Causal Explanations in Counterfactual Reasoning

Morteza Dehghani (morteza@northwestern.edu)
Department of EECS, 2145 Sheridan Rd
Evanston, IL 60208-0834 USA

Rumen Iliev (r-iliev@northwestern.edu)
Department of Psychology, 2029 Sheridan Rd
Evanston, IL 60208-2710 USA

Stefan Kaufmann (kaufmann@northwestern.edu)
Department of Linguistics, 2016 Sheridan Rd
Evanston, IL 60208-4090 USA

Abstract

This paper explores the role of causal explanations in evaluating counterfactual conditionals. In reasoning about what would have been the case if A had been true, the localist injunction to hold constant all the variables that causally influence whether A is true or not, is sometimes unreasonably constraining. We hypothesize that speakers may resolve this tension by including in their deliberations the question of what would *explain* the hypothesized truth of A. To account for our recent psychological findings about counterfactuals, an alternative approach based on Causal Bayesian networks is proposed in which the intervention operator utilizes the agent's beliefs about the explanatory power of the antecedent of the counterfactual. The results of three psychological experiments are reported in which the new method succeeds in predicting subjects' responses while the traditional method for evaluating counterfactuals in Bayesian networks fails.

Keywords: Counterfactual Reasoning; Causal Explanations; Causal Networks.

Introduction

Counterfactual reasoning plays an important role in causal inference, diagnosis, prediction, planning and decision making, as well as emotions like regret and relief, moral and legal judgments, and more. Consequently, it has been a focal point of attention for decades in a variety of disciplines including philosophy, psychology, artificial intelligence, and linguistics. The fundamental problem facing all attempts to model people's intuitive judgments about what would or might have been if some counterfactual premise A had been true, is to understand people's implicit assumptions as to which actual facts to "hold on to" in exploring the range of ways in which A might manifest itself. As Goodman (1955) stated it in his classic example: In asking what would have been the case if the match had been struck, people tend to infer from the presence of oxygen that it would have lit; but why not instead infer that there would have been no oxygen from the fact that it did not light? There is no principled logical difference behind asymmetries of this sort; something else seems to be at work.

Many formal theories of counterfactual reasoning are inspired by the model-theoretic accounts of Stalnaker (1968) and Lewis (1973). Minor differences aside, both crucially rely on a notion of *comparative similarity* between possible worlds relative to the "actual" world of evaluation. Simplifying somewhat, a counterfactual *'If had been A, would have been C'* ($A \mathrel{\Box\!\!\rightarrow} C$) is true if and only if C is true at all A-worlds that are maximally similar to the actual one. Stalnaker and Lewis account for various logical properties of counterfactuals by imposing conditions on the underlying similarity relation, but neither attempts a detailed analysis of this notion. Much of the subsequent work on modeling counterfactual reasoning can be viewed as attempts to make the notion of similarity more precise.

Recent years have seen an increased interest in the role of *causal (in)dependencies* in determining speakers' judgments about counterfactuals, driven in large part by advances in the formal representation and empirical verification of causal relations that had originated in statistics and artificial intelligence and had since had a major impact in psychology and other disciplines (Spirtes et al., 1993; Pearl, 2000). The formal vehicle of choice in this area is that of *Causal (Bayesian) Networks*, directed acyclic graphs whose edges represent the direction of causal influence and whose vertices are labeled with variables. Each distribution of values over these variables corresponds to a possible state of the system represented by the model. Causal networks are *partial* descriptions of the world, thus in general each state corresponds to a *class* of possible worlds in the Stalnaker/Lewis sense. The standard analysis of counterfactual reasoning about what would have been if some variable X had had value x relies on the notion of an external *intervention* which forces X to have value x but cuts all the causal links leading into X, ensuring that all those variables remain undisturbed whose values are not (directly or indirectly) caused by that of X. In effect, the counterfactual reasoning thus modeled is maximally "local" in the sense that for all variables Y which do not lie "downstream" of X in the direction of causal influence, the counterfactual *'If X had had value x, then Y would* still *have its actual value'* is invariably true.

