A Causal-Model Theory of Categorization

Bob Rehder (brehder@uiuc.edu)
Department of Psychology, University of Illinois, Urbana-Champaign, IL 61801

Abstract

In this article I propose that categorization decisions are often made relative to *causal models* of categories that people possess. According to this *causal-model theory of categorization*, evidence of an exemplar's membership in a category consists of the likelihood that such an exemplar can be generated by the category's causal model. *Bayesian networks* are proposed as a representation of these causal models. Causal-model theory was fit to categorization data from a recent study, and yielded better fits than either the prototype model or the exemplar-based context model, by accounting, for example, for the confirmation and violation of causal relationships and the asymmetries inherent in such relationships.

Several investigators have argued that category learning and categorization are strongly influenced by the theoretical, explanatory, and causal knowledge that people bring to bear (Murphy & Medin, 1985; Murphy, 1993; Heit, 1998). For example, manipulations of stimulus materials affect category learning by eliciting different aspects of people's background knowledge (e.g., Pazzani, 1989; Murphy & Allopenna, 1994). Performance on a variety of tasks has been correlated with the amount of relevant domain knowledge individuals possess (Keil, 1989; Medin, Lynch, Coley, & Atran, 1997). However, there has been relatively little development of this "theory-based" view of categories in terms of detailed theory and computational models (c.f. Heit, 1994). This state of affairs arises in part because of the uncertainty surrounding exactly what knowledge participants deploy in an experimental task. A few recent studies have addressed this problem by employing novel domains and teaching participants "background" knowledge as part of the experimental session (e.g., Ahn & Lassaline, 1996; Rehder & Hastie, 1999; Sloman, et al. 1998). For example, Rehder and Hastie taught participants about fictitious categories described as possessing causal relationships between binary-valued category attributes, and manipulated experimentally whether those causal relationships formed a common-cause or a common-effect causal schema (Figure 1). In the common-cause schema, one attribute (A1) is described as causing the three other attributes, whereas in the common-effect schema one attribute (A4) is caused by three other attributes. For example, one of the fictitious categories was Kehoe Ants, a species of ants described as living on an island in the Pacific Ocean, and one of that category's causal relationships was "Blood high in iron sulfate causes a hyperactive immune system. The iron sulfate molecules are detected as foreign by the immune system, and the immune system is highly active as a result." After learning about such categories and their causal relationships participants performed a transfer categorization task. Rehder and Hastie found that the presence of both a cause and its effect in an instance (e.g., an ant with iron sulfate blood and a hyperactive immune system) led to the instance receiving a higher category membership rating compared to control categories with no causal relationships. Because ratings were also higher when both the cause and effect were *absent* (normal blood and normal immune system), and *lower* when either the cause or the effect was present and the other absent (iron sulfate blood and normal immune system, or normal blood and hyperactive immune system), Rehder and Hastie concluded that participants were attending not merely to the presence of the cause/effect configuration, but rather to whether instances *confirmed* or *violated* causal relationships. Category membership ratings also reflected the asymmetries inherent in causal relationships. For example, a distinct characteristic of common-cause causal networks is that the effect attributes (e.g., A2, A3, and A4 in Figure 1) will be correlated, and indeed the categorization ratings of substantial numbers of common-cause participants were sensitive to whether those correlations were preserved or violated.

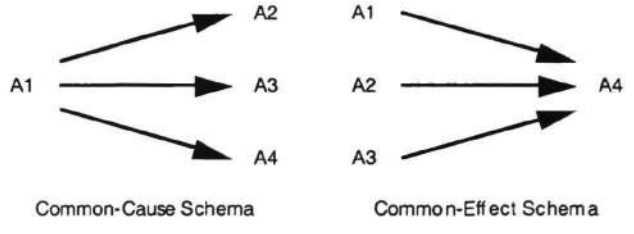

Figure 1

Although these results are suggestive of explicit causal reasoning in participants, it is important to consider whether they can be accounted for by the well-known *similarity-based* categorization models, such as the prototype model and the context model (Medin & Shaffer, 1978; Nosofsky, 1986). Similarity-based models are able to accommodate seemingly disparate categorization strategies by adjusting similarity parameters to differentially shrink or expand the dimensions of the stimulus space. In fact, Rehder and Hastie fitted these models to their transfer categorization data, and found that the models yielded only moderate-quality fits. The fits of instances that possessed many confirmations or many violations of causal relationships were particularly poor.

