Effects of Fact Mutability in the Interpretation of Counterfactuals

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 Road
Evanston, IL 60208-2710 USA

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

Abstract

This paper explores the relationship between fact mutability, intervention and human evaluation of counterfactual conditionals. Two experiments are reported that show the effects of causal strength and causal distance on fact mutability and intervention. Subjects' answers are compared to the predictions of three models of counterfactual reasoning in Artificial Intelligence. This comparison demonstrates that logical inferences and graph topologies are not sufficient for modeling all aspects of human counterfactual reasoning.

Keywords: Counterfactual Reasoning; Causal Networks; Norm Theory.

Introduction

Counterfactual reasoning has long been a subject of interest to philosophers (e.g. Leibniz, 1686; Hume, 1748; Goodman, 1947; Lewis, 1973; Stalnaker, 1968). More recently linguists, psychologist, and later on cognitive scientists, have become interested in the study of the concept of "what would have been" and how reasoning about events that almost happened provides us with knowledge that cannot be deduced from simple facts or indicative conditionals (e.g. Kahneman and Miller, 1986; Sternberg and Gastel, 1989). In the last two decades there have been several attempts to model counterfactual reasoning in AI (Ginsberg, 1986; Costello and McCarthy, 1999; Pearl, 2000; Hiddleston, 2005, among others). The advantage of such formal models is that they make precise predictions for particular cases. In this paper we briefly review two of these models and demonstrate using evidence from two psychological experiments that these models do not capture the full spectrum of human counterfactual reasoning. Specifically, we illustrate how causal distance and causal strength affect the interpretation of counterfactual statements. In conclusion, we argue that utilizing the psychological findings on fact mutability and similarity are crucial to building cognitively plausible computational models of counterfactual reasoning.

First, we briefly review the Stalnaker/Lewis theory of counterfactuals. Next, we discuss Causal Bayesian Networks (Spirtes, Glymour, and Scheines, 1993; Pearl, 2000) and the psychological evidence supporting this model. Also in the same section, we review Hiddelston's (2005) extension to Causal Bayesian Networks. We then discuss Kahneman and Miller's Norm Theory (1986) and two psychological experiments designed to test the correctness of the predictions of the AI models.

The Stalnaker/Lewis Theory

Many models 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 *i* of evaluation. Thus Lewis's truth conditions state that a counterfactual 'If it were that *Antecedent*, then it would be that *Consequent*' ($A \mathrel{\square\!\!\rightarrow} C$) is then true at world *i* "if and only if, if there is an antecedent-world accessible from *i*, then the consequent holds at every antecedent-world at least as close to *i* as a certain accessible antecedent-world" (p. 49). Assuming for simplicity that there is a set of *A*-worlds that are maximally similar to *i*, this means that the counterfactual $A \mathrel{\square\!\!\rightarrow} C$ is true if and only if *C* is true in all of those maximally similar *A*-worlds.

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. However, Lewis (1979), noting that his theory "must be fleshed out with an account of the appropriate similarity relation, and this will differ from context to context," gives an informal ranked list of general "weights or priorities" in determining similarity: first, to avoid big, widespread, diverse violations of law; second, to maximize the spatio-temporal region of perfect match of particular fact; third, to avoid small, localized violations of law; fourth, to secure approximate similarity of particular facts.

Despite the informality of these guidelines, one can discern a priority of laws over particular fact, and of "big" discrepancies over "small" ones. Much of the subsequent work on modeling counterfactual reasoning is based on similar intuitions and can be viewed as attempts to make the notion of similarity more precise. However, in view of

Lewis's emphasis on the inherent vagueness and context-dependence of the notion, it is not surprising that there is considerable variation in detail.

Ginsberg's Computational Implementation

Ginsberg (1986) models the interpretation of counterfactuals relative to an imaginary "large database describing the state of the world at some point." This is in line with AI conventions on the one hand, and the logical school of Premise Semantics (Veltman, 1978; Kratzer, 1981), on the other. In this framework, the move from the world of evaluation to antecedent-worlds corresponds to a *revision* of the database, and the problem of determining similarity is mirrored by the question of which facts to keep and which ones to give up in order to keep this revision *minimal*. Again following AI conventions, Ginsberg extends the term "possible world" to *partial* descriptions, which would correspond to sets of (total) possible worlds in the Stalnaker/Lewis framework.

