A Generative Model of Causal Cycles

Bob Rehder (bob.rehder@nyu.edu)
Jay B. Martin (jbmartin@nyu.edu)
Department of Psychology, New York University
6 Washington Place, New York, NY 10003 USA

Abstract

Causal graphical models (CGMs) have become popular in numerous domains of psychological research for representing people's causal knowledge. Unfortunately, however, the CGMs typically used in cognitive models prohibit representations of causal cycles. Building on work in machine learning, we propose an extension of CGMs that allows cycles and apply that representation to one real-world reasoning task, namely, classification. Our model's predictions were assessed in experiments that tested both probabilistic and deterministic causal relations. The results were qualitatively consistent with the predictions of our model and inconsistent with those of an alternative model.

We naturally reason about causally related events that occur in cycles. In economics, we expect that an increase in corporate hiring may increase consumers' income and thus their demand for products, leading to a further increase in hiring. In meteorology, we expect that melting tundra due to global warming may release the greenhouse gas methane, leading to yet further warming. In psychology, we expect that clinicians will affect (hopefully help) their clients but also recognize the clients often affect the clinicians.

Many psychologists investigate causal reasoning using a formalism known as *Bayesian networks* or *causal graphical models* (hereafter, *CGMs*). CGMs are one hypothesis for how people reason with causal knowledge. There are claims that causal learning amounts to acquiring the structure and/or parameters of a CGM (Cheng, 1997; Gopnik et al., 2004; Griffiths & Tenenbaum, 2005; 2009; Lu et al., 2008; Sobel et al., 2004; Waldmann et al., 1995). And, many models of causal reasoning assume that people honor the inferential rules that accompany CGMs (Holyoak et al., 2010; Lee & Holyoak, 2008; Rehder & Burnett, 2005; Rehder, 2003; 2009; Rehder & Kim, 2010; Shafto et al., 2008; Sloman & Lagnado, 2005; Waldmann & Hagmeyer, 2005). Unfortunately, because standard CGMs prohibit the presence of causal cycles, these models are unable to represent any of the cyclic events mentioned above.

In this article, we take the initial steps to extend CGMs using an 'unfolding' trick from machine learning (Spirtes, 1993). We discuss the implications of this approach to one class of reasoning problem, namely classification. There is a rich literature on how causal knowledge among the features of a category changes how people classify. We first review evidence for causal cycles among category features and one proposal for how they affect classification. We then report two experiments that test that account. Finally, we present our own model for extending CGMs to represent cycles in people's mental representations of categories.

Unfolding Cycles

One technique used to elicit people's beliefs about the causal structure of categories is the *theory drawing task*. Subjects are presented with category features and asked to draw directed edges indicating how those features are causally related. These drawings show that causal cycles are common. For example, Kim and Ahn (2002) found that 65% of subjects' representations of mental disorders such as depression included cycles. Sloman et al. found numerous cycles in subjects' theories of everyday biological kinds and artifacts.

In a first attempt to account for how cycles affect categorization, Kim et al. (2009) made two assumptions. The first was that causal knowledge affects classification in a manner specified by the *dependency model* (Sloman et al., 1998). On this account, features vary in their *conceptual centrality*, such that more central features provide more evidence for category membership. A feature's centrality is a function of its number of (direct and indirect) dependents (i.e., effects). Quantitatively, feature *i*'s centrality c_i can be computed from the iterative equation,

$$c_{i,t+1} = \sum d_{ij} c_{j,t} \qquad (1)$$

where $c_{i,t}$ is *i*'s weight at iteration *t* and d_{ij} is the strength of the causal link between *i* and its dependent *j*. For example, if a category has three features X, Y, and Z, and X causes Y which causes Z, then when $c_{Z,1}$ is initialized to 1 and each causal link has a strength of 2, after two iterations the centralities for X, Y, and Z are 4, 2, and 1. That is, feature X is more important to category membership than Y which is more important than Z. Qualitatively, the dependency model predicts this because X has two dependents (Y and Z), Y has one (Z), and Z has none.

