Causal Explanation and Fact Mutability in Counterfactual Reasoning

Morteza Dehghani Rumen Iliev Stefan Kaufmann

Abstract

Recent work on the interpretation of counterfactual conditionals has given much attention to the role of causal independencies. One influential idea from the theory of Causal Bayesian Networks is that counterfactual assumptions are made by *intervention* on variables, leaving all of their causal non-descendants unaffected. But intervention is not applicable across the board. For instance, *backtracking* counterfactuals, which involve reasoning from effects to causes, cannot proceed by intervention in the strict sense, for otherwise they would be equivalent to their consequents. We discuss these and similar cases, focusing on two factors which play a role in determining whether and which causal parents of the manipulated variable are affected: Speakers' need for an explanation of the hypothesized state of affairs, and differences in the 'resilience' of beliefs that are independent of degrees of certainty. We describe the relevant theoretical notions in some detail and provide experimental evidence that these factors do indeed affect speakers' interpretation of counterfactuals.

Counterfactual reasoning plays an important role in causal inference, diagnosis, prediction, planning and decision making, as well as in 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 (Goodman 1955). The problem has been approached from various directions, employing disparate theoretical frameworks and methodologies. Amidst this variety, some common themes can be discerned which in recent years have facilitated communication across disciplines and ushered in a confluence of theoretical and methodological views.

Especially important in this connection is the role of *causal relations*, which became amenable to new ways of mathematical modeling and empirical verification thanks to the work of Spirtes et al. (1993), Pearl (2000), Glymour (2001), Halpern and Pearl (2005), and others. Most relevantly in the present context, Pearl (2000) offered the notion of causal *intervention* as a mode of hypothetical update operative in counterfactual reasoning. Technically, intervention is an operation which forces a particular value on a variable while ensuring that the values of all its non-descendants in the causal structure remain intact. This idea has in some form or other informed many recent approaches to counterfactuals (Hiddleston 2005, Kaufmann 2004, 2005, 2009, Schulz 2007, and others). At the same time, it seems clear that causal intervention, for all its intuitive appeal, is only one piece of the puzzle and does not account for the full range of speakers' judgments. For instance, Sloman and Lagnado (2005) showed that people's judgments about counterfactual questions can vary widely depending on how exactly the premise is phrased. Even more fundamentally, it seems intuitively clear that intervention is only applicable in reasoning from causes to effects, as in (1a). Counterfactuals whose constituents are not related in this way, such as *backtracking* counterfactuals from effects to causes like (1b), would be equivalent to their consequents if they were interpreted by causal intervention in the above strictly local sense.

(1) a. If the sprinkler were on, the lawn would be wet.

b. If the lawn were wet, the sprinkler would be on.

That said, in extending the reach of the theory to cases like (1b), we believe that the right approach is not to pit causal intervention against alternative operations. Rather, we believe that speakers' judgments are always the result of an interplay between a number of tendencies, intervention among them, which may push in contradictory directions in particular cases. Viewed this way, much of the variation observed in speakers' judgments may be due to differences in the way they resolve such conflicts — ultimately differences in world knowledge, personal preferences, attention, and similar contextual factors. If this view is correct, then a better understanding of the relevant factors and the interaction between them is a crucial prerequisite in the quest for a comprehensive theory of counterfactual reasoning.

The present paper is a step towards that goal. Specifically, we investigated two forces which may counteract local causal intervention in speakers' reasoning about what would be the case if *A* were true: a desire to make the hypothesized truth of *A* plausible in the context of their overall beliefs, and an intrinsic preference to 'hold on to' certain facts more tightly than to others, independently of their role in the causal scenario. Both of these factors have an established place in cognitive science research under the headings of *causal explanation* and *fact mutability*, respectively. As stated above, we do not mean to pit these ideas against each other or against causal intervention; rather, we argue for a richer and more nuanced theoretical model of *selective intervention* which integrates these various factors. While more empirical research is needed before such a model can be fully developed, we offer a first semi-formal version and some empirical support in this paper.

The remainder of the paper is organized as follows. Section 1 gives a brief overview of the main strands of research feeding into this work and presents the notion of causal intervention as it has been proposed in the framework of causal networks. Section 2 discusses causal explanation and presents experimental results demonstrating its role in counteracting causal intervention. Section 3 discusses the nature and role of fact mutability. Section 4 concludes the paper with an outlook on directions for future work.

1 The problem of counterfactuals

For our purposes, counterfactuals are sentences of the form '*If had been A, would/might have been C*', formally represented as $A \Box\!\!\rightarrow C$ and $A \Diamond\!\!\rightarrow C$, respectively. Logically, the two operators are duals of each other. Here we are mostly concerned with the former. The semantic analysis of counterfactuals has been approached from many theoretical perspectives and in many formal frameworks. We cannot do justice here to the richness of approaches and insights produced by this work; for that, the reader is referred to the many survey books, papers and collections on the topic, including Bennett (2003), Edgington (1995), Roese and Olson (1995), among others. We only highlight some key ideas which form the background of our study, emphasizing the major themes common to all of them.