Overall, Ginsberg's theory is set up to minimize differences in facts and localize violations of rules. In particular, his second postulate ensures that violations are treated as exceptions to existing rules, rather than the workings of different ones. In this, the theory reflects Lewis's emphasis on minimizing violations of law while maximizing correspondence in particular fact.

Counterfactuals in Causal Bayesian Networks

Causal Bayesian Networks have recently gained considerable currency as a formal tool for representing domain knowledge and modeling causal and counterfactual reasoning (Pearl, 1998, 2000; Spirtes *et al.*, 1993). A Bayesian Network is a directed acyclic graph whose vertices represent variables and whose topology encodes independencies between those variables in the form of the *Markov Assumption*: The probability of a variable X is fully determined by the values of its immediate parents $pa(X)$ in the graph. A Bayesian Network is *causal* if all arrows are assumed to lead from causes to effects. Pearl (2000) assumes that causation is deterministic; uncertainty is modeled as a probability distribution over a distinct set of "exogenous" variables. Under this assumption, the values of $pa(X)$ jointly determine the *value* of X.

The causal interpretation makes a special update operation available, formally represented using the "*do* operator": The result of applying the operation $do(X=x)$ to a network with a vertex X results in a new network in which X has value x and all arrows leading into X are removed. In this network, the variables in $pa(X)$ are independent of X, hence unaffected by standard algorithms of belief propagation. The intention is that $do(X=x)$ represents an external *intervention* upon X which disrupts the causal process that normally determines the value of X, so the fact that $X=x$ does not warrant any inferences about its normal causes. Thus, $do(X=x)$ represents a *local* update whose effects are limited to the descendants of X in the graph.

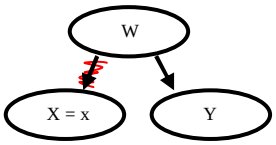

Figure 1: $do(X=x)$, an external action making X independent of its parent W

Pearl claims that the *do* operator "is a crucial step in the semantics of counterfactuals" (Pearl, 2000): A counterfactual with antecedent 'If it were that $X=x$' is interpreted in a causal network by first applying $do(X=x)$, then evaluating the consequent in the resulting modified network. This leads to the strong prediction that unless the consequent is not a (direct or indirect) effect of X, it should be unaffected by the intervention and retain its prior probability.

In a series of six psychological experiments, Sloman and Lagnado (2005) investigate whether human subjects' responses to counterfactual statements can be explained using the method of determining the truth of counterfactuals in Causal Bayesian networks and the $do(x)$ operator. Overall, they conclude that responses of the subjects were more or less compatible with the effect of the $do(x)$ operator. But they argue that "representing intervention is not always as easy as forcing a variable to some value and cutting the variable off from its causes. Indeed, most of the data reported here show some variability in people's responses. People are not generally satisfied to simply implement a 'do' operation. People often want to know precisely how an intervention is taking place." (Sloman and Lagnado, 2005).

Hiddleston's Theory of Counterfactuals

Hiddleston (2005) is a recent attempt to synthesize the insights of different strands of research into a theory which accounts for a number of problematic examples that were discussed in the philosophical literature. He adopts the basic idea of representing causal relations as directed acyclic graphs from Spirtes *et al.* (1993) and Pearl (2000), but his interpretation of these graphs dispenses with some of the assumptions about causality made by the latter. Unlike Pearl, he allows for non-deterministic causal laws, and he weakens the Markov Assumption to allow for cases in which a variable's taking on a different value is impossible given the actual values of its parents and the causal laws.

Hiddleston evaluates counterfactuals relative to models consisting of a causal graph and an assignment of values to all variables. One may think of such models as "possible worlds" and call the values that variables are assigned in them their "actual" values. Relative to such a model M, he introduces a notion of *positive parent* of a variable X (written $ppa(X)$), a subset of $pa(X)$, defined as those parents Y of X such that the conditional probability of X's taking on its actual value, given that all of X's parents (including Y) have their actual values, is strictly higher than the corresponding conditional probability in the event that Y

alone among X's parents takes on a different value. Thus the set of positive parents of X relative to the same causal structure may differ between alternative value assignments.