Kim et al.'s second assumption was that people reason with a simplified representation of cycles. Two reasons were provided for this assumption. First, because variables rarely cause each other constantly and simultaneously, it is likely that people assume that they influence each other in discrete time steps. Second, because it is implausible that people represent time steps extending into infinity, only a limited number of steps are likely to be considered. For example, consider the category in Fig. 1A in which feature C causes feature E and features X and Y are related in a causal cycle. Fig. 1B shows the cycle "unfolded" by one time step. The assumption is that in generation 1, X and Y mutually influenced one another, resulting in their states in generation 2 (X_2 and Y_2). Kim et al. proposed that feature importance would correspond to the predictions of the dependency model applied to the unfolded representation in Fig. 1B,

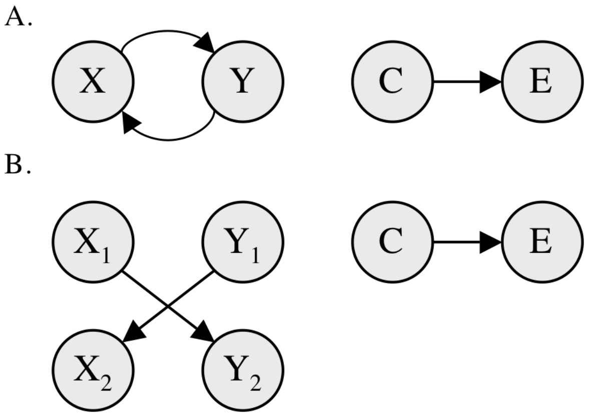

Figure 1. (A) Causal structure tested in Kim et al. (2009). (B) That representation unfolded one time step.

where the centralities of X and Y corresponded to their first generation instantiations (X_1 and Y_1). For the unfolded representation, features X_1, Y_1, and C are equally central because they each have one dependent (X_2, Y_2, and E, respectively) and more central than E, which has zero.[1]

To test this prediction, Kim et al. instructed subjects on artificial categories. For example, subjects learned about a mental disorder called *hadronuria* with four symptoms (e.g., *easily fatigued*, *lack of empathy, depersonalization,* etc.) that caused each other as in Fig. 1A (e.g., being easily fatigued tends to cause a lack of empathy). Subjects were then presented with test items described as having all category features except one and asked to rate the likelihood that it was a category member. The categories and the exact wording of the classification test were varied over five experiments. The results, shown in Fig. 2, confirmed the predictions. The test item missing only feature E was rated higher than the one missing C, suggesting that E was less important to category membership than C. (The phenomenon in which "more causal" features are more important to category membership is referred to as the *causal status effect*, Ahn et al., 2000). And, the ratings of the test item missing C did not differ from those missing only feature X or only Y. In another experiment, Kim et al. compared two-feature causal cycles with more complicated acyclic structures and found evidence they interpreted as consistent with their model.

Questions About the Model

The empirical results of Kim et al. (2009) are important insofar as they provide an initial assessment of how causal cycles affect classification. Moreover, their model is the first to address the difficult problem of how people represent and reason with causal cycles. Virtually all attempts to address cycles involve "unfolding" them in some manner (e.g., Spirtes, 1993), and our own model below will also incorporate this insight. Nevertheless, as it stands, the Kim et al. model faces a number of difficulties.

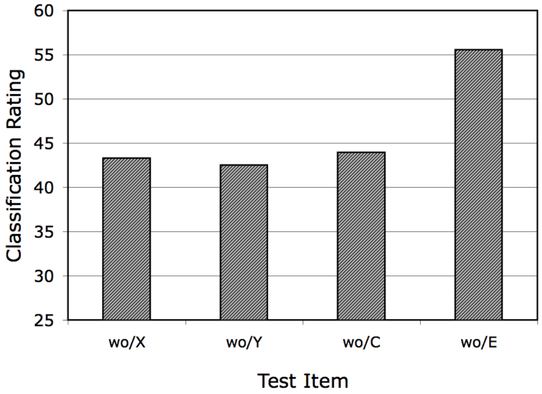

Figure 2. Classification results from Experiments 2-6 of Kim et al. (2009).

The first derives from the assumption of the dependency model that causal knowledge only affects the importance of *individual features*. In contrast, previous research has shown it has a larger effect on the *combinations* of features that are acceptable to category membership. For example, Rehder (2003; Rehder & Kim, 2006; 2010) have demonstrated *coherence effects* in which good category members are those that exhibit the pattern of interfeature correlations one expects to be generated by causal relations.