The failure of the similarity-based models to account for these data leads to a search for alternative categorization models that can account for people's apparent ability to reason causally while categorizing. In this article I propose that categorization decisions are often made relative to *causal models* of categories that people possess, and test *Bayesian networks* as a candidate representation of such models.

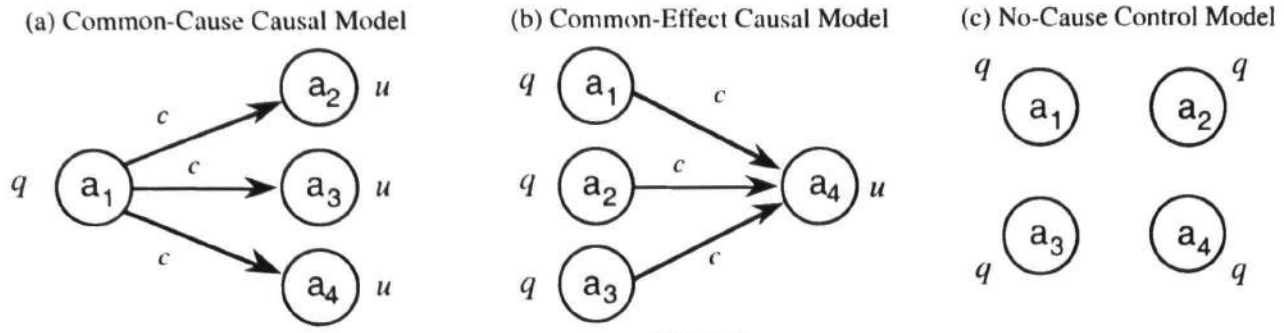

Figure 2

Below I present this *causal-model theory of categorization*, review the procedures employed in the Rehder and Hastie study, and compare data fits produced by the similarity-based models and the causal-model approach. When causal knowledge was present, causal models produced better fits than the similarity-based models by accounting for the categorization ratings of those instances especially affected by the confirmation or violation of causal relationships, and for the asymmetries inherent in causal relationships.

A Causal-Model Theory of Categorization

As in other categorization models, I assume that the classification process consists of both an evidence stage and a decision stage, and that the decision stage is given by a relative-ratio rule (Luce, 1963),

$$P(C_A|t) = E_A(t)/\Sigma_i E_i(t) \quad (1)$$

where $P(C_A|t)$ is the probability that instance t is classified into Category A, i ranges over the set of categories that t may belong to, and $E_i(t)$ is the evidence in favor of t belonging to category C_i. The core of the current proposal is that evidence of t's membership in category C_i consists of the likelihood that t could have been generated by C_i's causal model. Figure 2 presents Bayesian networks that serve as common-cause and common-effect causal models. Bayesian networks are directed acyclic graphs in which nodes represent variables (in this work, binary variables whose values are referred to as "present" and "absent", or "1" and "0"), and edges represent direct dependency relations among variables. In particular, Bayesian networks can be used to represent causal dependencies among variables in which the only direct causes of a variable are its immediate parents (Pearl, 1988). For example, the common-cause network (Figure 2a) has one variable (a_1) that directly causes three other variables (a_2, a_3, a_4). The common-effect network (Figure 2b) has three variables (a_1, a_2, a_3) that each directly causes one other variable (a_4). The no-cause control network (Figure 2c) specifies no causal relationships among variables.

For each variable with no edge into it (i.e., for each "exogenous variable"), Bayesian networks require specification of the probability that its value will be present. These probabilities are referred to as q. In addition, each edge requires the probability that the effect is caused when the cause is present, referred to as c. Finally, u is the probability that a caused (endogenous) variable is present even when its causes are absent. (u can be interpreted as the probability that an effect is brought about by some unspecified cause.) These probabilities can be treated as the parameters of the causal model that the Bayesian networks represents[1].

Given a causal model and its parameters, it is possible to calculate the likelihood that it will generate a particular value on a variable, or a particular set of values. Table 1 presents the equations for computing the likelihood for any settings of the four binary variables in the common-cause, common-effect, and no-cause control models. For example, the probability that the common-cause model will generate the values 1000 (that is, a_1 present, $a_{2\text{-}4}$ absent) is the probability that a_1 is present (q) times the probability that all three causal mechanisms fail to operate ($(1\text{-}c)^3$) times the probability that the effects are not otherwise caused ($(1\text{-}u)^3$).