Instead of modeling the interpretation of a counterfactual $A \square\!\!\rightarrow C$ whose antecedent is the assertion that some variable X has value x by "cutting the links" from $pa(X)$ to X, his rule involves a comparison between alternative "A-models" with the same causal structure as the original model M and in which $X=x$. Among those, a model is "*A-minimal*" if (i) the set of non-descendants Z of X such that both Z and Z's positive parents have the same values as in M is maximal; and (ii) the set of "causal breaks" - variables Z which take on a different value while the values of Z's positive parents are the same as in M - is minimal. The counterfactual $A \square\!\!\rightarrow C$ is true in M if and only if C is true in all of those A-minimal models. These conditions, as Hiddleston puts it, "force any causal break to be as minor and late as is lawfully possible."

Hiddleston leaves room for the incorporation of context dependence by allowing that depending on the situation, only a subset of all "causal breaks" may be *relevant*. The role of the context is not formalized in his models, however.

Norm Theory

One important perspective in research on counterfactual reasoning in psychology is Kahneman and Miller's Norm Theory (1986). According to this theory, the outcome of a situation is compared to a norm, which is constructed online based on past experiences and expectations, as well as on the particular outcome. Outcomes that are similar to the norm are considered normal, while outcomes that are significantly different are considered abnormal. Abnormal outcomes activate their normal counterparts, thus they are an invitation for counterfactual thinking. Kahneman and Miller (1986) argue that when constructing the norm, there are certain facts that are easier to mentally undo, or "mutate" from the actual world than others. People are more or less able to predict the availability of counterfactual alternatives to a given situation, "Some aspects of reality are more mutable than others, in the sense that alternatives to them come more readily to mind" (Kahneman, 1995). Hence, when thinking about causal situations, there are certain causes that tend to be easier to modify or mutate.

Fact Mutability and Intervention

According to the Norm Theory, when faced with multiple causes with different mutability rates, alternatives to causes which have higher mutability rates come more readily to mind and therefore these causes tend to be easier to mentally undo and give up their values. In other words, the higher the mutability rate of a fact, the less likely it is that its value will stay the same under an intervention.

Although deciding the mutability of facts is similar to the problem of determining similarity between possible worlds, fact mutability has not been taken into account in the philosophical and AI literature. We believe that this is an important psychological factor which should play a role in theories of counterfactual reasoning.

Two context free conditions which we believe affect fact mutability are causal distance and causal strength. Causal distance is the relative closeness of the antecedent of the counterfactual from its consequent in the causal graph. We predict that the closer the antecedent is to its consequent, the more mutable the antecedent is and therefore the easier it is to mentally undo it. The second condition is the strength of the causal connection between the antecedent and the consequent. Psychological evidence suggests that people not only use causal structures but also utilize beliefs about causal strengths (e.g. Kushnir and Gopnik, 2005; Waldmann & Hagmayer, 2001). ΔP (Jenkins and Ward, 1965) and power PC (Cheng, 1997) both utilize conditional probabilities to compute causal strength. In Bayesian Networks, these parameters can be calculated from the conditional probabilities. However, none of the models discussed above utilizes causal strength when evaluating counterfactuals. We predict that if an effect has multiple causes, the ones which have a stronger causal connection to it are more easily undone than the ones which have a weaker connection. This is consistent with our prediction about causal distance; the connection with direct causes is stronger than that with distant causes.

There are also context dependent factor which affect the mutability of facts. For example, actions are more mutable than failures to act (Kahneman & Miller, 1986). We investigate the affect of these context dependent factors on the evaluation of counterfactuals in a separate paper (in preparation).

Experiments

In the following experiments we present scenarios containing facts with different mutability rates and investigate how these different rates affect subjects' responses to counterfactual questions. We then compare these responses to the predictions of the normative models discussed in the previous section focusing on the predictions of Causal Bayesian Networks.

Experiment 1

In this experiment we investigated how the distance of the antecedent from its consequent in the causal graph can change the mutability of facts when analyzing a backward counterfactual statement, where the effect is the antecedent of the counterfactual and the cause is the consequent. 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 distance of the cause from the effect should not play a role when evaluating counterfactuals. We predict immediate causes to be more mutable than distant causes. Therefore, it should be easier for people to undo immediate causes than distant causes.

In the following three scenarios we asked questions about different variables *A*, *B* and *C*. In each scenario, *A* causes *B* and *B* causes *C*, representing the *chain* network topology in Figure 2. We then asked the following counterfactual questions:

(1) If *C* had not happened, would *B* have happened?
(2) If *C* had not happened, would *A* have happened?
(3) If *B* had not happened, would *C* have happened?
(4) If *B* had not happened, would *A* have happened?