It is an empirical question whether and to what extent speakers' judgments about particular counterfactuals actually reflect such highly localized reasoning. There is no doubt that the method of blocking the flow of information from effects to causes in modeling counterfactual inference captures

Dehghani, M., Iliev, R. & Kaufmann, S. (2009). Causal Explanations in Counterfactual Reasoning. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society, (CogSci).*

an intuitively real asymmetry. Yet psychological experiments have so far yielded only mixed support for the strong version of this idea (Sloman and Lagnado, 2005) and even systematic violations in some cases (Dehghani et al., 2007). It appears that while causal locality is an important factor in counterfactual inference, it interacts with other tendencies which may outweigh or overrule it in some cases. We believe that a systematic investigation of those other tendencies and of their interaction with causal locality is the key to further progress towards a better understanding of the notion of similarity at work in counterfactual inference.

In this paper we explore the role of *causal explanation* in evaluating counterfactuals. The basic idea is that in reasoning about what would have been the case if A had been true, the localist injunction to hold constant all the variables that causally influence whether A is true or not is sometimes unreasonably constraining, particularly when the hypothesized truth of A is a very unlikely outcome given their actual values. We hypothesize that speakers may resolve this tension by including in their deliberations the question of what would *explain* the hypothesized truth of A – that is, whether an alternative state of the causes of A would make its hypothesized truth less surprising.

In the following, we first discuss causal explanation and Gärdenfors' definition of the important notion of *explanatory power*. Next, we propose an alternative operator for the Bayesian network framework which would allow this framework to make more precise prediction about counterfactual conditionals, taking into account both the causal structure and the explanatory goodness of events. In the experiments section of the paper, we discuss three psychological experiments showing systematic violations of the Bayesian network framework by subjects who were asked to evaluate counterfactual statements, and how our new operator can account for these violations.

Causal Explanation

There is abundant evidence that humans have a deeply entrenched inclination towards providing and acquiring explanations (Keil, 2006; Lombrozo and Carey, 2006). This need to answer the "why" question is not limited to proposing naïve theories about the relationships between objects or events. Rather, the tendency to search for missing links and to understand properties has been shown to be linked to variety of cognitive processes, including predictions (Heider, 1958), diagnosis (Graesser and Olde, 2003), categorization (Murphy and Medin, 1985), and attention allocation (Keil et al., 1998).

There are different types of explanations, some of which (e.g., mathematical proofs) are not necessarily related to causal links. Typically, however, causality plays a major role in explanations. When we explain a fire by the action of an arsonist, we rely on a construal of the situation in which the arsonist's action is a cause and the fire is its effect. In many cases, however, the causal analysis of a situation presents a complex picture and agreeing on the best explanation (let alone the "correct" one) can be challenging if not impossible. Rarely it is the case that real-world events have clear, equivocal causes and effects. For one thing, in many cases the casual links are probabilistic rather than deterministic. Moreover, effects often have more than one relevant cause, and distinguishing between a focal cause and mere enabling conditions can be difficult (McGill, 1993). Furthermore, the process of finding the focal cause may be heavily context-dependent (Einhorn and Hogarth 1986; see also the papers in Collins et al. 2004).

Nevertheless, we believe that causal explanations play a crucial role in the interpretation of counterfactual conditionals. This is obviously the case in *backward* counterfactual reasoning, i.e., reasoning from a hypothesized effect to its causes, in answering counterfactual questions like (1a).

(1) a. If the Iraq war had not happened, would the 9/11 attacks have happened?
    b. If the 9/11 attacks had not happened, would the Iraq war have happened?

But explanations are also likely to be implicitly involved in our evaluation of *forward* counterfactuals like (1b): Even a speaker who does not believe that the Iraq war was a direct effect of the 9/11 attacks may answer the question quite differently depending on his or her beliefs about what actually caused the attacks, what would have prevented them, and how whatever may explain their non-occurrence would have affected the events leading up to the war.

