Biases in Refinement of Existing Causal Knowledge

Woo-kyoung Ahn
Department of Psychology
University of Louisville
Louisville, KY 40292
wkahn001@homer.louisville.edu

Raymond J. Mooney
Department of Computer Sciences
University of Texas
Austin, TX 78712
mooney@cs.utexas.edu

Abstract

This study describes a psychological experiment on biases that people exhibit in refining probabilistic causal knowledge. In the experiment, the effect of background knowledge was shown by manipulating the causal structure of prior knowledge provided to the subjects. It was found that later training instances affected the refinement of the background knowledge in different ways depending on the causal model initially given to the subjects. The two biases found in the current experiment are (1) knowledge refinement was conservative in the sense that background knowledge was modified as little as possible to account for the observed data and (2) weakening of an existing causal relationship resulted in automatic strengthening of a related causal relationship.

Introduction

How do people revise their existing knowledge given new observations that do not clearly fit with their initial knowledge? In most natural situations, at least some prior knowledge of relevant causal mechanisms is available to explain external stimuli. In certain cases, some relevant causal knowledge is available but new causal relationships may need to be inferred in order to fully account for the observed data. The current study investigates the specific nature of knowledge refinement as a learner acquires new training instances that cannot be fully explained by existing knowledge.

Need for psychological studies on biases in knowledge refinement

Research in both machine learning and psychology has revealed the important role that prior knowledge plays in learning. Psychological research has demonstrated that subjects' learning is greatly affected by their naive theories and existing domain knowledge (Murphy & Medin, 1985; Ahn, Brewer, & Mooney, 1992; Pazzani, 1991). Depending on the presence or absence of relevant background knowledge, subjects learn different concepts from the same examples (Wisniewski, 1989). Meanwhile, machine learning research has developed algorithms that learn more accurate concepts from fewer examples when given relevant background knowledge in the form of an approximate domain theory (Mooney, 1993; Pazzani, 1991; Pazzani & Kibler, 1992; Towell, Shavlik, & Noordewier, 1990).

Although these studies have clearly demonstrated that prior knowledge greatly affects human learning, most of this research have ignored how prior knowledge is modified by experience. For example, Pazzani (1991) presented empirical results on how prior causal knowledge influenced categorization and also developed a computational model of this process; however, he did not address the issue of how existing causal knowledge itself is affected by conflicting data.

Nonetheless, theories concerning biases of knowledge revision are in great demand in knowledge engineering. It is generally agreed that the primary difficulty with developing robust knowledge-based systems is the knowledge acquisition bottleneck (i.e., the complexity of extracting and encoding the domain knowledge needed to perform the task). Knowledge-based systems are typically developed by first interviewing an expert in order to obtain an initial set of rules. Next, the knowledge base is incrementally improved in a laborious process referred to as knowledge-base refinement. Typically, a set of sample problems is used to detect errors in the knowledge base and corrections are determined during a time-consuming consultation with the expert. Recent research in theory refinement attempts to automate the laborious process of knowledge refinement by using various machine learning techniques to automatically revise an existing, approximate knowledge base to fit a set of empirical data (Ginsberg, Weiss, & Politakis, 1988; Ourston & Mooney, 1990; Towell et al., 1990; Koppel, Feldman, & Segre, 1994).

The current study provides initial data revealing important biases that humans display in revising their existing knowledge. Specifically, it focuses on the revision of probabilistic causal knowledge, in which underlying causes (e.g. diseases) probabilistically manifest certain effects (e.g. symptoms). Such knowledge can be formally represented as a Bayesian network (Pearl, 1988). The question is how the strength of existing causal relationships and the addition of new causal links are

affected by new evidence that is not fully consistent with existing causal knowledge.

Main Claim

Our general predictions are that changes are conservative (i.e., background knowledge is modified as little as possible to account for the observed data) and that changes in the strengths of known causal links are preferred to inferring new causal connections. Therefore, as long as existing knowledge is consistent with the observed data, a new piece of causal knowledge will not be acquired even if this new causal explanation would be more parsimonious.