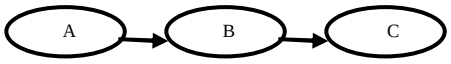

Figure 2: *Chain* topology

According to Pearl's (2000) model, given the above network $P(A \mid do(C)) = P(A \mid do(B)) = P(A)$ and $P(B \mid do(C)) = P(B)$, therefore the truth-value of the cause should not change given any intervention on its effect.

Method

36 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 on a computer screen, and subjects were asked to rate the likelihood of each question from 0 to 10, 0 being "definitely no" and 10 being "definitely yes."

Three different scenarios were designed to follow the same causal structure where *A* causes *B*, *B* causes *C* and both *A* and *C* definitely happened. Logical abbreviations in the parenthesis are added here for the reader's convenience; they were not shown to the subjects.

Scenario 1

Ball *A* causes Ball *B* to move.
Ball *B* causes Ball *C* to move.
Balls *A*, *B* and *C* definitely moved.

Question 1: If Ball *C* had not moved, would ball *B* still have moved? ($C \square\!\!\rightarrow B$)
Question 2: If Ball *C* had not moved, would ball *A* still have moved? ($C \square\!\!\rightarrow A$)
Question 3: If Ball *B* had not moved, would ball *C* still have moved? ($B \square\!\!\rightarrow C$)
Question 4: If Ball *B* had not moved, would ball *A* still have moved? ($B \square\!\!\rightarrow A$)

Scenario 2

Tom's alarm clock did not ring on Wednesday morning, resulting in Tom being late for work. Because Tom was late that morning, Tom's boss gave him extra work for the weekend.

Question 1: If Tom had not had extra work for the weekend, would he have been on time on Wednesday morning? ($C \square\!\!\rightarrow B$)
Question 2: If Tom had not had extra work for the weekend, would his alarm have rung on Wednesday morning? ($C \square\!\!\rightarrow A$)
Question 3: If Tom had not been late on Wednesday morning, would he have extra work for the weekend? ($B \square\!\!\rightarrow C$)
Question 4: If Tom had not been late on Wednesday morning, would his alarm have rung on time? ($B \square\!\!\rightarrow A$)

Scenario 3

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 person: a missing child. Soon, they find the 5-year-old girl, but when she gets onboard, the boat sinks

Question 1: If the boat had not sunk, would they have found the child? ($C \square\!\!\rightarrow B$)
Question 2: If the boat had not sunk, would the captain have decided to search for the child? ($C \square\!\!\rightarrow A$)

Results

The results are summarized in Table 1. The Moving Balls scenario replicated one of Sloman and Lagnado's (2005) findings: forward counterfactuals are treated differently from backwards counterfactuals. The responses to the forward counterfactual question "If *B* had not happened would *C* have happened?" ($B \square\!\!\rightarrow C$) were lower than those to any of the backward counterfactuals (mean=2.8, SD = .5, $F(1,35) = 9.53$, $p<.05$), suggesting that the hypothetical absence of the cause has a stronger influence on the state of the effect than vice versa.

A repeated measure ANOVA was conducted comparing the mean of the answers to the three backward counterfactuals, revealing significant: A post hoc test revealed that $B \square\!\!\rightarrow A$ was significantly lower than both $C \square\!\!\rightarrow A$ and $C \square\!\!\rightarrow B$ ($F(1,35) = 7.89$, $p<.05$) However, there was no reliable difference between the estimated probabilities of $C \square\!\!\rightarrow A$ and $C \square\!\!\rightarrow B$.

The same pattern was observed in the Alarm Clock scenario: there was a significant difference between

Scenario/Questions	$C \square\!\!\rightarrow B$	$C \square\!\!\rightarrow A$	$B \square\!\!\rightarrow A$	$B \square\!\!\rightarrow C$
Moving Balls	5.9	5.6	4*	2*
Alarm Clock	5.6	5.3	7.9*	2.9*
Lifeboat	4.5*	6.2	N/A	N/A

Table 1: Subjects' mean response to backward counterfactual questions

backward and forward counterfactuals ($F(1,35)=21.23$, $p<0.05$). No significant difference was detected between $C \square\!\!\rightarrow A$ and $C \square\!\!\rightarrow B$, but the estimated $B \square\!\!\rightarrow A$ was much higher than any of the other two ($F(1,35)=9.92$, $p<.05$).