Coherence effects are likely to have contributed to the Kim et al. results in Fig. 2. For example, a test item missing only feature X was likely given a low classification rating not just because of the importance of feature X but also because its absence violates two causal relations: X is absent despite the presence of Y and vice versa. Moreover, because they differ in how many causal links they violate, differences in the ratings of the four test items may reflect not only the importance of the features themselves but differences in the items' coherence.

A second difficulty concerns how feature importance varies with the strengths of the causal links. According to the dependency model, a feature's centrality increases not only with its number of dependents but also with the strength of the links with those dependents (the *d*s in Eq. 1). For example, as the causal link between feature C and E in Fig. 1 grows stronger, so too should the causal status effect (the centrality of C relative to E). However, research has shown that the causal status effect grows *smaller* as the strength of the causal links grows larger (Rehder & Kim, 2010).

These predictions are now tested in two new experimental tests of causal cycles. Following Kim et al., subjects were instructed on a novel category with interfeature causal relations as in Fig. 1A. But unlike Kim et al. we assess whether coherence effects obtain with causal cycles by testing participants not only on test items missing a single feature but also those missing two features (one missing both X and Y and one missing both C and E) as well as the category "prototype" (all four features present). In addition, across two experiments we manipulate the strength of the causal links. In Expt. 1, we instruct subjects that the causal relations are probabilistic by telling them that a cause produces

[1] Note that dependency model's original formulation makes it technically inapplicable to certain causal networks, including the one in Fig. 2B. However, Kim et al. proposed new variants of the dependency model (e.g., *alpha centralities*) that address these issues Nevertheless, these variants inherent the same qualitative properties (and problems) as their predecessor (see below).

its effect 75% of the time. In Expt. 2, the causal relations are described as deterministic (100%) instead. We assess how the relative importance of individual features (e.g., C vs. E vs. the cycle features) varies with causal link strength.

Experiments 1 and 2

Method

Materials. Three novel categories each with four features were tested: Myastars (a type of star), Romanian Rogos (a type of automobile), and Lake Victoria Shrimp. Each typical feature was described as occurring in most category members whereas the opposite, non-typical value was described as occurring in some category members. For example, participants were told "Most Myastars have high density whereas some have low density."

Participants learned three causal links arranged as in Fig. 1. Each link specified the cause and effect (e.g., "Ionized helium causes the star to be very hot.") and some detail regarding the causal mechanism (e.g., "Ionized helium participates in nuclear reactions that release more energy than the nuclear reactions of normal hydrogen-based stars, and the star is hotter as a result."). In addition, a sentence described the strength of the causal link (e.g., "Whenever a Myastar has high density, it will cause that star to have a large number of planets with probability X%." where X was 75 and 100 in Expts. 1 and 2, respectively). Features 1-4 of each category were assigned to the causal roles shown in Fig. 1 in four different ways (XYCE, XYEC, CEXY, and ECXY), so that each feature appeared in each role an equal number of times across subjects.

Procedure. Participants first studied several screens of information about the category. Initial screens presented the category's cover story and which features occurred in "most" versus "some" category members. Features were presented in one order for half the subjects and in the reverse order for the other half. The fourth screen presented the three causal links, including the mechanism and causal strength information. The fifth screen was a diagram similar to the one in Fig. 1. Subjects were required to pass a multiple-choice test to ensure they learned this information.

Participants then rated the category membership of seven test items. On each trial, the four dimension values were listed in one of the two counterbalanced orders. Responses were entered by positioning a slider on a scale with end labeled “Sure that it isn't” (a category member) and the other end “Sure that it is.” The position of the slider was encoded as a number in the range 0-100. Each test item was presented twice in separate blocks and the order of the trials within a block was randomized for each participant.

Participants. There were three between-subject factors: the four assignments of features to roles X, Y, C, and E, the two feature presentation orders, and which of the three categories was learned. Subjects were randomly assigned to these 4 x 2 x 3 = 24 between-participant cells subject to the constraint that an equal number appeared in each cell. 96 New York University undergraduates, split evenly between Expts. 1 and 2, received course credit for participating.