Ahn, Kalish, Medin, and Gelman (1995) have also demonstrated a similar point using an information-seeking paradigm in causal reasoning. In this study, subjects received event descriptions and were instructed to ask questions in order to explain the events. The subjects tended to seek out information that would provide evidence for or against hypotheses about underlying mechanisms with which they were already familiar. In contrast, previous psychological models on causal attributions have emphasized that the most critical and necessary information in causal attributions is information about covariation between candidate causes and effects (e.g., Cheng & Novick, 1992; Kelley, 1967, 1971). These models tended to focus on the bottom-up processes of acquiring novel causal relationships rather than on the top-down processes of utilizing existing causal knowledge. However, Ahn et. al's results showed that people did not seek out a novel causal relationship between arbitrary factors by relying solely on covariation information. Rather, people attempted to seek out evidence for causal mechanisms with which they were already familiar, a result which supports the idea that people are conservative in learning new causal relationships.

Experiment

The current experiment investigates under what conditions people add new causal connections to prior domain knowledge as opposed to modifying the strength of existing connections.

Methods

Subjects received background knowledge in the form of one of three causal models as shown in Figure 1. In these models, A, B, C and/or D are symptoms caused by two new diseases, X and Y. In all of the conditions, subjects were told that disease X caused symptoms A and B 70% of the time (indicated by solid arrow lines in the figure) and symptom C 20% of the time (indicated by dotted arrow lines in the figure) and that disease Y caused symptoms B and C 70 % of the time and symptom A 20% of the time. The difference between the three causal structures lies in the causal relationship between symptom D and the diseases. In the indirect-cause condition, symptom D was caused by symptom B, and therefore also indirectly caused by diseases X and Y; in the direct-cause condition, symptom D was caused directly by diseases X and Y; and in the no-cause condition, there was no known cause for D. Finally, the subjects were told that both diseases were equally likely to occur a priori.

After learning these causal structures, the subjects judged the likelihood of various and in the no-cause factors given various configurations of other factors. There were six test items; P(X|B), P(Y|B), P(X|B and no D), P(Y|B and no D), P(X|A), and P(Y|C). For example, for P(X|B), the subjects were asked, "What is the probability that a person who exhibits symptom B has disease X? _____%"

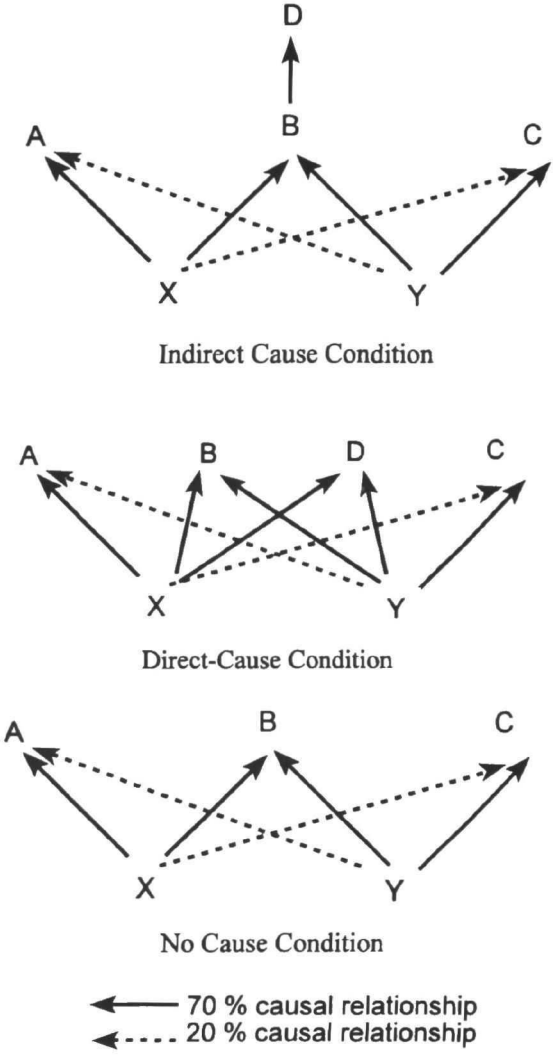

Figure 1. Causal models used in each condition

Then, subjects in all conditions received information on a set of training cases which were described as data gathered afterwards. These training cases were constructed in such a way that symptom D was more associated with disease X than with disease Y. The actual description of the training cases was;