When the variables of a causal model are interpreted as attributes of a category, and the edges are interpreted as causal relationships between category attributes, the equations of Table 1 can be used to compute the likelihood that the category will generate a particular exemplar, that is, a particular set of attribute values.

The causal-model approach to categorization assumes that categorizers estimate for each candidate category the likelihood that the category generated the exemplar, and then combine this evidence in accord with the relative-ratio rule (Eq. 1) in order to reach a categorization decision.

The Experiments

In addition to Kehoe Ants, Rehder and Hastie (1999) employed five other fictitious categories: one other biological kind, two nonliving natural kinds, and two artifacts. For each of the four binary attributes the base rates of its two values was stated to be 75% and 25%. To relate these attribute values to the networks in Figure 2, the 75% and 25% values are henceforth referred to as "present" and "absent" (or "1" and "0"), respectively. Causal relationships were written such that the "presence" of one attribute caused the presence of another.

[1]The networks shown in Figure 2 exhibit a number of restrictions, any of which can be lifted. First, all causal links are assumed to have the equal strength (i.e., the same c) because in the Rehder and Hastie study all relationships were pretested to equate their plausibility. Second, exogenous variables are assumed to have the same base rate (q), because all participants were explicitly told that attributes exhibited equal base rates (see description of the experiments below). Third, endogenous variables are assumed to have the same u for the same reason. Fourth, in the common-effect network I assume that the probabilities that each cause will bring about its effects are independent, that is, the common-effect causal model is related to a "fuzzy or gate" (Pearl, 1988). These restrictions have the advantage of reducing the number of free parameters, facilitating comparison with similarity-based models (see below).

Table 1

Exemplar	Common Cause Causal Model	Common Effect Causal Model	No-Cause Control Model
0000	$q'u'^3$	q'^3u'	q'^4
0001	$q'u'^2u$	q'^3u	qq'^3
0010	$q'u'^2u$	$qq'^2c'u'$	qq'^3
0100	$q'u'^2u$	$qq'^2c'u'$	qq'^3
1000	$qc'^3u'^3$	$qq'^2c'u'$	qq'^3
0011	$q'u'u^2$	$qq'^2(1-c'u')$	$q^2q'^2$
0101	$q'u'u^2$	$qq'^2(1-c'u')$	$q^2q'^2$
0110	$q'u'u^2$	$q^2q'c'^2u'$	$q^2q'^2$
1001	$qc'^2u'^2(1-c'u')$	$qq'^2(1-c'u')$	$q^2q'^2$
1010	$qc'^2u'^2(1-c'u')$	$q^2q'c'^2u'$	$q^2q'^2$
1100	$qc'^2u'^2(1-c'u')$	$q^2q'c'^2u'$	$q^2q'^2$
0111	$q'u^3$	$q^2q'(1-c'^2u')$	q^3q'
1011	$qc'u'(1-c'u')^2$	$q^2q'(1-c'^2u')$	q^3q'
1101	$qc'u'(1-c'u')^2$	$q^2q'(1-c'^2u')$	q^3q'
1110	$qc'u'(1-c'u')^2$	$q^3c'^3u'$	q^3q'
1111	$q(1-c'u')^3$	$q^3(1-c'^3u')$	q^4

Note. $q'=(1-q)$. $r'=(1-r)$. $s'=(1-s)$.

In each of three experiments, each participant learned of one category, and the category's causal schema was manipulated as a between-subjects factor: Participants were given either a common-cause or a common-effect set of causal relationships, or no causal relationships.