Results

Initial analyses revealed no effects of which category subjects learned, the assignment of features to causal roles, or feature presentation order in either experiment, and so the classification results are presented in Fig. 3A (Expt. 1) and Fig. 3B (Expt. 2) collapsed over these factors.

Expt. 1 results: Probabilistic links. As expected, the prototype item received a high rating (95.3) indicating that it was viewed as a very likely category member. In addition, the ratings of the four test items with one missing feature (light gray bars in Fig. 3) were similar to those of Kim et al. (2009) shown in Fig. 2: The item missing only the effect feature E (the “wo/E” item) was rated higher than the one missing only the cause C (“wo/C”), which in turn was rated about the same as those missing one of the cycle features.

As expected, the two test items missing two features (either missing both X and Y, or both C and E) were rated lower (69.4) than the prototype. Importantly, however, they were rated higher than the items missing one feature (33.9). This result reflects an effect of coherence: Although they have only two typical features, these items are consistent with the causal relations (i.e., causally-related features are either both present or both absent). In contrast, items with only one missing feature violate the causal relations and they receive a lower rating as a result.

A one-way ANOVA revealed an effect of test item, $F(6, 282) = 67.71$, $MSE = 448.5$, $p < .0001$. The item missing E was rated significantly higher than the one missing C (36.9 vs. 31.7), $t(47) = 3.64$, $p < .001$, which in turn did not differ from the cycle features, $t(47) = 1.28$, $p > .20$. Items missing one feature were rated lower than those missing two, $t(47) = 6.07$, $p < .0001$, which in turn was rated lower than the prototype, $t(47) = 7.32$, $p < .0001$.

Expt. 2: Deterministic links. Fig. 3B shows that in Expt. 2 the prototype again received the highest rating. However, the use of deterministic causal relations in Expt. 2 resulted in a different pattern of ratings among the test items missing a single feature. Whereas in Expt. 1 the item missing C was rated lower than the one missing E (a causal status effect) and the same as those missing X or Y, in Expt. 2 it was rated *higher* than both the wo/E item (i.e., a *negative* causal status effect) and those missing a cycle feature. The latter result contradicts the Kim et al. model's central claim that the cause feature C and the cycles features X and Y should be equally important to category membership (because both have one dependent; see Fig. 2B).

As in Expt. 1, the two items that were missing two features but maintained coherence were rated much higher than the four items missing just one feature (66.6 vs. 21.1).

An ANOVA revealed an effect of test item, $F(6, 282) = 97.80$, $MSE = 491.4$, $p < .0001$. The item missing feature C was rated higher than the one missing E, $t(47) = 3.92$, $p < .001$, which in turn did not differ from X and Y, $t < 1$. In addition, features missing one feature were rated lower than those missing two, $t(47) = 8.41$, $p < .0001$, which in turn were rated lower than the prototype, $t(47) = 7.76$, $p < .0001$.

Fitting the Kim et al. cycles model. We also quantitatively fit the variant of the Kim et al. model that computes alpha centrality (see Footnote 1) to the classification ratings from Expts. 1 and 2. Specifically, the ratings were predicted according to the formula,

$$rating(t_i) = \beta_0 + \left[\sum_j \left(c_{\alpha,k}\left(t_{i,j}; d_{75}, d_{100}, \alpha\right) f_j\right)\right]^{\gamma} \quad (1)$$

where t_i is a test item, d_{75} and d_{100} are the causal link strengths in Expts. 1 and 2, respectively, and α represents a feature's starting centrality. $c_{\alpha,k}$ returns feature j's alpha centrality in category k (i.e., the dependency network in Fig. 1B) and f_j codes whether j is present (1) or absent (0) in t_i. Finally, the purpose of β_0 and γ is to map the model's predictions onto the rating scale: γ scales those predictions according to an arbitrary power function; β_0 then translates the result. The parameters were unconstrained except that $d_{100} \geq d_{75}$ representing the stronger causal links in Expt. 2 vs. 1.

The model was simultaneously fit to the 14 classification ratings from Expts. 1 and 2 by identifying values for parameters d_{75}, d_{100}, α, β_0, and γ that minimized squared error. The best fitting parameters were $d_{75} = d_{100} = 0$, $\beta_0 = 36.8$, $\gamma = 2.88$ (when the ds = 0, parameter α has no effect on the predictions). The predicted ratings are presented in Fig. 3 superimposed on the data.