70% of patients with symptom D had disease X; 30% had disease Y.
75% of patients with symptoms D and B had disease X; 25% had disease Y.
60% of patients with symptoms D and C had disease X; 40% had disease Y.
20% of patients with symptom C but **not** D had disease X, 80% had disease Y.

After that, the subjects judged the likelihood of the test items again. While making these judgments, the subjects were provided with the figure of the initial causal structure and the description of the new training cases. Therefore, in making all these judgments, there was no demand on memory.

Prediction

The basic prediction was that subjects in the indirect-cause condition would not construct a new causal link betw X and D because D could already be explained as an indi effect of X through B. On the other hand, subjects in no-cause condition would infer a new direct causal connec between X and D in order to account for the data. 7 difference would be reflected in the increase in the estimat P(X|D, not B); More specifically, the increase should be in the indirect-cause condition than the no-cause condit since the absence of B would indicate that D did not cc about as a side-effect of X causing B; whereas the absenc B would not affect a direct connection between X and D the no-cause condition.

As in the indirect-cause condition, the subjects in direct condition would not need to construct a new l between X and D because this relationship is alre explained by the existing causal link. Therefore, the incre in the estimate of P(X|D, not B) should be similar in indirect-cause and the direct-cause condition. In addition the no-cause condition had acquired a direct causal l between X and D through the training instances, their sec rating on P(X|D, not B) should be similar to the second rat of the direct-cause condition. As a result, the direct-ca condition serves as a baseline group for the other t conditions.

Table 1. Results of the No-cause condition

	P(X|B)	P(Y|B)	P(X|D,noB)	P(Y|D,noB)	P(X|A)	P(Y|C)
Test1	55.3	55.3	19.4	19.4	62.4	60.6
Test2	65.0	42.4	58.5	37.9	61.2	65.6
Test1 - Test2	9.7	-12.9	39.1	18.5	-1.18	5.0

Table 2. Results of the Indirect-cause condition

	P(X|B)	P(Y|B)	P(X|D,noB)	P(Y|D,noB)	P(X|A)	P(Y|C)
Test1	57.4	55.0	37.9	33.2	64.7	65.3
Test2	68.4	41.4	41.9	29.4	70.0	59.1
Test1 - Test2	11.1	-13.7	3.9	-3.8	5.3	-6.2

Table 3. Results of the Direct-cause condition

	P(X|B)	P(Y|B)	P(X|D,noB)	P(Y|D,noB)	P(X|A)	P(Y|C)
Test1	57.7	53.5	45.8	44.6	65.0	50.2
Test2	58.2	42.9	53.7	36.9	58.8	71.2
Test1 - Test2	0.5	-10.5	7.9	-7.7	-6.2	20.92

Results and Discussion

Tables 1-3 summarize the results. Each table shows mean probability judgments by each condition. In each condition, the second row indicates mean ratings on the first test, the third row indicates the mean ratings on the second test after the training instances, and the fourth row indicates the differences between the second and the first tests.

For each test item, an ANOVA was conducted with the condition as a between-subject variable and the two tests as a within-subject variable. The focus of the current study is the interaction between the increase of the probability estimates and the three conditions; that is, does the increase or decrease of the causal strength changes as a function of existing causal knowledge? Only three out of the six test items resulted in a reliable interaction effect at $p < .05$; P(X|D, not B), P(Y|D, not B), and P(Y|C). The following figures illustrate the direction of the interaction effect on these three items.