After learning about a category and its causal schema, participants were exposed to exemplars of the target category (e.g., Kehoe Ants) in the guise of a training classification task. Participants classified a series of exemplars into the target category or a contrast category labeled "other", with feedback provided on every trial. In the *Neutral-Data Experiment*, target-category exemplars exhibited no correlations between attributes. That is, the inter-attribute correlations that might be expected from the causal relationships learned by the common-cause or common-effect participants were *not* reflected in the target category exemplars (as a result participants in all conditions observed the same training exemplars). In the *Congruent-Data Experiment* target-category exemplars exhibited the inter-attribute correlations implied by the causal relationships: the common-cause participants saw common-cause correlations and the common-effect participants saw common-effect correlations (and the no-cause control participants saw no inter-attribute correlations). Finally, in the *No-Data Experiment* participants were presented with no training exemplars and hence received no empirical information about the target category. Table 2 presents the number of instances of each exemplar presented to participants as members of the target and contrast categories in the Neutral-Data and Congruent-Data experiments. Note that the target category samples exhibited the 75%/25% attribute base rates that participants were told the category possessed. The attribute base rates in the contrast category samples were 25%/75%.

All three experiments concluded with participants performing three transfer tasks: a categorization task, a similarity rating task, and a property induction task. (The similarity and induction results are not discussed further.) During the categorization task, participants rated on a 100-point scale the category membership of 32 exemplars, consisting of all possible 16 examples that could be formed from four binary attributes, each presented twice. No feedback was provided.

Table 2

Exemplar	Neutral-Data Experiment Target	Neutral-Data Experiment Contrast	Congruent-Data Experiment Target CC	Congruent-Data Experiment Target CE	Congruent-Data Experiment Target Control	Congruent-Data Experiment Contrast
0000	0	11	4	1	0	15
0001	0	3	2	0	1	5
0010	0	4	2	1	0	5
0100	0	4	2	1	0	5
1000	0	4	0	1	1	5
0011	1	1	1	1	2	2
0101	1	1	1	1	2	1
0110	1	1	1	2	2	2
1001	1	1	0	1	2	2
1010	1	1	0	2	1	2
1100	1	1	0	2	2	2
0111	4	0	0	5	5	1
1011	4	0	3	5	5	0
1101	4	0	3	5	5	0
1110	3	0	3	2	5	1
1111	11	0	26	18	15	0

Note. CC=Common-Cause Schema. CE=Common-Effect Schema. Control=No-Cause Control Schema.

216, 234, and 180 University of Colorado undergraduates participated in the Neutral-Data, No-Data, and Congruent-Data experiments, respectively. Average category membership ratings for each exemplar in each condition of the Neutral-Data, No-Data, and Congruent-Data experiments are presented in the Appendix.

Model Fitting Procedure

To fit the transfer categorization data from Rehder and Hastie (1999), causal models must be assumed for both the target category (e.g., Kehoe Ants) and the contrast category. It was assumed that for the target category participants employed a causal model appropriate to the causal schema they learned: a common-cause model for the common-cause schema, a common-effect model for the common-effect schema, and a no-cause control model when they learned of no causal relationships. It was also assumed that all participants employed a no-cause control model for the contrast category, because no causal knowledge was provided about that category. As a result, fits of data from the common-cause and the common-effect schema conditions involve four parameters (q, c, and u for the target category, and q for the contrast category), and fits to the no-cause control condition involve two (one q each for the target and contrast category). Data fitting involved finding the set of parameters that minimized squared error, with predictions being given by the equations in Table 1.

The transfer categorization data were also fitted to the prototype and context model (Nosofsky, 1992). The prototype model assumes that evidence of category membership consists of the (additive) similarity of a stimulus to the category's prototype. In fitting the prototype

Table 3

Model Parameters	Common Cause Schema: No Data	Common Cause Schema: Neutral Data	Common Cause Schema: Congruent Data	Common Effect Schema: No Data	Common Effect Schema: Neutral Data	Common Effect Schema: Congruent Data	No-Cause Control Schema: No Data	No-Cause Control Schema: Neutral Data	No-Cause Control Schema: Congruent Data
Prototype Model									
s_1	0.03	0.00	0.00	0.25	0.04	0.11	0.10	0.00	0.00
s_2	0.57	0.00	0.19	0.41	0.00	0.21	0.32	0.00	0.10
s_3	0.53	0.01	0.28	0.53	0.05	0.25	0.33	0.00	0.08
s_4	0.57	0.00	0.33	0.16	0.00	0.00	0.35	0.00	0.05
RMSE	0.097	0.080	0.071	0.077	0.090	0.078	0.023	0.071	0.033
Context Model									
s_1	–	0.21	0.29	–	0.36	0.37	–	0.22	0.26
s_2	–	0.31	0.44	–	0.33	0.42	–	0.29	0.33
s_3	–	0.32	0.44	–	0.35	0.41	–	0.27	0.29
s_4	–	0.34	0.52	–	0.26	0.24	–	0.30	0.37
RMSE	–	0.046	0.045	–	0.067	0.051	–	0.029	0.025
Causal Models									
	Common Cause Causal Model			Common Effect Causal Model			No-Cause Control Model		
Target Category									
q	0.52	0.74	0.63	0.47	0.65	0.60	0.56	0.66	0.63
c	0.30	0.31	0.29	0.22	0.40	0.35	–	–	–
u	0.36	0.55	0.47	0.31	0.38	0.39	–	–	–
Contrast Category									
q	0.37	0.30	0.39	0.35	0.32	0.38	0.41	0.31	0.36
RMSE	0.039	0.025	0.038	0.037	0.018	0.024	0.030	0.025	0.030