1.1 Ordering semantics

Virtually all semantic theories of conditionals are based on a simple intuition: A sentence '*If A, then C*' asserts that C follows if A is added to a certain body of premises. Ramsey (1929) introduced an early version of the idea for indicative conditionals.[1] This proposal was later adapted to counterfactuals by Stalnaker (1968), further developed by Lewis (1973), and became the foundation of what is now known as the Stalnaker/Lewis possible-worlds semantics of counterfactuals. Some differences aside, Stalnaker and Lewis both appeal to a notion of *comparative similarity* between possible worlds: Relative to an 'actual' world i of evaluation, this notion is formalized as an ordering relation on the worlds in the model. In Lewis's implementation, a counterfactual $A \mathrel{\Box\!\!\rightarrow} C$ is then true

[1] Ramsey's original formulation was phrased in terms of "degrees of belief in [C] given [A]" and is often considered the starting point of the thesis that the probabilities of conditionals are the corresponding conditional probabilities (Adams 1965, 1975, Jeffrey 1964, Lewis 1976, Stalnaker 1970 and many others; see also the surveys in Bennett 2003, Edgington 1995, Eells and Skyrms 1994). Recent psychological studies have furnished solid empirical support for that idea (Evans and Over 2004, Oaksford and Chater 2007, Over et al. 2007). It should be kept in mind, though, that Ramsey in his paper reserved the idea for indicative conditionals of the form 'if [A] will [C]' and took a quite different approach to counterfactuals. Nor is the probabilistic thesis nowadays as widely accepted for counterfactuals as it is for indicatives. For these reasons, and because a theory of the role of causal relations in the interpretation of counterfactuals does not crucially depend on a probabilistic reading of the Ramsey Test, we do not discuss that research in detail in this paper.

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' (Lewis 1973, p. 49). Assuming for simplicity that there is a set of A-worlds that are maximally similar to i, this means that the counterfactual $A \mathrel{\Box\!\!\rightarrow} C$ is true if and only if C is true in all of those maximally similar A-worlds. For instance, in evaluating our example (1a) at a world i at which the sprinkler is off, one would consider those worlds at which the sprinkler is on and that are otherwise as similar to i as this condition permits: The lawn is there, the laws governing gravity and the evaporation of water are as we know them, and so on. The counterfactual is true if the lawn is wet at all those 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 a preference for avoiding 'big' discrepancies over 'small' ones. Much of the subsequent work on analyzing counterfactuals in possible-worlds semantics is motivated by similar intuitions and can be viewed as attempts to make the notion of similarity more precise.

1.2 Premise semantics

A conceptually different but related approach is to interpret counterfactuals relative to sets of sentences.[2] Such sets may be interpreted in various ways, for instance as theories, belief states, or state descriptions of some mechanism. The move from the world of evaluation to hypothetical antecedent-worlds corresponds in this framework to the assembly of a new set of sentences, and the problem of determining similarity is mirrored by the question of which true sentences to keep and which ones to give up.

Goodman (1955) offered an early version of this approach. He noted that while counterfactuals $A \mathrel{\Box\!\!\rightarrow} C$ generally assert that some connection holds between A and C, it is rarely the case that the latter logically follows from the former. Thus even if it is true that '*if the match had been scratched, it would have lit*', the scratching does not in itself guarantee the lighting: In addition, oxygen has to be present, the match has to be dry, etc. 'The first problem' in the interpretation of counterfactuals, Goodman writes, 'is … to specify what sentences are meant to be taken in conjunction with the antecedent as a basis for inferring the consequent'.[3]

The problem is hard because speakers consistently exclude certain sentences for non-logical reasons. For instance, that the match did not light is no less a fact than that there was oxygen; yet it

[2] Some versions of Premise Semantics are spelled out in terms of sets of propositions, which are formally distinct from sentences. This difference does not affect the main ideas discussed in this section.

[3] Goodman's second problem — that of defining 'natural or physical or causal laws' — will not concern us in this paper. Following much recent work, we take causal relations to be given as part of the model.

does not follow that '*if the match had been scratched, it would have been wet*'. Vexed by facts of this sort, Goodman resorted to circularity: His rule calls for the selection of those true sentences that would not be false if the antecedent were true. Nevertheless, his suggestions inspired much subsequent work aimed at giving general criteria for the selection of sentences to add to the antecedent. In philosophy and linguistics, this line of thought is known as *Premise Semantics* (Kanazawa et al. 2005, Kratzer 1981, 1989, 2002, Pollock 1981, Rescher 1964, Veltman 1976, 2005). Lewis (1981) clarified the relationship between the early formal implementations of this approach and his own ordering semantics, showing that the two result in very similar semantic theories.

Similar ideas were developed in Artificial Intelligence. Ginsberg (1986) interprets counterfactuals relative to an imaginary 'large database describing the state of the world at some point'. Ginsberg assumes that this database contains both statements about specific facts and rules in conditional form. The rules of interpretation are set up in such as way as to minimize differences in facts and localize violations of rules, treating the latter as exceptions to existing rules rather than the workings of different ones.

The distinction between parochial facts and general rules or laws is a concern common to most theories in Goodman's tradition. It is also reflected in Lewis's emphasis on minimizing violations of law while maximizing correspondence in particular fact.

1.3 Causal networks

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. These developments originated in statistics and artificial intelligence and have since had a major impact in psychology and related disciplines (Glymour 2001, Pearl 2000, Spirtes et al. 1993).

The formal vehicle of choice in this area is that of *Causal Bayesian Networks*. A Bayesian Network is a directed acyclic graph whose vertices are labeled with variables. Each distribution of values over the variables corresponds to a possible state of the system represented by the model. Bayesian 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 topology of the network encodes independence assumptions between the variables in form of the *Markov Assumption*: The probability distribution over the values of a variable $\mathbb{A}$ is fully determined by the values of its immediate parents $pa(\mathbb{A})$[4,5]. A Bayesian Network is *causal* if all arrows are assumed to lead from causes to effects.

[4] We distinguish typographically between variables like $\mathbb{A}$ and their values $A, \bar{A}$. We assume for simplicity that all variables are Boolean.