As can be inferred from the figures, the indirect-cause and the direct-cause condition did not significantly increase or decrease their estimates for P(X|D, no B) and P(Y|D, no B), whereas the no-cause condition significantly increased their estimates. Also, prior to training, the estimate for P(X|D, no B) in the no-cause condition is significantly less than the indirect-cause condition, whereas after training it is significantly greater than the indirect-cause condition. These results indicate that only the no-cause condition established a new causal link between X and D and a somewhat weak, direct link between Y and D. Although the subjects in the indirect-cause condition could have established a new causal link between X and D based on the same training instances, they presumably applied their existing knowledge to account for the association between X and D (i.e., X causes B and therefore causes D). Because of this interpretation of the association between X and D, it mattered much more for the indirect-cause condition not to have symptom B compared to the no-cause condition, since B's absence blocks the known causal path between X and D.

Another interesting but unexpected result came from the changes in the estimate of P(Y|C). The direct-cause condition's estimate significantly increased after the training instances compared to the other conditions. This is an

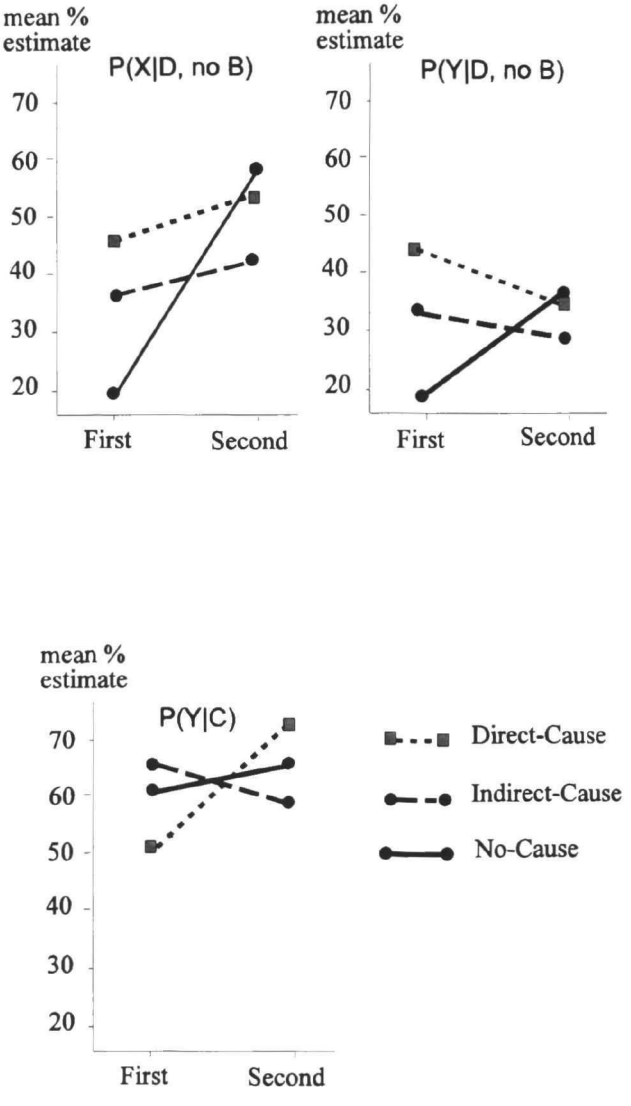

Figure 2. Results from the three conditions

unexpected finding because as far as the link between disease Y and symptom C is concerned, all three conditions were initially the same and they all received exactly the same training instances.

Our interpretation is as follows: Note that in the training instances, the association between symptom D and disease Y is somewhat weakened whereas the association between symptom C and disease Y is somewhat strengthened. In the direct-cause condition, which is the only condition who started out with a direct link between Y and D, this causal link must have been weakened by the training instances. Then, this weakened link might have actually increased diagnostic values of other symptoms. In other words, as one symptom became less diagnostic of disease Y, the other symptom automatically became more diagnostic.