In the third scenario we asked only two questions, $C \square\!\!\rightarrow A$ and $C \square\!\!\rightarrow B$, which came out to be the significantly different ($F(1,35)=12.10$, $p<.05$).

Discussion

Causal Bayesian networks predict that an intervention on the effect should not affect the value of its cause(s). Therefore, answers to backward counterfactual questions in the first and third scenario should all be Yes (10) and in the second scenario should all be No (0). Hiddleston's and Ginsberg's models predict the answers to all the questions in the three scenarios to be No (0).

Subjects' answers to $B \square\!\!\rightarrow A$ were consistently different from answers to $C \square\!\!\rightarrow A$. We believe that this difference may be due to the distance between the cause and the effect nodes. That is, the closer the antecedent is to its consequent, the easier it is to undo its value. Therefore, it is easier to undo the value of A in $B \square\!\!\rightarrow A$ than in $C \square\!\!\rightarrow A$. Following this hypothesis, in the first scenario where A is true (10) in the consequent of counterfactuals, it seems that subjects more often undid the value of A in $B \square\!\!\rightarrow A$ and altered it to false (0) than in $C \square\!\!\rightarrow A$. In the second scenario the same trend was observed: A is false in the consequent of counterfactuals, and subjects more often undid its value to true in $B \square\!\!\rightarrow A$ than in $C \square\!\!\rightarrow A$.

Also, in the last scenario significant difference was observed between answers to $C \square\!\!\rightarrow B$ and $C \square\!\!\rightarrow A$ which again agrees with the above hypothesis and is due to the fact that B is closer to C than A. Therefore, it is easier to undo, or give up, its value.

All of the observed differences were consistent with our prediction. However, no difference was detected between $C \square\!\!\rightarrow B$ and $C \square\!\!\rightarrow A$ in the first two scenarios. We believe this might be due to the differences of the context in which the conditionals were being evaluated, as B in the third scenario seems to be a more salient cause that in the other two scenarios. How different context affects the evaluation of counterfactuals is part of our ongoing research.

In conclusion, the experiments show that the degree to which people are willing to mutate the antecedent of the counterfactual varies with the location of the antecedent and the context of the scenario. This conclusion contradicts the predictions of the three models discussed in this paper.

In the next part we describe another experiment which questions the applicability of the $do(x)$ operator, and the logical operations of the two other models.

Experiment 2

In addition to causal distance, another factor which we believe may influence counterfactual reasoning is the causal strength between the antecedent and the conclusion of counterfactual statements. We predict that the stronger the causal effect is, the easier it is to undo it. Hence, when faced with two causes with a common effect, people more easily undo the cause which has the connection to the effect.

The topology of the network in the following scenario is a *collider* (Figure 3): There are two causes which affect the same effect. We hypothesize that a difference between A and B in the strength of their respective causal connection with C will affect which of the two is given up in counterfactual reasoning about C, contrary to the assumption in the Causal Networks literature that both should be equally unaffected.

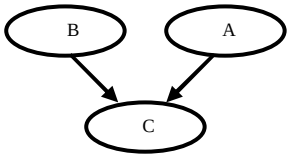

Figure 3: *Collider* Topology

Method

The same participants were presented with another scenario which was very similar to the Lifeboat scenario, with the only difference that this time there were two persons in the water.

Scenario 1

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.

Questions:

(1) If the boat had not sunk, would they have found the child? ($C \square\!\!\rightarrow B$)

(2) If the boat had not sunk, would they have found the cook? ($C \square\!\!\rightarrow A$)

Results

The mean for $C \square\!\!\rightarrow A$ was 6.5 while the mean for $C \square\!\!\rightarrow B$ was 7.0. The difference between the two questions was significant ($F(1,35) = 7.19$ $p<.05$).

Discussion

Causal Bayesian networks predict that the answers to both of the questions should be Yes (10), and Hiddleston's and Ginsberg's models predicts that the answers should be No (0).