Fig. 3 illustrates the two difficulties with the Kim et al. model we identified earlier. First, because the model predicts that an object's degree of category membership is the sum of its features weighted by their centrality, it is constrained to predict a lower rating for the items missing two features (wo/X&Y and wo/C&E) than those missing one (e.g., wo/X). That is, the model is unable to account for the coherence effect found in both Expts. 1 and 2.

Second, the model is unable to predict the positive status effect found in Expt. 1 and the negative one found in Expt. 2. This is so because the model predicts that that effect should be larger for stronger causal links. Because this pattern is opposite than the one observed, the best model fit compromises by yielding a causal status effect (the difference between the wo/C and wo/E item) of *zero* in both experiments (produced by values of 0 for the causal strength parameters d_{75} and d_{100}). Finally, the model is unable to account for the higher rating given to the wo/C item relative the items missing a cycle feature. In summary, Fig. 3 illustrates how the Kim et al. model is unable to reproduce the key qualitative results from these experiments.

A Generative Model of Causal Cycles

We now present our own model of causal cycles based on the *generative model* proposed by Rehder and colleagues (Rehder, 2003; Rehder & Kim, 2006) and Directed Cyclic Graphs (DCG) proposed by Spirtes (1993). The generative model assumes that interfeature causal relations are represented as probabilistic causal mechanisms and that classifiers consider whether an object is likely to have been produced or generated by those mechanisms. Objects likely to have been generated are considered to be good category members and those unlikely to be generated are poor ones.

One advantage of the generative model is that it provides an account of coherence effects: A population of category members should exhibit the expected pattern of correlations between causally related features. Thus a likely category member is one that maintains those correlations.

In addition, Rehder & Kim (2010) showed that the generative model predicts the observed sensitivity of the causal status effect to causal strength. Specifically, if a causal link is deterministic, then an effect should be at least as prevalent among category members as its cause (and more prevalent when the effect has additional causes), and thus the cause can be less important to category membership decisions than the effect. In contrast, if a causal link is probabilistic, the effect can be less prevalent than the cause, in which case it should have less weight on classification decisions.

We now present an extension to the generative model that addresses causal cycles. Importantly, this proposal builds on the basic insight provided by Kim et al. (2009) regarding the "unfolding" of causal cycles one time step. However, our account will enjoy the advantages of the generative model, specifically a DCG, including an account of coherence effects and correct predictions regarding the strengths of causal links.

The generative model assumes that a category's causal knowledge is represented as a type of parameterized CGM. For example, a CGM associated with the category in Fig.