[5] Pearl (2000) assumes that causation is deterministic; uncertainty is encoded as a probability distribution over a distinct set of 'exogenous' variables modeling our ignorance. In this case, the values of $pa(\mathbb{A})$ jointly determine the *value* of $\mathbb{A}$. Such an approach may be warranted in the applications for which Pearl intended his models, but we do not adopt it here. People's commonsense reasoning treats at least some processes as genuinely subject to chance. A theory whose goal is to model common-sense reasoning should respect these intuitions, regardless of whether the underlying processes are in fact deterministic or not (Hiddleston 2005, p. 639). Even

<FIGURE 1 GOES HERE>

Under this approach, the standard analysis of counterfactual reasoning about what would have been if some variable $\mathbb{A}$ had had value A involves a special kind of update, formally represented using the 'do' operator: The result of applying the operation $do(\mathbb{A}= A)$ to a network with vertex $\mathbb{A}$ results in a new network in which $\mathbb{A}$ has value A and all arrows leading into $\mathbb{A}$ are removed (see Fig. 1). In the modified network, the non-descendants of $\mathbb{A}$ are independent of $\mathbb{A}$, hence unaffected by standard algorithms of belief propagation. The intention is that $do(\mathbb{A}= A)$ represents an external *intervention* upon $\mathbb{A}$ which disrupts the causal process that normally determines $\mathbb{A}$'s value, so the fact that $\mathbb{A}= A$ does not warrant any inferences about its normal causes. Thus the update is maximally local in the sense that for all variables $\mathbb{X}$ which do not lie 'downstream' of $\mathbb{A}$ in the direction of causal influence, the counterfactual 'If $\mathbb{A}$ had had value A, $\mathbb{X}$ would (still) have its actual value' is invariably true.

For example, consider the network structure on the left in Figure 1 and suppose the interpretation and the distribution of values over the variables are as given in (2).

(2) a. $\mathbb{X} = \overline{X}$ '*It is not raining*'

more fundamentally, determinism is a special case of non-determinism, but not *vice versa*, and there is and possibly can be no proof that the world is deterministic. Any claim that the outcome of a coin toss is really determined by an array of circumstantial factors which we are merely not in a position to examine would be 'based on faith rather than on evidence' (Hausman 1998, p. 185).

b. $\mathbb{A} = \bar{A}$ *'Bill is not carrying an umbrella'*

c. $\mathbb{Y} = \bar{Y}$ *'The streets are not wet'*

(3) a. If Bill were carrying an umbrella, it would not be raining.

b. If Bill were carrying an umbrella, the streets would not be wet.

An intervention as shown in the right-hand side in the figure would verify both of the counterfactuals in (3). The underlying intuition here is that one cannot make it rain by giving Bill an umbrella.

It is an empirical question when and to what extent speakers' judgments about particular counterfactuals reflect such highly localized reasoning. Perhaps the reader will agree that the counterfactuals in (3a,b) are not clearly and unambiguously true in the situation described. These are the kind of data that this paper is about. There is no doubt that the method of blocking the flow of information from effects to causes captures an intuitively real asymmetry. At the same time, evidence from both philosophical considerations and psychological experiments suggests rather pursuasively that the *do* operator alone does not begin to capture the full richness of real-life counterfactual reasoning. We conclude this section by discussing two such arguments.

1.4 Hiddleston's synthesis

Hiddleston (2005) presents a version of Premise Semantics which incorporates key notions from the causal-networks approach. Some of the details of his proposal are not directly relevant to our main concern. We only outline his solution to one particular problem with the do operator which is relevant to our discussion below.

In some cases, the application of $do(\mathbb{A}= A)$ leaves too many variables untouched. For one thing, the model resulting from the intervention may describe an impossible state if the values of $\mathbb{A}$'s parents logically imply that $\mathbb{A} = \bar{A}$. But even when the values of $\mathbb{A}$'s parents only causally, not logically, force $\mathbb{A} = \bar{A}$, it is not clear that leaving those values unaltered is the right strategy. There is an important difference between an alternative model with $\mathbb{A}= A$ which is inconsistent with prevailing causal law, (i.e., where the conditional probability of $\mathbb{A}= A$, given the values of $\mathbb{A}$'s parents, is 0), and one which is merely unlikely (i.e., where the conditional probability is low but positive). To avoid the former, Hiddleston allows the hypothetical information that $\mathbb{A}= A$ to affect variables 'upstream' from $\mathbb{A}$ along deterministic links. In a network in which such links are the exception rather than the rule, these repercussions may still be contained to a small set of $\mathbb{A}$'s non-descendants.[6]

[6] Another problem Hiddleston addresses is that $do(\mathbb{A}= A)$ may leave too few variables untouched if the counterfactual assumption raises the probability of the actual value of a causal descendant of $\mathbb{A}$. (Hiddleston's Example #3 is a case in point.) Again considering a nondeterministic network, suppose $\mathbb{A} = \bar{A}$ has a descendant $\mathbb{C}$ = $C$and the conditional probability of the actual event $\mathbb{C}$ = C, given $\mathbb{A}= A$, is high but strictly less than 1. Under the standard approach, applying $do(\mathbb{A}= A)$ and

Hiddleston 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 there. Specifically, his 'stripped-down' version allows for nondeterministic causal laws and does not make the Causal Markov Assumption. Counterfactuals are evaluated relative to models consisting of a causal graph and an assignment of values to all variables. As a stand-in for the elusive notion of 'actual cause', he introduces a notion of *positive parents* of a variable, defined as those parents which have a 'direct positive influence' on its actual value. Technically, $\mathbb{Y}$ is a positive parent of $\mathbb{X}$ if the conditional probability of $\mathbb{X}$'s taking on its actual value, given that all of $\mathbb{X}$'s parents (including $\mathbb{Y}$) have their actual values, is strictly higher than the corresponding conditional probability in the event that $\mathbb{Y}$ alone among $\mathbb{X}$'s parents takes on a different value. (A closely related notion plays a central role in Gärdenfors's definition of causal explanation — see below.) Since this definition depends on the values in addition to the causal structure, the set of positive parents of $\mathbb{X}$ relative to the same causal structure may differ between different value assignments.