In order to make this a bit more precise, we turn to Gärdenfors's (1988) formal definition of causal explanation. Gärdenfors defines an *epistemic state* as a triple $K = \langle W, P, B \rangle$, where W is a set of possible worlds with a common domain of individuals; P a function which, for each $w \in W$, defines a probability measure P_w on sets of individuals in w; and B a probability measure on subsets of W, representing degrees of belief. The distinction between P and B allows Gärdenfors to represent beliefs about probabilities, defined as expectations of P relative to B, and thus to include statements about probabilities in his object language. We ignore this feature for simplicity.

An agent who may need an explanation for an event E most likely already holds E to be true. Whether and how urgently an explanation is needed depends on the degree of belief given to E in a *contracted* state $K_{\overline{E}}$, an epistemic state in which E is not known but which is otherwise as similar to K as possible. There is no unique contraction in general, but Gärdenfors proposes as "typical" the case that $K_{\overline{E}}$ is the agent's last epistemic state prior to learning E (p. 176).

If, according to $K_{\overline{E}}$, E is certain to happen, then the fact that E did indeed happen does not require an explanation. An explanation is only required to the extent that E is unexpected according to $K_{\overline{E}}$. In Gärdenfors's terms, an agent asking for an explanation for E expresses a *cognitive dissonance* between E and the rest of her beliefs. This cognitive dissonance is measured by the *surprise value* of E. Sintonen

(1984) argues that the role of the explanans is to reduce the cognitive dissonance and provide *cognitive relief*, which he measures as the reduction of surprise provided by the update of $K_{\overline{E}}$ with the explanans.

Gärdenfors imposes two conditions on explanations relative to a belief state K. First, if an explanation is needed for E, it should increase the degree of belief in E according to $K_{\overline{E}}$. Thus C is considered a worthy explanation if $B_{\overline{E}}(E|C) > B_{\overline{E}}(E)$. A measure of the *explanatory power* of C, defined as the difference $B_{\overline{E}}(E|C) - B_{\overline{E}}(E)$, is used to predict speakers' choices between alternative explanations. The second condition for a worthy explanation is that it should not already be known in state K, i.e., $B(C) < 1$.

In order to define a *causal* explanation, Gärdenfors first gives a definition for a *cause*. In his formalism, C is a *cause* for E relative to the epistemic state only if it satisfies the following two conditions: (i) $P(E) = 1$, $P(C) < 1$ and (ii) $P_{\overline{C}}(E|C) > P_{\overline{C}}(E)$. In addition to the two conditions described for *explanations*, he restrict *causal explanations* by imposing the following additional condition on them: C is a *causal explanation* for event E, if C is a cause of E *in relation to* P^{+}_{C}.

Note that Gärdenfors' notions of goodness of explanation and explanatory power crucially refer to the conditional probability of the explanandum given the explanans, in relation to the unconditional probability of the explanans. Notably missing is the prior probability of the explanans. Chajewska and Halpern (1997) argue convincingly that this is a serious omission, as this prior probability can play a role in choosing between different alternative explanations.

Whatever the shortcomings of Gärdenfors' account are, the importance of the conditional probability of the explanandum given the explanans is itself doubtless an important factor in causal explanation. In this paper, we focus on this conditional probability in both the theoretical and the experimental part. We emphasize, however, that this limitation is not intended to deny that other factors should be considered, and we plan to consider them in the next phase of this work.

Selective Intervention

In this section, we propose an alternative approach for evaluating counterfactual conditionals in causal networks. In essence, it is a weakened version of the operation associated with Pearl's (2000) '*do*' operator. Recall that Pearl postulates a three-step procedure in reasoning counterfactually about the event that $X = x$: (i) Abduction: updating the exogenous variables according to the available evidence; (ii) Intervention: setting $X := x$ and cutting all causal links into X; and (iii) Prediction. Our modification concerns the second step, Intervention. As before, $do(X = x)$ involves forcing the variable X to have value x. However, rather than cutting all causal links into X and thus blocking any consequences of the intervention for X's non-descendants, the links are cut selectively following an analysis of the possible causal explanations for the hypothesized event that $X = x$.