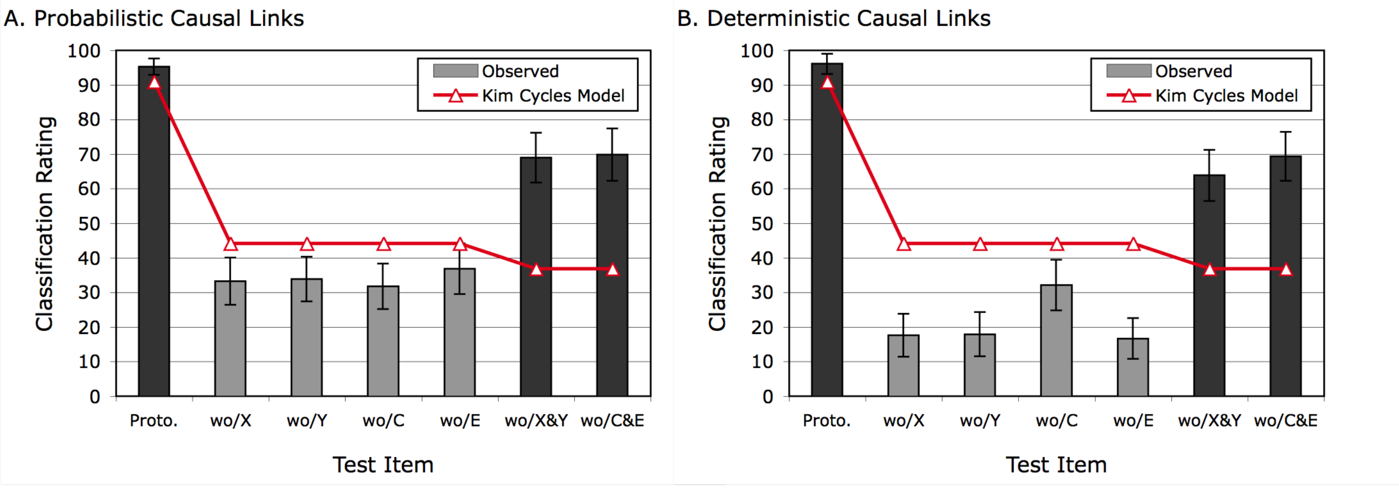

Figure 3. Classification ratings from (A) Experiment 1 (probabilistic causal links) and (B) Experiment 2 (deterministic causal links). Fits of the Kim et al. model are superimposed on the data. Error bars are 95% confidence intervals.

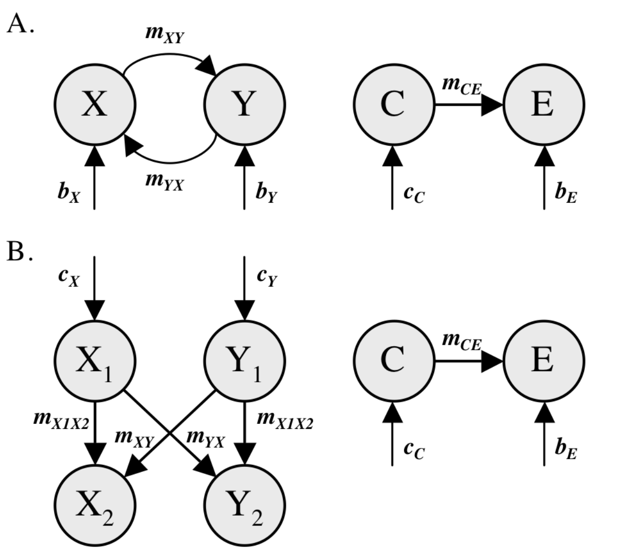

Figure 4. (A) An improper CGM of the category in Figure 1A. (B) That CGM unfolded one time step.

1A is presented in Fig. 4A. The causal mechanism between feature C and E is assumed to operate (i.e., to bring about E) with probability m_{CE} when C is present and any other potential background causes of E collectively operate with probability b_E. C and E's background causes are assumed to form a "fuzzy-or" network that together produce E in members of category k conditioned on C with probability,

$$p_k(E=1 \mid C=1) = m_{CE} + b_E - m_{CE}b_E \qquad (2)$$

When C is absent it has no effect on E.

$$p_k(E=1 \mid C=0) = b_E \qquad (3)$$

The probability of the root cause C is a free parameter c_C.

As mentioned however, graphs with cycles are not proper CGMs because the standard inferential procedures that accompany CGMs are undefined. Accordingly, we work instead with the unfolded representation in Fig. 4B. In this representation, the state of variable X_2 is a fuzzy-or function of X_1 and Y_1,

$$p_k(X_2=1 \mid X_1=1,Y_1=1) = m_{X_1X_2} + m_{YX} - m_{X_1X_2}m_{YX} \qquad (4)$$

$$p_k(X_2=1 \mid X_1=0,Y_1=1) = m_{YX} \qquad (5)$$

$$p_k(X_2=1 \mid X_1=1,Y_1=0) = m_{X_1X_2} \qquad (6)$$

$$p_k(X_2=1 \mid X_1=0,Y_1=0) = 0 \qquad (7)$$

An analogous four equations specify how Y_2 is a fuzzy-or function of X_1 and Y_1.

To represent that a variable present in generation 1 is guaranteed to be present in generation 2, we assume,

$$m_{X_1X_2} = m_{Y_1Y_2} = 1 \qquad (8)$$

Finally, we assume no information is available concerning the presence of root causes X_1 and Y_1, that is, $c_X = c_Y = .5$. That is, we apply the principle of indifference, or in Bayesian terms, a non-informative prior.