model, the prototypes of the target and contrast categories were assumed to be 1111 and 0000, respectively. The context model assumes that evidence of category membership is the total (multiplicative) similarity of a stimulus to all stored category exemplars. In fitting the context model, it was assumed that all exemplars presented during training classification were stored in memory with their category membership. Because of the absence of training exemplars in the No-Data experiment, the context model was not fit to that experiment's data. Both the prototype and context model produce four parameters (s_1, s_2, s_3, and s_4) in the range [0,1] representing the saliency of each dimension (where smaller numbers mean more salient).

Although the relative-ratio rule (Eq. 1) is typically used to predict the probability of category membership, in the Rehder and Hastie study participants produced continuously-valued category membership ratings rather than binary-valued categorization decisions. Accordingly, the rule is used to predict category membership ratings, which are divided by 100 to bring them into the range [0-1].

Results

Fits of the prototype model, the context model, and causal-model theory to the Rehder and Hastie (1999) transfer categorization data are presented in Table 3 as a function of causal schema and experiment. In all common-cause and common-effect conditions, causal models produced better fits than either the prototype model or the context model in terms of residual error variance (RMSE), and did so with the same number of free parameters as the similarity-based models (four). In the no-cause control experimental conditions, participants were not instructed on the presence of inter-attribute causal relationships, and the prototype model and the context model produced adequate data fits in those conditions (with one exception: the prototype model fit in the Neutral-Data experiment). However, causal models produced equally good fits in those conditions, and did so with two fewer free parameters[2].

One reason for the poorer fits of the prototype and context models in the common-cause and common-effect conditions is their inability to account for the category membership ratings of those exemplars especially sensitive to the confirmation or violation of causal relationships. To illustrate, Figure 3 presents context model and causal model fits in the Neutral-Data experiment for such exemplars. As is apparent from the figure, in the common-cause condition the context model *underestimates* the category membership ratings of exemplars that possess three confirmations of causal relationships: 1111 (the common-cause, a_1, and its effects all present) and 0000 (common-cause and its effects all absent). It also *overestimates* the ratings of exemplars that possess three violations of causal relationships: 1000 (the common-cause is present but its effects are absent) and 0111 (the common-cause absent, but its effects are present). Analogously, in the common-effect condition the context model underestimates the ratings of exemplars that possess three confirmations of causal relationships (1111 and 0000), and overestimates the ratings of exemplars that possess three violations (0001 and 1110).

In comparison, in both the common-cause and common-effect conditions causal model theory yielded quite good fits of these exemplars (Figure 3). The causal models' parameter

[2] Participants in the No-Cause Control condition of the No-Data Experiment exhibited a substantial response bias in favor of the target category, and hence data fits in that condition include the addition of a bias parameter b=.58 (b=.50 means zero bias).