It was explicitly mentioned in the scenario that a few additional pounds would be enough to sink the lifeboat, therefore both the cook and the child were potential causes for the sinking of the boat. However, the results show that it was easier for the subjects to undo A (cook was found). This result suggests that the subjects were less likely to cut the link between the weaker cause and the effect compared to the link between to the stronger cause and the same effect. Thus causes with stronger effects are more mutable and more likely to be intervened on than weak causes.

Conclusion

Although the models discussed are able to correctly predict people's judgments about many counterfactual questions, they fail to capture certain aspects of human counterfactual reasoning. We argue that causal strength and causal distance influence the interpretation of counterfactual conditionals. None of the models reviewed in this paper utilize these two factors.

While intervention may not necessarily remove the causal dependencies between the antecedent of the counterfactual and its immediate causes, how people choose the location of intervention is related to psychological factors including the mutability of facts involved in the case.

In a series of new experiments we aimed at a direct comparison between the context sensitivity of Norm Theory and the context neutrality of the Bayesian Networks approach in analyzing counterfactuals conditionals (in preparation). Among the important predictions that the Norm Theory makes is that when people analyze counterfactuals, they are more likely to undo actions that lead to some type of consequence rather than inactions that lead to the same consequence. In other words, actions are more mutable compared to inactions. We have found evidence for this effect in a set of experiments which follow the same causal network but set up a different context. Generally, if the causal network is a collider, people tend to undo the value of the node which represents an action. This result follows the main claims of this paper that not only the causal structure of the graph is important, but the content of each node and the context in which the conditional is being evaluated affect speakers' evaluation of counterfactual conditionals.

Determining the mutability of facts and performing similarity analyses between potential worlds are two important steps in evaluating the truth of counterfactual conditionals. These two context dependent factors are influenced by a variety of psychological factors and are not fully captured by the popular models of counterfactual reasoning in AI. In future work, we plan to continue our experimental studies on how psychological findings on similarity can be related to fact mutability and in general to Stalnaker and Lewis' notion of comparative similarity, and work towards a formal theory in which the role of these factors can be implemented.

Acknowledgments

This research was supported in part by a grant from an AFSOR MURI. The authors would like to thank Ken Forbus and Lance Rips for their insights throughout this work.

References

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

Costello, T. & McCarthy, J. (1999). Useful Counterfactuals. *Electronic Transactions on Artificial Intelligence*.

Ginsberg, M. L. (1986). Counterfactuals. *Artificial Intelligence* 30:35-79.

Goodman, N. (1983). *Fact, Fiction, and Forecast*. Cambridge, Mass.: Harvard University Press.

Griffiths, T. L. & Tenenbaum, J. B. (2005). Structure and Strength in Causal Induction. *Cognitive Psychology* 51(4):334-384.

Hiddleston, Eric (2005). A Causal Theory of Counterfactuals. *Noûs*, 39(4):632-657.

Hume, D. (1748). *An Enquiry concerning Human Understanding*.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. Psychological Monographs, 79.

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

Kahneman, D. (1995). Varieties of Counterfactual Thinking. In N. J. Roese. & J. M. Olson J. M. *What Might Have Been: The Social Psychology of Counterfactual Thinking*.(pp. 375-396). Mahwah, NJ: Lawrence Erlbaum.

Kratzer, A. (1981). Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic* 10:201-216.

Kushnir, A. & Gopnik, A. (2005). Young Children Infer Causal Strength from Probabilities and Interventions. *Psychological Science* 16 (9):678-683

Leibniz, G.E. (1686). *Discourse on Metaphysics and Other Essays*.

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

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow, *Noûs* 13:455-76.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Francisco.

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

Rescher, N. (1964). Hypothetical reasoning. Amsterdam: North-Holland.

Sloman, S.A., & 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 N. Rescher (ed.), *Studies in Logical Theory* (pp. 98-112). Oxford: Balckwell.

Sternberg, R. J., & Gastel, J. (1989b). If dancers ate their shoes: Inductive reasoning with factual and counterfactual premises. *Memory and Cognition* 17:1-10.

Veltman, F. (1976). Prejudices, presuppositions and the theory of counterfactuals. In Groenendijk, J. and M. Stokhof (eds.) *Amsterdam Papers in Formal Grammar*, Vol. 1 (pp. 248-281). University of Amsterdam.

Waldmann, M. R. & Hagmayer Y. (2001). Estimating Causal Strength: The Role of Structural Knowledge and Processing. *Cognition* 82(1):27-58.