Before going on to describe the operation in more detail, we point out two questions it raises, whose importance we acknowledge but whose investigation would go beyond the scope of the present study. First, what determines whether or not the intervention is total or selective? Clearly one case in which some modification to the causes of X is called for is when the actual values of those causes rule out X's having value x. On a clear day, one cannot consistently entertain the question of what would happen if there was a thunderstorm without imagining there being clouds in the sky. But we believe, and some of our experimental data below lend evidence to this view, that speakers manipulate the causes of X not only when $X = x$ is impossible under their actual values, but also when it is merely unlikely. Such cases seem to require an appeal to a notion like the "cognitive dissonance" discussed above, which presumably triggers a search for explanation when it exceeds a certain threshold. What that threshold is and whether and how it is determined by the details of the model and/or the nature of the system represented by the model is an empirical question which requires more investigation.

While the first question concerns the conditions under which selective intervention is triggered, the second question is the mirror image of the former and concerns cases in which the variable X has multiple parents in the causal structure: Once we allow for causal links into X to be left intact, what determines whether any links will be cut at all? Leaving all links intact amounts to conditionalizing the entire network on the observation that $X = x$. But this is at odds with the intuition that in counterfactual reasoning the intervention is, though perhaps not radical, still "minimal" in the sense that as many facts are held constant as is reasonably possible. Again, our experimental data below show evidence that even when participants leave some links intact, they do cut others. To accommodate this observation, what seems to be required is an appeal to the notion of "cognitive relief" mentioned above: Speakers are content with manipulations on the causes that provide *sufficient* relief, again with respect to some threshold which may be depend on facts about the model and/or the situation modeled.

This paper does not offer a precise answer to either of these questions. What our experiments show is that selective intervention does occur; exactly how it is triggered and constrained is a question left for future work.

With these caveats, we now give an informal outline of the idea behind selective intervention, illustrated with an example. The goal is, first, to determine whether an explanation for the hypothesized event is required, and second, if an explanation is required, which parents of the variable in question the explanation should make reference to.

Chain and Fork Topologies. Consider first the simple case of a causal *chain*, i.e., a graph in which the vertices are linearly ordered. Suppose some variable X with parent Y is the target of the counterfactual premise that $X = x$. The question here is whether this premise would call for an explanation. Assuming that X is known to have some value $x' \neq x$,

the first step is to undo this value setting and recalibrate the network with X as an unobserved variable. This is reminiscent of Pearl's preparatory *abduction* step; however, here the goal is not to update the network with an observation, but to "downdate" it with the removal of the observation that $X = x'$. The next step is to assess in this new network the degree of surprise or cognitive dissonance that an observation of $X = x$ would cause in the actual state. This depends on the value of Y if observed, and on the probability distribution over Y's values otherwise. If the surprise is deemed tolerable, then no explanation is required and the intervention proceeds as usual by cutting the link $Y \to X$. Otherwise, the link $Y \to X$ is left intact and the actual value of Y, if observed, is un-set, allowing changes in X to affect Y. In either case, the intervention concludes by setting $X := x$ and updating the network (with or without the link into X).

Now, clearly the decision to leave the link $Y \to X$ intact cannot be the end of the story. It raises the question of how to treat the link into Y, the link into Y's parent, and so on. Cutting no links at all results in a loss of the useful distinction between intervention and observation, hence of the ability to distinguish between counterfactual and non-counterfactual reasoning. We assume that the decision whether to cut the link into Y is made by applying the above decision procedure again – this time treating the set $\{Y, X\}$ as unobserved in downdating the network, then asking how surprising an observation of $X = x$ would be, and cutting the link into Y if the surprise would be tolerable. In general, the same procedure is applied recursively to the ancestors of X until a link is cut.

Selective intervention in the case of a causal *fork*, in which Y has multiple outgoing links, essentially follows the same logic as the causal chain. In a fork network, if an observation of $X = x$ results in a degree of surprise, then an explanation seems required to account for the change. In that case, the link $Y \to X$ will remain intact. However, if the surprise is deemed tolerable and no explanation seems required then the intervention cuts the link $Y \to X$.