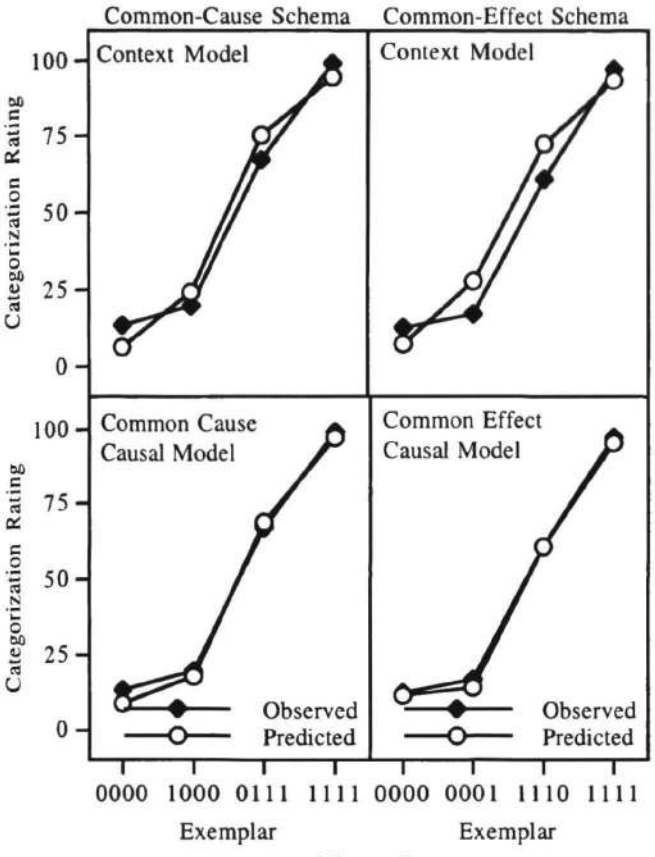

Figure 3

values support the view that their superior data fits were due to participants' use of causal relationships when generating categorization ratings. For example, parameter c, which reflects the strength of the causal relationship between a cause and an effect attribute, was large and positive in all common-cause and common-effect schema conditions (Table 3). One effect of those parameter values is to make the generation of exemplars that confirm causal relationships (1111 and 0000 for both the common-cause and common-effect models) more likely, and the generation of exemplars that violate causal relationships (1000 and 0111 for the common-cause model, and 0001 and 1110 for the common-effect model) less likely. Because the likelihood of generation controls category membership ratings (Eq. 1), causal models account for the sensitivity to the correlations between causally-connected features shown in Figure 3.

The context model can also exhibit sensitivity to inter-attribute correlations, but only when category exemplars are present in memory that manifest those correlations. Such exemplars were absent in both the Neutral-Data and No-Data experiments, and hence participants' sensitivity to correlations between causally-connected features in those experiments must be attributed to the causal knowledge they learned. However, exemplars that manifested the appropriate correlations were present in the Congruent-Data experiment and hence that experiment established the most favorable conditions for the context model. Nevertheless, even in the Congruent-Data experiment causal model theory yielded the best fits, apparently because participants weighed the critical inter-attribute correlations more heavily than is predicted by the context model's multiplicative-similarity rule.

Discussion

The failure of the prototype and context models to account for the Rehder and Hastie (1999) categorization data implies that participants were not rating the category membership of exemplars on the basis of (only) similarity, a result that led Rehder and Hastie to suggest that their participants were engaging in explicit causal reasoning while categorizing. In this article "causal reasoning" is rendered computationally explicit in the form of causal-model theory. In fact, the good fits of the categorization data produced by causal models, together with the specific parameter values responsible for those fits, support the claim that people can engage in causal reasoning while categorizing when causal knowledge is present that enables that reasoning.

An alternative way to account for the current data in terms of similarity is to argue that the feature space is expanded to include higher-order features encoding the confirmation or violation of causal relationships, and that similarity is computed in that expanded space. However, Rehder and Hastie also collected ratings of the similarity of pairs of category members, and found that such ratings were insensitive to whether exemplars matched on those higher-order features (also see Rehder & Hastie, 1998), suggesting that the presence of causal knowledge did not result in an expansion of the feature space. Such higher-order features also fail to account for the asymmetries inherent in causal relationships. For example, when the direction of causality in the common-cause and common-effect models (Figure 2) is reversed, the result is substantially worse fits of the data of many common-cause and common-effect participants, respectively. In other words, when classifying exemplars many participants did not just evaluate each causal link in isolation but rather considered interactions among links produced by the entire *network*, where the nature of those interactions is determined by the links' direction of causality.

