents not-D, not-A causes D, not-A allows D, not-A prevents D, or none of the above*).

Table 6. Exp. 2 predictions and (%) results (Black = observed; Striped = TD model; White = MT model; C=CAUSE, A=ALLOW, P=PREVENT; C\_N = A CAUSE not-B)



### Results

As Table 6 illustrates, the results provided further support for the transitive dynamics model. To compare the dynamics model with the model theory, the responses to each argument were correlated with the predictions of each theory. A paired *t*-test indicated that the average Spearman correlation between people’s responses and the transitive dynamics model ( $M = .59$ ) was higher than the average Spearman correlation between people’s responses and the model theory ( $M = .36$ ),  $t(14) = 2.69$ ,  $p = 0.018$ . The findings were the same when analyzed nonparametrically using the Wilcoxon Signed Ranks Test,  $Z = 2.41$ ,  $p = 0.016$ . The

results provide further support for the hypothesis that certain causal arguments are consistent with more than one conclusion since the main difference between the predictions of the dynamics model and the model theory was in the possibility of more than one response (in the case of the dynamics model). In sum, the transitive dynamics model offers a framework that extends from initial perception of causal relations (Wolff, 2007) to causal reasoning.

## Conclusions

Together, the dynamics model and the transitive dynamics model offer an account of how people initially acquire individual relationships and then combine those relationships to form new causal relationships and structures. As noted in Wolff (2007), people might acquire causal relationships from the perception of configurations of forces. However, the dynamics model also offers an explanation for how causal relationships might be acquired from simply hearing or reading statements of causation (e.g., *CO<sub>2</sub> emissions are causing global warming*). This is possible because the dynamics model is also a model of causal meaning (Wolff & Song, 2003). An account of the acquisition of causal relations through language is critical since language is arguably the main source of causal knowledge (Sloman, 2005). Clearly, people do not learn that *CO<sub>2</sub> emissions cause global warming* on the basis of their own personal experience.

The results from Experiments 1 and 2 also support a new account of negation. The transitive dynamics model offers an explanation for how the absence of an influence can be a cause (e.g., *Lack of nutrition causes hair loss*). An absent cause is a force whose absence allows another force to produce an effect. Negation in the dynamics model extends the breadth of the model from 4 to 32 different possible expressions of causation. In addition, the dynamics model, unlike the model theory, can differentiate relations of synonymy from relations of implication.

In the transitive dynamics model, individual causal relationships are deterministic, but causal relationships involving non-contiguous factors can become probabilistic. This is especially the case when the transitive dynamics model is extended to causal structures involving converging and diverging causal relations (not discussed in this paper). The dynamics model offers an explanation for how people's estimates about the probabilities of events might be encoded in the structure of their causal network.

Finally, the dynamics model offers an account of how causal knowledge might be shared across people. People regularly engage in causal reasoning without full knowledge of the causal interactions involved. To make reasonable inferences and decisions, they may

compensate for their incomplete representations by drawing on the knowledge of others through verbal communication. The dynamics model explains how verbal communication might be enough to begin the process of causal reasoning and knowledge creation.

## Acknowledgments

This work was supported by NSF Grants DGE-0536941 and DGE-0231900 to Aron K. Barbey and by an award from the University Research Committee of Emory University to Phillip Wolff.

## References

- Barbey, A.K. & Wolff, P. (2006). Causal reasoning from forces. In *Proceedings of the 28th Annual Conference of the Cog. Science Society* (p. 2439). Mahwah, NJ: Erlbaum.
- Genter, D. & Wolff, P. (2000). Metaphor and knowledge change. In E. Dietrich, & A. Markman, *Cognitive Dynamics: Conceptual Change in Humans and Machines* (pp. 295-342). NJ: Lawrence Erlbaum.
- Goldvarg, E. & Johnson-Laird, P. (2001). Naive causality: a mental model theory of causal meaning and reasoning. *Cognitive Science*, 25, 565-610.
- Gopnik, A., Glymour, C., Sobel, D., Shulz, L., Kushnir, T. & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psych. Review*, 111, 1-31.
- Griffiths, T. L. & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cog. Psych.* 51, 334-384.
- Johnson-Laird, P. (2001). Mental models and deduction. *TRENDS in Cognitive Science*, 5, 434-442.
- Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, 37, 243-282.
- Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48, 127-162.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.
- Redington, M., Chater, N. & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, 425-469.
- Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford: OUP.
- Sloman, S. A. & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science*, 29, 5-39.
- Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, 12, 49-100.
- Tenenbaum, J. B. & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Dietterich & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (pp. 59-65). Cambridge, MA: MIT Press.
- Waldmann, M. R. & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 216-227.
- Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*.
- Wolff, P. & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cogn. Psych.*, 47, 276-332.