Instead of modeling the interpretation of a counterfactual with antecedent $\mathbb{A}= A$ by cutting links into $\mathbb{A}$, Hiddleston's rule involves a comparison between alternative 'A-models'. Among those,

recalibrating the probabilities in the modified network would introduce a low but positive probability for the event that $\mathbb{C} = \bar{C}$. But intuitions are that in such cases, the counterfactual $(\mathbb{A}= A) \mathrel{\Box\!\!\rightarrow} (\mathbb{C} = C)$ is true. In order to account for this, Hiddleston's rules ensure that the values of some descendants of $\mathbb{A}$ are held fixed.

his rule identifies a subset of ‘A-minimal’ or ‘minimally altered’ models. The counterfactual is true if and only if its consequent is true in all A-minimal models.

In a recent paper, Rips (2009) compares the predictions of both Causal Bayesian Networks and Hiddleston’s theory for counterfactual conditionals, and reports that neither provides a fully adequate account of counterfactual reasoning. He proposes a more constrained version of Hiddleston’s theory, in which a heuristic is applied to constrain the search space of A-minimal models.

Of particular interest for our present purposes is Hiddleston’s contention that $\mathbb{A}$’s parents are sometimes affected by the counterfactual update with $\mathbb{A}= A$[7]. Our experimental data lend empirical support to this claim. Importantly, though, we show that counterfactual updates may affect non-descendants even in cases in which the link into A from the parent in question is not deterministic.

1.5 Intervention vs. observation

[7] The need to allow for some “backward” propagation of counterfactual assumptions for the sake of consistency was noticed before (cf. for instance the “causal ramps” of Bennett 2003, Mårtensson 1999, and references therein). We highlight Hiddleston’s more recent work here because of its explicit connection to Causal Bayesian Networks.

Above in Section 1.3, we discussed the do operator as the formal embodiment of a particular kind of causal inference. Its main feature is that the manipulation of the variable in question only has consequences for that variable's causal descendants, while all inferences about its causes are blocked. In this sense, do encodes a radically *localized* response to the update. At the opposite ends stands conditionalization, the *global* standard update operation on Bayesian Networks, in which the entire network is recalibrated according to Bayes's Rule (Pearl 1988). Intuitively, these two operations are associated with different ways in which the manipulated variable may come to have its new value and in which the reasoning subject may be involved. The terms used to mark the difference include 'intervention' *vs.* 'observation' (Spirtes et al. 1993, Pearl 2000) and 'doing' *vs.* 'seeing' (Sloman and Lagnado 2005). Borrowing from a well-known textbook example (Pearl 2000), the difference can be linked to that between 'turning the sprinkler on' and 'finding it on': In the former case one would not draw any conclusions about the weather, whereas in the latter one would likely do so. The importance of the distinction and the psychological reality of the two modes of update is often emphasized in work on causal reasoning, learning, decision making, and counterfactual inference (Glymour 2001, Hagmayer and Sloman 2009, Sloman and Lagnado 2005, Steyvers et al. 2003, Waldmann and Hagmayer 2005, among others).

Often in the literature, the problem is presented as a dichotomy: A counterfactual question contextualized in a causal scenario will prompt people to reason either as interveners or as observers, and the empirical question is which factors drive this choice. However, at least in the case of counterfactual reasoning, this black-and-white view seems to be inappropriate. In an influential study, Sloman and Lagnado (2005) start by distinguishing the two opposing modes of 'seeing' and 'doing', and explore the prediction that 'a counterfactual event will not be treated as diagnostic of its normal causes', i.e., that participants' judgments will conform to the strict locality of the do

operator. Overall, Sloman and Lagnado find that while there is good evidence that participants are capable of reasoning by local intervention, they do not always do so, nor do their judgments fall neatly into a bimodal distribution. In concluding, however, Sloman and Lagnado are careful not to assert that deviations from the predictions of the *do* operator are *eo ipso* evidence of reasoning by observation. Rather, they suggest that participants may have some flexibility in conceptualizing interventions in particular cases: 'Representing intervention is not as easy as forcing a variable to some value and cutting the variable off from its causes. . . People are not generally satisfied to simply implement a *do* operation. People often want to know precisely how an intervention is taking place.'[8]

[8] In fact, Sloman and Lagnado did not test how participants respond to observations. Some stimuli did mention observation explicitly; (ia) is a case in point. Even that sentence is a counterfactual, however (about a *hypothetical* observation), and a difference in responses between (ia) and (ib) (see Sloman and Lagnado 2005, for details) is not surprising under the hypothesis that all counterfactuals are interpreted by intervention, since the two involve very different counterfactual premises.

(i) a. What is the probability that C would have happened if we observed that B did not happen?

b. What is the probability that C would have happened if B had not happened?

c. Someone intervened and made B false. What is the probability that C would be true?

Other stimuli, such as (ic), were not as clearly counterfactual: The indicative premise may induce a tendency to suppose that the intervention in question was observed. However, these cases had in

Partly motivated by those findings, in our investigation we do not set up an *a priori* dichotomy between intervention and observation, asking instead *what kind of* intervention is made in particular cases. As we discuss below, we believe that interventions may come in different degrees of locality, sparing selected non-descendants of the manipulated variable. Once such flexibility is allowed, we are no longer dealing with an all-or-nothing affair, and one can in principle conceive of observation as a limiting case of (radically non-local) intervention. Counterfactual hypotheses and true observations may tend towards opposite ends of the spectrum, but in this paper we are only interested in the former.