According to the generative model, the probability that an object t is a member of category k, $p_k(t)$, is given by joint probability over its observed features X_2, Y_2, C, and E. For the model in Fig. 4B,

$$p_k(t) = p_k(X_2)p_k(Y_2)p_k(E \mid C)p_k(C) \qquad (9)$$

where $p_k(X_2)$ and $p_k(Y_2)$ are computed by summing over the possible states of the unobserved variables X_1 and Y_1,

$$p_k(X_2) = \sum_{X_1=0,1} \sum_{Y_1=0,1} p_k(X_2 \mid X_1Y_1)p_k(X_1)p_k(Y_1)$$

$$p_k(Y_2) = \sum_{X_1=0,1} \sum_{Y_1=0,1} p_k(Y_2 \mid X_1Y_1)p_k(X_1)p_k(Y_1)$$

where $p_k(X_1 = 1) = c_X$ and $p_k(Y_1 = 1) = c_Y$.

Fitting the generative model. We now fit this model to the results of Expts. 1 and 2 in a manner analogous to the Kim et al. cycles model. Specifically, the observed ratings are predicted according to the formula,

$$rating(t_i)/100 = \beta_0 + p_k(t_i; c_C, b_E)^{\gamma} \qquad (10)$$

The causal strength parameters m_{XY}, m_{XY}, and m_{CE} were set to .75 for Expt. 1 and 1.0 for Expt. 2. The model was fit to the results of Expts. 1 and 2 by identifying values for parameters c_C, b_E, β_0, and γ that minimized squared error. c_C was constrained to the range [.50, 1]; b_E was constrained to [0, 1]. The best fitting parameters were c_C = .70, b_E = .07, β_0 = .18, and γ = .45. The predicted ratings are presented in Fig. 5 superimposed on the empirical ratings.

As is apparent from the figure, the generative model suc-

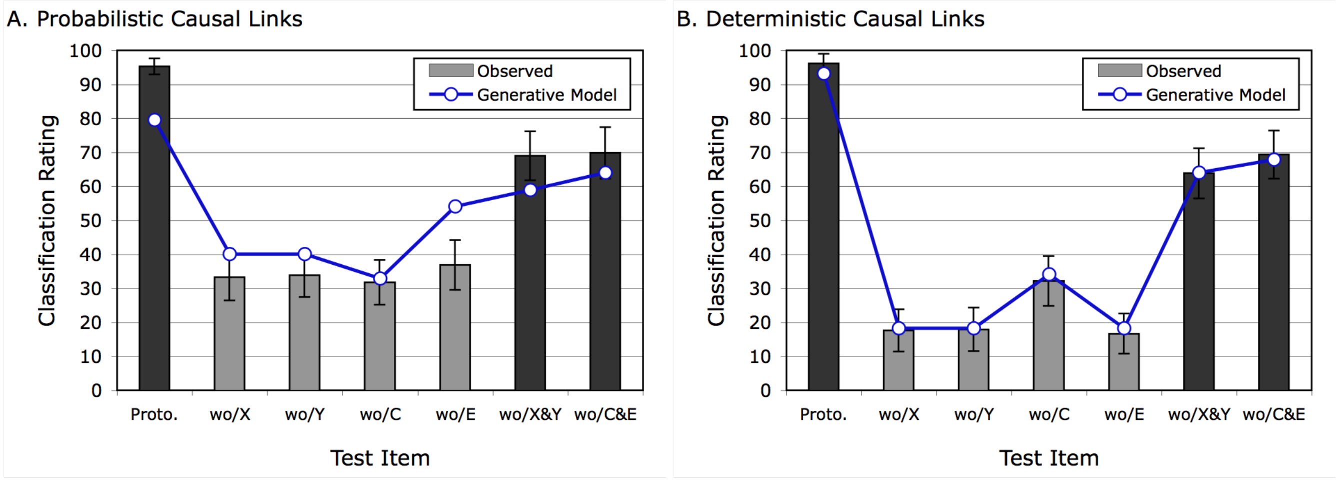

Figure 5. Classification ratings from (A) Experiment 1 (probabilistic causal links) and (B) Experiment 2 (deterministic causal links). Fits of the generative model are superimposed on the data. Error bars are 95% confidence intervals.

ceeds in reproducing the two key phenomena from this study. First, it exhibits coherence effects: The items missing two features have a higher classification probability than those missing one. This is the case because, even though the item missing both X and Y, and the one missing both C and E, have fewer features, they are coherent in light the category's causal relations.