Collider Topology. The case in which X has multiple incoming links is not fundamentally different from the chain topology. Now, however, instead of merely asking whether or not to leave the incoming link intact, the question is *how many* and *which* links to keep. Once again, if the values of X's parents jointly make the event $X = x$ unsurprising, the intervention proceeds as usual by cutting all links. Otherwise the intervention is selective, but keeping the number of intact links into X to a minimum. Leaving a single link $Y \to X$ may be sufficient: In this case, Y is affected by the hypothesized $X = x$, but all other parents of X remain unaffected. As before, whether it is "sufficient" in this sense to leave a single link intact depends on the agent's tolerance for cognitive dissonance. If no single link into X in itself provides enough relief, then all pairs of links into X are considered, and so on, up to the entire set of incoming links.

Some assumptions underlying the above description are still subject to empirical verification. For instance, the way we set up the search predicts that the agent will not even consider leaving two links intact if she can achieve sufficient cognitive relief by leaving only one. This may lead her to forego considerable relief in case the best way to leave one link is sufficient but inferior to options that involve leaving more links. In other words, we predict a strong preference for cutting links, hence for keeping the intervention local. Whether this prediction is borne out will have to be determined in future studies.

Illustration. We use the second experiment of Dehghani et al. (2007) to demonstrate how selective intervention is applied. Subjects were presented with a scenario in the collider topology ($A \to C \leftarrow B$) in which one cause is explicitly mentioned to be stronger than the other (in the above sense of conditional probability):

> A lifeboat is overloaded with people saved from a sinking ship. The captain is aware that even a few additional pounds could sink the boat. However, he decides to search for the last two people: a missing child and a missing cook. Soon, they find both people, but when they get onboard, the boat sinks.

The subjects were then asked the following counterfactual questions:

(2) If the boat had not sunk, ...
  a. ...would they have found the child? ($C \mathrel{\Box\!\!\to} B$)
  b. ...would they have found the cook? ($C \mathrel{\Box\!\!\to} A$)

There was a significant tendency for subjects to reply 'Yes' to the first question and 'No' to the second question. From the perspective of the procedure outlined in this section, this result suggests that the link $B \to C$ was cut whereas the link $A \to C$ was left intact. Assume that the boat's not sinking would have come as a considerable surprise to subjects given what they knew about the scenario, triggering the quest for a causal explanation of the boat's staying afloat. By considering each of A and B separately, subjects can easily check that the boat is more likely to stay afloat if the cook is not found (but the child is) than if the child is not found (but the cook is). Therefore the link from the cook's whereabouts (A) to the boat's fate (C) is left intact. As a result, after the update of the network with C, the posterior probability that the cook was not found is high, prompting subjects to answer 'Yes' to (2a) and 'No' to (2b).

In the next section, we compare the results of three new psychological experiments to the predictions of Causal Bayesian Networks and the method discussed in this section.

Experiments

The following experiments involve scenarios which contain facts with different frequencies of occurrence and investigate how these different rates affect subjects' responses to counterfactual questions. We then compare these responses to the predictions of Causal Bayesian Networks with and without the selective invention modification. Note that the subjects

were randomly divided into two groups for each question. Therefore, Group A from Experiment 1 does not correspond to subjects in Group A in Experiment 2 or Group A in Experiment 3.

Experiment 1

In the first experiment, we examine how in a collider topology the likelihood of causes affect people's evaluation of counterfactual statements. Sloman and Lagnado (2005) suggest that people are more likely to keep the state of the consequent intact when the effect is part of the antecedent of the counterfactual statement. Therefore, if the effect has been intervened on, the status of the cause(s) should not change and hence the likelihood of occurrence of effects should not play a role when evaluating counterfactuals. We predict the link between the cause with the highest explanatory power and its effect will be preserved, while the other links will be severed. Therefore, people should more often undo the cause with the highest explanatory power than the other cause.

Method. 58 Northwestern undergraduate students were presented with a series of scenarios, and after each scenario they were asked to evaluate the likelihood of a number of counterfactual statements. The questions were presented in form of a questionnaire, and subjects were asked to rate the likelihood of each question from 0 to 10, 0 being "definitely no" and 10 being "definitely yes". A scenario that described a test situation was presented between subjects. Group A was presented with the first question, while group B was presented with the second question.