1.6 Interim summary

While this section could not do justice to the full range of approaches, it identified some ideas and problems common to them. There is a wide and growing consensus that causal independencies play a central role in the interpretation of counterfactuals; however, the question facing all accounts which address this fact is what exactly that role is. We believe with Hiddleston, Sloman and Lagnado and others that the *do* operator is a rather crude device, since the non-descendants of counterfactual antecedents are not always isolated from the intervention. In the next section, we discuss one hypothesis about what may trigger deviations from the strict do operation: the speakers' desire to find a causal explanation for the hypothesized truth of the antecedent.

common the curious mix of moods exhibited by (ic), in which the indicative premise is followed by a subjunctive question, which we believe may counteract such a tendency.

2 Causal explanation

There is abundant evidence that humans have a deeply entrenched inclination towards seeking and providing 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, it has been shown to be linked to a variety of cognitive processes, including prediction (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 causality. Typically, however, causality does play 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. To be sure, in many cases 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 is it the case that real-world events have clear, unequivocal causes and effects. In many cases the causal 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; see also the papers in Collins et al. 2004). Furthermore, the process of finding the focal cause may be heavily context-dependent (Einhorn and Hogarth 1986).

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 from a hypothesized effect to its causes, in answering questions like (4a).

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

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

But explanations are also likely to be implicitly involved in our evaluation of *forward* counterfactuals like (4b): Even a speaker who does not believe that the Iraq war was a direct effect of the 9/11 attacks may answer the question differently depending on his or her beliefs about what caused the attacks, what might have prevented them, and how whatever would explain their non-occurrence would have affected the events leading up to the war.

2.1 Gärdenfors (1988)

To make this a bit more precise, we turn to a simplified version of Gärdenfors's (1988) formal account of causal explanation. We adopt his notation in this section, using A and $\bar{A}$ as shorthand for the event that $\mathbb{A} = A$ and $\mathbb{A} = \bar{A}$, respectively. An *epistemic state* is a pair $K = \langle W, B \rangle$, where W is a set of possible worlds with a common domain of individuals and B a probability measure on subsets of W, representing degrees of belief[9]. The following three fundamental operations on belief states underlie Gärdenfors's theory of belief dynamics:

[9] In fact, Gärdenfors defines a belief state as a triple $\langle W, P, B \rangle$, where W and B are as above and P maps each world $w \in W$ to a probability measure P_w on sets of individuals (i.e., extensions of predicates) in w. The distinction between P and B allows Gärdenfors to represent beliefs about

(5) a. **Expansion**: K_E^+ is the result of learning E in K.

b. **Contraction**: K_E^- is the result of retracting E from K.

c. **Revision**: K_E^* is the result of learning E in K while making (minimal) adjustments to avoid contradiction.

Each of these operations is a function mapping pairs of belief states and sentences to belief states[10]. It is not clear that all three need to be taken as basic in a theory of belief change. For

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.

[10] The assumption that they are functions has two important consequences. The first is that all three are defined for any pair K, E. Thus for instance, expansion is defined even if E is believed to be false in K. In this case, K_E^+ is the *inconsistent* belief state, written as $K_\perp$. In contrast, the revision K_E^* maintains consistency in such a case by removing any beliefs from $\mathbb{K}$ which conflict with E, but otherwise staying as close to K as possible. This idea of minimal change is also behind the contraction of K by E. The second consequence of the functionality of the three operations is that there is always a *unique* output state. This is a hard requirement to meet: When E cannot be consistently added to K, there may be multiple ways to adjust K, each of which may be "minimal" in some sense. Consequently, Gärdenfors's goal is not to define the three operations, but to impose

instance, Gärdenfors is sympathetic to the idea that revision can be defined in terms of contraction and expansion according to the *Levi Identity* $K^*_E = [K^-_{\bar{E}}]^+_E$ (Gärdenfors 1988, p. 69); in words, the result of revising K with E is the same as that of first retracting $\bar{E}$ and then adding E. We return to this assumption below.

For a belief state $K = <W , B>$, each of the three operations is defined in terms of corresponding pointwise operations on the components W, B. For instance, K^+_E is the pair $K = <W^+_E , B^+_E>$, where W^+_E is the result of restricting W to those worlds at which E is true and B^+_E is obtained from B by conditionalization on E[11]. The definitions of notions like *explanation* and *cause* make crucial reference to the probabilistic component. We briefly discuss them in this section. It is important to keep in mind throughout that in Gärdenfors's theory the question whether one sentence counts as an explanation or a cause of another cannot be answered out of context, but only relative to a given belief state K. The evaluation of either question relative to K, however, crucially involves conditions on belief states other than K.

In order for a sentence C to count as an explanation for E relative to K, the following conditions must be met (p. 204-5). First, E should be held to be true in K, for otherwise the question what explains E would not arise or be hypothetical at best. On the other hand, *C* should *not* be held

certain constraints or "postulates" which any fully specified set of operations must meet in order to constitute a realistic model of the dynamics of belief.

[11] If $B(E) = 0$, then B^+_E is defined to be $B_\perp$, the "absurd" probability function assigning 1 to all sentences.

to be true. One rationale for this latter condition is that it rules out E as an explanation for itself. More broadly, it ensures that the explanans is relevant in K, in the sense that it is not already derivable from K (p. 178). The third condition is a hypothetical one in K: If E were not already held to be true, then C would raise E's probability. This is evaluated relative to K_E^-, the closest belief state to K in which E is not held to be true[12]. Formally, the condition is that $B_E^-(E|C) > B_E^-(E)$.

For illustration, consider again the variables $\mathbb{X}$, $\mathbb{A}$, $\mathbb{Y}$ in the network shown in Figure 1 above, interpreted as in (2'), and suppose that in K, $\bar{A}$ is believed but nothing is known about the other variables.