Second, the model reproduces the sign of the causal status effect in the two experiments. Because the causal links are probabilistic in Expt. 1, the effect is less probable then the cause (see Eq. 2), and so an item missing the effect is more probable then one missing the cause. Because the links are deterministic in Expt. 2 (and E has some alternative causes, represented by $b_E = .07$), the effect is *more* probable than the cause, and so an item missing the effect is *less* probable then one missing the cause.

Because it accounts for the effects of both coherence and causal status, the correlation between the model's predicted ratings and the observed ones was .97. Because it accounts for neither of these effects, the corresponding correlation for the Kim et al. model was .58. The better fit of the generative models was accomplished with fewer free parameters (4) as compared to the dependency model (5).

Nonetheless, Fig. 5A reveals a few discrepancies between the predicted and observed ratings. Whereas the fit to Expt. 2 is extremely good ($R > .99$), for Expt. 1 the model overpredicts the wo/E test items (i.e., it predicts that the causal status effect should be larger than it is) and underpredicts the prototype. Further analysis suggests that this may have been due to subjects treating the strength of the causal links in Expt. 1 as > .75. For example, a value of .90 for the m parameters in Expt. 1 yields a better fit ($R > .99$). Determining whether the discrepancies in Fig. 5A reflect a fundamental difficulty for the model or subjects' difficulty in representing probabilities veridically awaits further research.

General Discussion

We have presented a new model of how people represent causal cycles and applied that model to classification data. Recall that CGMs are used extensively throughout cognition research but are unable to represent cycles among variables. We have built on the insight provided by Kim et al. (2009) that, in people's minds, cycles may be broken by unfolding them. Our model, however, is a type of generative model that has been shown to exhibit a number of other favorable qualities, such as accounting for coherence and causal status effects. By applying the generative model to an "unfolded" representation of cycles, we preserve these advantages.

This work is at an early stage. At this point our claim is only that our model captures the qualitative trends in human classification judgments when causal cycles are present. Additional work will need to test our model's predictions with other causal structures, including cyclic and non-cyclic structures with more than just two features.

A perhaps more fundamental issue concerns the number of time steps a causal cycle is unfolded. Both Kim et al. and we assume one time step, but one might question whether people's representation of cycles is so simplified. An alternative would be to assume that cycles are unfolded through a large number of time steps but that causal strengths get weaker with each subsequent step, allowing the model to converge to a steady state. Tests of these and other possibilities await additional theoretical and empirical work.

Acknowledgments

This work was supported by the Air Force Office of Scientific Research, Grant No. FA9550-09-NL404.

References

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology, 41*, 361-416.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., & Kushnir, T. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 3-23.

Griffiths, T. L., & Tenebaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 334-384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction *Psychological Review, 116*, 56.

Holyoak, K. J., Lee, J. S., & Lu, H. (2010). Analogical and category-based inferences: A theoretical integration with Bayesian causal models. *JEP: General, 139*, 702-727.

Kim, N. S., & Ahn, N. S. (2002). Clinical psychologists' theory-based representation of mental disorders affect their diagnostic reasoning and memory. *JEP: General, 131*, 451-476.

Kim, N. S., Luhmann, C. C., Pierce, M. L., & Ryan, M. M. (2009). Causal cycles in categorization. *Memory & Cognition, 37*, 744-758.

Lee, H. S., & Holyoak, K. J. (2008). The role of causal models in analogical inference. *JEP: LMC, 34*, 1111-1122.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*, 955-984.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *JEP: LMC, 29*, 1141-1159.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of object categories. *Cognitive Psychology, 50*, 264-314.

Rehder, B. & Kim, S. (2010). Causal status and coherence in causal-based categorization. *JEP: LMC, 36,* 1171-1206.

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenebaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition, 109*, 175-192.

Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, 22, 189-228.

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science, 29*, 5-39.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28*, 303 -333.

Spirtes, P. (1993). Directed Cyclic Graphs, Conditional Independence, and Non-Recursive Linear Structural Equation Models. Technical Report CMU-PHIL-35, Dept. of Phil., CMU.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *JEP: LMC, 31*, 216-227.