Despite differing assumptions regarding the form of category representations (prototypes, exemplars, rules, etc.), traditional similarity-based models assume that such representations are built with information taken from the *data* people observe (i.e., exemplars). Causal model theory diverges sharply from this approach by assuming instead that category representations are formed from the category *knowledge* people possess. The superior fits of causal model theory reported here for the Neutral- and Congruent-Data experiments reflects the greater importance of category *knowledge* versus category *data* on categorization in those experiments. Causal model theory also applies when no exemplars of the category have been observed at all (e.g., the No-Data experiment), a domain beyond the purview of the traditional models. In contrast to the traditional models, causal model theory is thus applicable to the many real-world categories about which people know far more than they have observed first hand.

Causal models have been implicated in other domains. For example, Glymour (1998) argues that people's ability to estimate the strength of causal influences controlling for other causes (i.e., Cheng's, 1997, causal power theory) is equivalent to estimating the conditional probability associated with an edge in a Bayesian network (in this article, parameter c). Waldmann, Holyoak, and Fratianne

(1995) demonstrated that the speed of category learning depends on the match between the correlational structure of the learning data and the learner's causal model of the category. An important area of development for causal model theory is to specify the learning algorithm by which learners *integrate* their category knowledge (in the form of causal models) with data (i.e., observations of the category).

In this article I have demonstrated how the claim that causal knowledge affects categorization can be formalized as an explicit computational model, how it can be fitted to empirical data, and how it can be rigorously tested against other models. Bayesian networks were utilized as a device with which causal knowledge was represented and evidence in favor of category membership was calculated. Future work may advance causal model theory by specifying the processes by which likelihood functions (e.g., Table 1) are computed (or approximated). The success of parallel network algorithms in implementing complex reasoning processes in other domains (Pearl, 1988, Thagard, 1989) make the prospect for this development promising.

Acknowledgements

This work was conducted while the author was supported by NSF Grant SBR 97-20304.

References

Ahn, W., & Lassaline, M. E. (1996). *Causal structure in categorization: A test of the causal status hypothesis (Part I)*. Unpublished manuscript.

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

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines*, 8, 39-60.

Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1264-1282.

Keil, F C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*. (pp. 103-189). NY:Wiley.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85,207-38.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology*, 32, 49-96.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 904-919.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes*. (pp. 149-167). Hillsdale, NJ: Erlbaum.

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 416-432.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufman.

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

Rehder, B., & Hastie, R. (1998). The differential effects of causes on categorization and similarity. In *The Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 893-898). Madison, WI.

Rehder, B., & Hastie, R. (1999). *The essence of categories: The effect of underlying causal mechanism on induction, categorization, and similarity*. Submitted for publication.

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

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General, 124*, 181-206.

Appendix – Transfer Categorization Ratings from Rehder & Hastie (1999)

	Common Cause Schema			Common Effect Schema			No-Cause Control Schema		
Exemplar	No Data	Neutral Data	Congruent Data	No Data	Neutral Data	Congruent Data	No Data	Neutral Data	Congruent Data
0000	50.2	13.4	27.1	40.7	12.1	18.1	33.1	9.9	8.8
0001	41.4	19.9	32.8	34.6	16.8	24.1	41.4	20.4	26.8
0010	41.7	20.1	33.0	38.9	22.5	30.3	42.8	20.1	28.8
0100	39.9	19.9	36.6	41.1	22.6	30.9	43.4	17.2	23.2
1000	37.5	19.9	29.8	44.4	25.4	32.0	46.7	19.2	29.1
0011	43.0	47.3	43.4	51.9	50.4	49.1	56.5	50.5	46.3
0101	40.9	46.7	44.5	52.5	52.5	48.5	54.1	51.1	49.0
0110	40.7	43.8	47.2	46.2	40.8	38.7	55.2	49.2	47.3
1001	50.8	51.1	48.9	56.5	47.8	52.7	59.7	56.8	53.3
1010	51.4	50.5	50.8	48.6	38.9	40.0	60.3	56.2	49.1
1100	52.4	54.4	50.4	52.4	41.3	44.5	60.4	54.2	55.0
0111	45.5	67.2	50.0	66.1	84.5	76.1	69.6	80.7	72.4
1011	65.3	86.2	67.9	68.1	84.5	78.2	72.8	84.2	79.5
1101	65.2	84.2	70.9	71.0	84.8	76.8	74.6	83.4	74.7
1110	67.2	86.5	70.7	53.1	61.3	49.8	73.4	82.6	76.1
1111	90.0	98.6	97.8	90.0	97.2	92.3	88.6	95.0	92.2