(2') a. $\mathbb{X}$ 'It is / is not raining'

b. $\mathbb{A}$ = $\bar{A}$ 'Bill is not carrying an umbrella'

c. $\mathbb{Y}$ 'The streets are / are not wet'

Under Gärdenfors's conditions, given reasonable background assumptions about the probabilistic dependencies between the variables, $\bar{X}$ ('It is not raining') is an explanation for Bill's not carrying an umbrella: Hypothetically assuming that nothing were believed about $\mathbb{A}$ (i.e., retracting this information from K), learning that it is not raining would make it more likely that Bill is not carrying an umbrella. For the same reason, however, $\bar{Y}$ ('The streets are not wet') would also qualify as an explanation for $\bar{A}$: Since nothing is known about the weather, learning that the streets are not wet

[12] Presumably, since otherwise K_E^- is as close to K as possible, C is not held true in K_E^- .

would make it more likely that it is not raining, hence that Bill is not carrying an umbrella. Thus $\bar{X}$ and $\bar{Y}$ count as explanations for $\bar{A}$.

But in order for an explanation C to be a *causal explanation* of E in K, it must in addition count as a cause of E. Gärdenfors has an analysis of this latter notion, but his definition comes into play here only indirectly, for two reasons.

The first reason is that according to Gärdenfors, C can only count as a cause of E relative to a belief state in which both C and E are taken to be true. Now, the assumption that C counts as an explanation for E in K implies that C is *not* taken to be true in K. Instead, the question whether C counts as a cause of E is evaluated hypothetically: If C were believed to be true, would it then count as a cause of E? According to Gärdenfors, the most natural way to test this is to add C to K, then ask whether it counts as a cause of E in the resulting belief state K_C^+.

Now, Gärdenfors does not simply require that C count as a cause of E in K_C^+ . This is the second and more important reason why his definition of that relation figures only indirectly in his treatment of causal explanation. Instead, he assumes that speakers' intuitions about causal relations are antecedently given and affect their reasoning in specific ways, so that his formal definition is already fulfilled just in case they think of C as a cause of E. More specifically, Gärdenfors assumes that *whenever C counts as a cause of E, retracting C removes the basis for the belief in E.* Formally, Gärdenfors's condition is that $[K_C^+]_C^- = K_E^-$. He then goes on to show that if this condition is met and C counts as an explanation of E in K, then C counts as a cause of E in K_C^+ according to his formal definition[13], hence as a causal explanation of E in K.

In our example, this means that while the belief that Bill is not carrying an umbrella would remain intact upon hypothetically assuming that it is not raining, it would not survive the subsequent retraction of that information: If it is believed that it is not raining and that Bill is not carrying an umbrella, then hypothetically giving up the former would also remove the belief in the latter. It is for that reason that dry weather counts as a causal explanation for Bill's umbrella-lessens in K. On the other hand, according to the theory, if it is believed that the streets are dry and that Bill is not carrying an umbrella, then hypothetically giving up the belief in the former does *not* remove the basis for the belief in the latter, hence the streets' being dry does not count as a causal explanation for Bill's not carrying an umbrella.[14]

[13] C counts as a cause of E in K if and only if (i) both C and E are held to be true in K; and (ii) C raises the probability of E relative to K_C^-, i.e., $B_C^-(E|C) > B_C^-(E)$. If $[K_C^+]_C^- = K_E^-$, this comes down to the condition that $B_E^-(E|C) > B_E^-(E)$, which is already satisfied since C counts as an explanation for E in K (see above).

[14] It is important to keep in mind here that we have a particular causal structure in mind, namely one in which Bill's behavior is only affected by the weather, not by the state of the streets. If we assume instead that Bill checked in the morning whether the streets were wet before deciding whether to take his umbrella, the causal model is different, and so are the predictions.

The idea that causal beliefs affect belief contractions is important for the purposes of this paper, and we return to it below. Generally, Gärdenfors maintains, in retracting a sentence C one gives up C's effects as well. In contrast, the contraction does not affect sentences that are merely correlated with C.[15]

Before moving on, we note that the degrees of belief relevant in determining whether C counts as a (causal) explanation for E also play a role in other parts of Gärdenfors's theory. If B_E^- is high (i.e., E is likely in K_E^-), then E does not require an explanation in K. An explanation is only required to the extent that E is *unexpected* or *surprising* according to K_E^-. The "surprise value" of E is inversely related to $B_E^-(E)$. In Gärdenfors's terms, an agent asking for an explanation expresses a *cognitive dissonance* between E and the rest of her beliefs.[16] Gärdenfors proposes a measure of the *explanatory power* of C relative to E in K, defined as the difference $B_E^-(E|C) - B_E^-(E)$, and

[15] One may question the generality of Gärdenfors's proposal that $[K_C^+]_C^- = K_E^-$ if C counts as a cause of E in K_C^+. For instance, it may well be that in K_C^+, C counts as a cause of other sentences besides E, all of which presumably would be given up in $[K_C^+]_C^-$. Our goal here is not to improve upon Gärdenfors's theory, however, and the simple tenets discussed here are sufficient to illustrate the ideas.

[16] 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_E^- with the explanans.

suggests that this measure may predict speakers' choices between alternative explanations (p. 185).[17]

2.2 Application to counterfactual reasoning in causal networks

If a sentence A functions as a counterfactual antecedent (i.e., if strictly speaking the counterfactual premise is '$\mathbb{A} = A$'), it is unlikely to be held true by the speaker; indeed, counterfactuals are typically used when the antecedent is believed to be false. The crucial step in evaluating a counterfactual with antecedent A is to construct an alternative belief state K_A^*, the result of revising K with A, definable in terms of the Levi Identity as $K_A^* = [K_{\overline{A}}^-]_A^+$. As mentioned above, we adopt the assumption that this operation is sensitive to causal relations: Any beliefs about $\mathbb{A}$'s causal consequences are given up.[18]

[17] Note that Gärdenfors' notions of goodness of explanation and explanatory power only refer to the conditional probability of E given C in relation to the unconditional probability of E. Notably missing is the prior probability of C. Chajewska and Halpern (1997) argue that this prior probability also plays a role in choosing between different alternative explanations. Whatever the shortcomings of Gärdenfors' account may be, however, the importance of the conditional probability of the explanandum given the explanans is doubtless an important factor in causal explanation. We focus on this conditional probability.