Scenario

90% of the time ball A moves, ball C moves.
10% of the time ball B moves, ball C moves.
Balls A, B and C definitely moved.

(A) If ball C had not moved, would ball A have moved? ($C \mathrel{\Box\!\!\rightarrow} A$)

(B) If ball C had not moved, would ball B have moved? ($C \mathrel{\Box\!\!\rightarrow} B$)

Results. The mean for $C \mathrel{\Box\!\!\rightarrow} A$ was 3.36 while the mean for $C \mathrel{\Box\!\!\rightarrow} B$ was 6.13. The difference between the two questions was highly significant ($t(42) = -2.95, p < 0.005$).

Discussion. Causal Bayesian networks predict that the answers to both of the questions should be Yes (10), as intervening on C will result in cutting the link from both of its parents. However, the participants more often answered 'No' to the first question and 'Yes' to the second. This results suggests that the $B \rightarrow C$ link was cut, but the $A \rightarrow C$ link was left intact. Ball C not moving results in a degree of surprise, given that it has been explicitly mentioned in the scenario that ball C definitely moved. By considering both A and B separately, subjects can easily see that $B_{\overline{C}}(C|A) > B_{\overline{C}}(C|B)$. That is, ball C is more likely not to move if ball A doesn't move (and ball B does) than if A does move (and ball B doesn't). Therefore, A contains higher explanatory power for C not moving. As the results of this, the link between A and C is left intact. After updating the network, the posterior probability of A moving becomes low, resulting in the answer to the first question to be 'No'. However, the posterior probability of B not moving becomes high resulting in the answer to the second question to be 'Yes'. Comparing these predictions to subjects' responses reveals that selective intervention seems to be more consistent with subjects' answers than the normal '*do*' operator.

Experiment 2

In this experiment we investigate how intervening on an effect in a fork topology changes the status of the common cause. In a fork network, changing the value of the effect for which the cause is the sole explanation, creates a degree of surprise and hence, an explanation seems required to account for the change. Therefore, we predict the link between the common cause and the child for which the cause is the best (only) explanation is preserved while the other link is dropped.

Method. The same participants were presented with a scenario in the fork topology. Group A was presented with the first question, while group B was presented with the second question.

Scenario

Ball A causes Ball B to move 5% of the time.
Ball A causes Ball C to move 100% of the time.
A, B and C definitely moved.

(A) If ball B had not moved, would ball C have moved? ($B \mathrel{\Box\!\!\rightarrow} C$)

(B) If ball C had not moved, would ball B have moved? ($C \mathrel{\Box\!\!\rightarrow} B$)

Results. The mean for $B \mathrel{\Box\!\!\rightarrow} C$ was 8.38 while the mean for $C \mathrel{\Box\!\!\rightarrow} B$ was 4.00. The difference between the two questions was highly significant ($t(27) = 4.34, p << 0.001$).

Discussion. Causal Bayesian networks predict that an intervention on the effect should not change the value of its cause(s). Therefore, in the case of the fork topology, intervening on one of the child nodes should not effect the value of other nodes in the graph. Hence, according to this theory, the answer to both questions should be 'Yes' (10). The results suggest that the subjects however, selectively intervened on B dropping the $A \rightarrow B$ link and not on C, keeping the $A \rightarrow C$ link intact. Given that it has been explicitly mentioned that ball A definitely moved and ball A is the only cause for C moving, Ball C not moving would call for an explanation. Hence, the link $A \rightarrow C$ would be preserved, as ball C is more likely not to move if ball A does not move. However, A is not the only reason for B moving and therefore there could be other explanations for B not moving. The intervention proceeds as usual by cutting the link $A \rightarrow B$. Analyzing the above questions in the new network reveals that the answer to the first question should be 'Yes' (10), while the answer to the second question should be 'No' (0).

Experiment 3

The same schema as the second experiment was used in this experiment. Only the probability of A causing B was increased to 95%, while the other probability was not changed.
Method. The same participants were presented with another scenario. Group A was presented with the first question, while group B was presented with the second question.