This new phenomenon can be considered a converse of the "discounting effect" or "explaining away" for the revision of causal strengths. According to the discounting effect proposed by Kelley, people tend to discount a candidate cause if we find out that one cause is already responsible for the effect. For example, if Mary finds out that her brother took her radio away, she might wonder whether it was because his radio was broken or he was mad at her. Finding out that her brother's radio was actually broken, she is less likely to believe that her radio was taken because he was mad at her. In Artificial Intelligence, this phenomenon is called "explaining away" and is computationally implemented in Bayesian networks (Pearl, 1988), providing a normative account of this psychological principle.

The current results on P(Y|C) in the direct-cause condition seem the converse of the discounting effect. That is, initially, one starts out with a belief that a cause has two effects. As one of the causes is weakened through later observations, people automatically boost up the strength of the other causal relationship. To give a more real-life example of this phenomenon, suppose one initially believed that having an extra X-chromosome caused a person to have a high-pitched voice and to be agreeable. If the later observations indicated that there was no genetic ground for being agreeable, then having a high-pitched voice would gain a more diagnostic value for the existence of an extra X-chromosome.

Conclusion

The current study had demonstrated two interesting biases in refinement of causal background knowledge as a function of its initial causal structure and training instances. First, the refinement occurred in a conservative manner. People would not construct a new causal link as long as their existing causal knowledge can explain the new training instances even when this causal explanation was less direct and parsimonious. This phenomenon is consistent with processes underlying stereotype formation; Even if there can be many alternative ways of accounting for one's behavior, people would rather apply their existing knowledge than take a new perspective on the observation and learn a new possible causal connection. Second, weakening an existing causal strength might actually strengthen the causal strength of an alternative effect.

In the future, we hope to explore additional biases that people exhibit when revising probabilistic causal knowledge by examining the effect of different types of data on a larger variety of initial causal structures. In addition to inferring the revisions subjects made to their knowledge based on their subsequent judgements, we plan to more directly inquire into the exact changes they make to prior causal knowledge in order to account for conflicting data.

We also hope to develop and test a computational model of revising probabilistic causal knowledge based on revising both the parameters and structure of a Bayesian network to make it consistent with a set of training data. This model will attempt to integrate methods for revising existing causal strengths (Schwalb, 1993) with methods for inducing new causal structures (Cooper & Herskovits, 1992).

References

Ahn, W., Brewer, W. F., & Mooney, R. J. (1992). Schema acquisition from a single example, *Journal of Experimental Psychology: Learning, Memory, & Cognition, 18,* 391-412.

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S.A. 1995). The role of covariation versus mechanism information in causal attribution, *Cognition, 54,* 299-352.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365-382.

Copper, G, G., & Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning, 9,* 309-347

Ginsberg, A., Weiss, S. M., & Politakis. (1988). Automatic knowledge based refinement for classification systems. *Artificial Intelligence, 35,* 197-226.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, pp.192-238). Lincoln: University of Nebraska Press.

Kelley, H. H. (1971). *Attribution in social interaction.* Morristown, NJ: General Learning Press.

Koppel, M., Feldman, R. R., & Segre, A. M. (1994). Bias-driven revision of logical domain theories. *Journal of Artificial Intelligence Research, 1,* 1-50.

Mooney, R. J. (1993). Integrating theory and data in category learning. In G. Nakamura, R. Taraban, & D. L. Medin (Eds.) *Categorization by humans and*

machines: The Psychology of learning and motivation, 29, 189-218, Orlando, FL: Academic Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review., 92*, 289-316.

Pazzani, M. J. (1991). A computational theory of learning causal relationships, *Cognitive Science, 15*, 401-424.

Pazzani, M. J., Kibbler, D. (1992). The utility of background knowledge in inductive learning. *Machine Learning, 9*, 57-94.

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

Schwalb, E. (1993). Compiling Bayesian networks into neural networks. In *Proceedings of the Tenth International Conference on Machine Learning*, 291-297, Amherst, MA.

Towell, G. G., Shavlik, J. W., & Noordewier, M. O. (1990). Refinement of approximate domain theories by knowledge-based artificial neural networks. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, 861-866. Boston, MA.

Wisniewski, E. J. (1989). Learning from examples: The effect of different conceptual roles. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society,* 980-986. Ann Arbor, MI.