[18] See the discussion of Hiddleston in Footnote 6 above for arguments that this strategy may not be applicable in complete generality in the interpretation of counterfactuals.

Most important for us is the question whether A requires an explanation in K_A^*, and if so, how speakers go about looking for an explanation. The first is a matter of "cognitive dissonance": An explanation is needed if or to the extent that A is unlikely in$[K_A^*]_A^-$. Given our assumptions, this is a state in which neither A nor $\bar{A}$ is known and in which nothing is believed about $\mathbb{A}$'s causal descendants. Suppose, then, that A is unlikely in this state, thus an explanation for A is called for in K_A^*. Our contention is that in this case the speaker does not base his hypothetical reasoning on K_A^*, but will instead introduce further revisions to arrive at a state K_A^{**} that is similar to K_A^* but in which A does not require an explanation. It is fair to assume that speakers will first look for an explanation among the parents of $\mathbb{A}$ in the causal network. Formally, this requires first suspending any beliefs about the (actual) values of those parents that are carried over from K to K_A^*, then asking which values would reduce the need for an explanation for A.

For concreteness, consider first the case that $\mathbb{A}$ has only one parent, as in the *chain* shown in Fig. 2 or the *fork* underlying the illustration in Fig. 1. Whether an explanation is needed in K_A^* depends on the conditional probability of A, given the value of $\mathbb{A}$'s parent $\mathbb{X}$. In fact, the Markov Assumption ensures that this is the only relevant probability. If it is high enough for the surprise to be tolerable, no explanation is required and the value of $\mathbb{X}$ remains untouched. Otherwise, the comparative explanatory power of $\mathbb{X}$'s possible values determines what value it receives in the alternative model K_A^{**}. Now, deciding on a value for $\mathbb{X}$ cannot be the end of the story. It raises a similar question about $\mathbb{X}$'s parent $\mathbb{Z}$, and so on up the chain, for "cognitive dissonance" can arise anywhere in the causal network if a variable has a value that is very unlikely given the values of its parents. We assume that the decision of how to treat $\mathbb{X}$'s parent is made by applying the above decision procedure again, testing for cognitive dissonance and, if necessary, setting $\mathbb{Z}$ to the value with the greatest explanatory power. The same procedure is applied recursively to the ancestors of $\mathbb{Z}$ until one is left intact because the cognitive dissonance has become tolerable. Clearly this idea raises the question as to how exactly explanatory power is calculated when multiple variables are involved, and whether it is somehow offset by a 'cost' for the increasing non-locality of the update. These questions are interesting both theoretically and empirically, but we set them aside in this paper.

<FIGURE 2 GOES HERE>

The case in which $\mathbb{A}$ has multiple parents, as in the *collider* topology shown in Fig. 2, is not fundamentally different. Here, instead of merely asking whether or not to leave the parent unaffected, the question is *how many* and *which* of them to keep. If the values of the parents jointly make $\mathbb{A}$ surprising, an explanation is required and involves changing the value of at least one of the parents. This may be sufficient; but whether it is depends on the agent's tolerance for cognitive dissonance. If no change to one parent provides enough relief, then all pairs of parents are considered, and so on, up to the entire set of parents of $\mathbb{A}$.[19]

[19] Notice that this predicts that the agent will not consider changing two parents if she can achieve sufficient cognitive relief by changing only one. An alternative possibility would be to compare all possible combinations of values of $\mathbb{A}$'s parents and choose the one with the highest explanatory

In the graph-theoretic terminology of causal networks, leaving all parents' values intact corresponds to the application of the do operator, cutting all links into $\mathbb{A}$. The decision to change some but not all of the parents' values roughly amounts to a selective application of the operator, leaving the links from some parents intact and thus allowing $\mathbb{A} = A$ to affect those parents.[20] In the following sections, we will at times employ the language of causal networks and describe our results in terms of a decision as to whether or not to cut links. Generally, cutting a link from a parent amounts to leaving its value intact; leaving a link in pace amounts to letting the update affect the parent.

We have outlined an account of the interpretation of counterfactuals whose predictions differ from the simple invocation of the do operator in the theory of Causal Networks. In the following subsection, we use the latter as our baseline and discuss experimental results for scenarios in which the two theories make different predictions.

power. In dismissing the latter possibility, we predict a strong preference for leaving parents unchanged, thus keeping K_A^{**} as close as possible to K_A^{*}. Whether this prediction is empirically borne out will have to be determined in future studies.

[20] In a nondeterministic network, the correspondence is only approximate. Letting $\mathbb{A} = A$ propagate up the link may shift the probabilities of the various values of the parent, but does not enforce a choice. Our adoption of Gärdenfors's framework implies that one of the parents' values is chosen.

2.3 Experiments on causal explanation

The following experiments involve scenarios which contain facts with different frequencies of occurrence or with different causal strengths. We investigate how these differences affect subjects' responses to counterfactual questions. Our focus throughout the experimental sections of this paper is on the comparison between participants' responses with the prediction of Causal Bayesian Networks and our theory of selective intervention outlined above.