Scenario
Ball A causes Ball B to move 95% of the time.
Ball A causes Ball C to move 100% of the time.
A, B and C definitely moved.

(A) If ball B had not moved, would ball C have moved? ($B \mathrel{\Box\!\!\rightarrow} C$)

(B) If ball C had not moved, would ball B have moved? ($C \mathrel{\Box\!\!\rightarrow} B$)

Results. The mean for $B \mathrel{\Box\!\!\rightarrow} C$ was 7.04 while the mean for $C \mathrel{\Box\!\!\rightarrow} B$ was 3.68. The difference between the two questions was highly significant ($t(40) = 3.06, p < 0.005$).
Discussion. The predictions of both methods remain the same as the previous experiment. Even though,the difference between A causing B and A causing C is only 5%, subjects clearly distinguished between the two causal links. We believe this is due to the fact that A is the sole explanation for C moving. However, B could have potentially been moved by other causes and A not moving did not trigger a need for an explanation. This experiment highlights the fact that intervention seems to have a clear qualitative effect, cut or no cut, rather than a gradient one.

Conclusions

We proposed an alternative approach for evaluating counterfactual conditionals based on the relationship between causal explanations and the causal topology of the graph. This approach consists of a modification to Pearl's intervention operator. As before, $do(X = x)$ involves forcing the variable X to have value x. However, rather than cutting all causal links into X and thus blocking any consequences of the intervention for X's non-descendants, we proposed selective intervention in which the links are cut selectively following an analysis of the possible causal explanations for the hypothesized event that $X = x$.

The result of our experiments show that selective intervention does occur and causal explanations seem to play a role in this selection. However, causal explanation may not be the only factor influencing this process. Previously the relationship between fact mutability, intervention and evaluation of counterfactual have been explored by Dehghani et al. (2007). Kahneman and Miller (1986) claim that there are certain facts that are easier to mentally undo, or mutate than others. We have previously shown that the more mutable a fact is the more likely it is that the link from its cause would not be cut. We believe that the notions of explanatory power and mutability can and should be combined in a more comprehensive and plausible account of selective intervention.

References

Chajewska, U. and Halpern, J. Y. (1997). Defining explanation in probabilistic systems. In *Proceedings of the UAI-97*, pages 62–71.

Collins, J., Hall, N., and Paul, L. A., editors (2004). *Causation and Counterfactuals*. The MIT Press.

Dehghani, M., Iliev, R., and Kaufmann, S. (2007). Effects of fact mutability in the interpretation of counterfactuals. In *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*, Nashville, TN.

Einhorn, H. J. and Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, 99:3–19.

Gärdenfors, P. (1988). *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge, MA.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Harvard University Press., Cambridge, Mass.

Graesser, A. and Olde, B. (2003). How does one know whether a person understands a device? the quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*, 95:52436.

Heider, F. (1958). *The psychology of interpersonal relations*. John Wiley & Sons., New York.

Kahneman, D. and Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93:136–153.

Keil, F. C. (2006). Explanation and understanding. *Annual Reviews of Psychology*, 57:227–254.

Keil, F. C., Smith, C., Simons, D. J., and Levin, D. T. (1998). Two dogmas of conceptual empiricism: implications for hybrid models of the structure of knowledge. *Cognition*, 65:103–35.

Lewis, D. (1973). *Counterfactuals*. Blackwell, Oxford.

Lombrozo, T. and Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, 99:167204.

McGill, A. L. (1993). Selection of a causal background: Role of expectation versus feature mutability. *Journal of Personality and Social Psychology*, 64:701–707.

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

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, London.

Sintonen, M. (1984). *The Pragmatics of Explanation*. North-Holland, Amsterdam.

Sloman, S. A. and Lagnado, D. (2005). Do we 'do'? *Cognitive Science*, 29:5–39.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag, New York.

Stalnaker, R. (1968). A theory of conditionals. In Rescher, N., editor, *Studies in Logical Theory*, pages 98–112. Blackwell, Oxford.