2.3.1 Experiment 1

In the first experiment, we examine whether in collider and fork topologies the explanatory power of causes (defined in terms of conditional probability) affects the interpretation of counterfactuals whose antecedent calls for a revision of an effect. Sloman and Lagnado (2005) suggest that when the effect is explicitly mentioned in the antecedent, its causes are more likely to be left unchanged (i.e., the links into the effect are likely to be cut). If there is indeed such a tendency, then causes should be particularly resistant to revision in stimuli with this property. Therefore, if causes are in fact affected in the interpretation of such stimuli, this would constitute evidence in favor of our theory.

We predict that in a collider, the cause with the highest explanatory power for the effect will be affected (i.e., the link will be preserved), while the other causes remain intact (i.e., those links will be severed). Our stimuli mention one of the causal parents in the consequent, thus participants' responses give us direct evidence as to whether that parent is affected or not. In a fork network, the stimuli are counterfactuals which refer to effects in both the antecedent and the consequent. Here the observation is indirect, allowing us to infer whether the common cause is affected. Our stimuli describe binary forks in which one link is deterministic and the other is of varying causal strength (as measured by the conditional probability). In this way we can examine whether the treatment of the parent depends solely on the qualitative difference between determinist and non-deterministic links as Hiddleston would have it, or whether it is also affected by the gradient measure of causal strength á la Gärdenfors.

2.3.1.1 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’. Logical abbreviations in parentheses are added here for the reader’s convenience; they were not shown to the subjects. Each participant randomly received either version (A) or version (B) of each of the scenarios. Note that since the assignment was random, the (A) or (B) groups were different sets of participants in each scenario.

Scenario 1 (Collider)

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? $(C \mathrel{\Box\!\!\rightarrow} B)$

Scenario 2 (Fork with weak link)

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? $(C \mathrel{\Box\!\!\rightarrow} B)$

Scenario 3 (Fork with strong link)

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? ($B \Box\!\!\rightarrow C$)

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

2.3.1.2 Results. In Scenario 1, the mean for $C \Box\!\!\rightarrow A$ was 3.36 while the mean for $C \Box\!\!\rightarrow B$ was 6.13. The difference between the two questions was highly significant ($F(43) = 8.65$, $p < 0.01$). In Scenario 2 there was a highly significant difference between the means for $B \Box\!\!\rightarrow C$ and $C \Box\!\!\rightarrow B$ which respectively were 8.38 and 4.00 ($F(45) = 22.14$, $p < 0.01$). In Scenario 3, the mean for $B \Box\!\!\rightarrow C$ was 7.04 while the mean for $C \Box\!\!\rightarrow B$ was 3.68. The difference between the two questions was highly significant ($F(47) = 9.78$, $p < 0.01$). There was also a difference in responses to $B \Box\!\!\rightarrow C$ across Scenarios 2 and 3 ($F(54) = 2.78$, $p = 0.10$).

2.3.1.3 Discussion. First, some discussion is in order regarding the nature of the stimuli. One might argue that the scenarios are too underspecified to support conclusions about our theory. To see this, consider for instance Scenario 1 and note that it specifies only the conditional probabilities of Ball C's moving given Ball A's moving and given Ball B's moving, and stipulates that all three balls moved. According to our hypothesis, the counterfactual premise that C did not move calls for an explanation; but to evaluate the explanatory power of a possible explanation, say, of A's not moving, one would need the conditional probability of C's not moving given that A does not move and B does. This conditional probability is not given and cannot be calculated from the scenario. Now, while this is indeed a potential problem, we nevertheless consider our setup both unavoidable and sound. It is unavoidable because a full specification of all probabilities would have added artifice and complexity to this already rather abstract scenario. It is sound because (or to the extent that) participants tend to make the implicit 'closed- world' assumption that Ball C's movement has no other causes than the ones mentioned, combined with the common-sense assumption that Ball A's or B's being at rest would not cause Ball C to move. With these assumptions in mind, we turn to our interpretation of the data.

In the collider network in Scenario 1, the baseline account in terms of the *do* operator predicts that the answers to both of the questions in the first scenario should be 'Yes' (10), as intervening on $\mathbb{C}$ 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 suggests, first of all, that participants did not simply cut all links. This result is particularly interesting in view of the fact that neither of the links was presented as deterministic, thus a state of affairs in which both A and B move while C does not move is not impossible. Secondly, we see a clear preference for leaving the $\mathbb{A} \rightarrow \mathbb{C}$ link intact and cutting $\mathbb{B} \rightarrow \mathbb{C}$ (i.e., letting the manipulation affect $\mathbb{A}$ but not $\mathbb{B}$). This is predicted by our hypothesis that participants look for an explanation. In this scenario, Ball A's not moving provides a much stronger explanation than Ball B's not moving would.

For Scenarios 2 and 3, again the *do* operator would cut the link from the common cause, leaving its value intact, thus the sister of the antecedent variable should not be affected and the answer should be 'Yes' (10) in all cases. The results suggest, however, that the subjects left the common cause $\mathbb{A}$ intact when $\mathbb{B}$ was manipulated, but not when $\mathbb{C}$ was - in other words, they cut the $\mathbb{A} \rightarrow \mathbb{B}$ link but not the $\mathbb{A} \rightarrow \mathbb{C}$ link. In both scenarios, the differences between these two links is that the associated dependencies are non-deterministic and deterministic, respectively. The fact that the deterministic link was left intact strongly suggests a reluctance to consider counterfactual alternatives which are causally impossible. Interestingly, the tendency to cut the non-deterministic link was related to the strength of the probabilistic dependency (5% and 95%, respectively).

This latter result suggests that the degree of surprise plays an important role in determining whether causal parents are affected: If Ball B's not moving is unlikely, a stronger adjustment is made to the status of Ball A. Thus overall, these results show that the gradient measure of surprise drives the interpretation in addition to the qualitative question of causal legality.

2.3.2 Experiment 2

In the second experiment, we use two scenarios which include causes with different strengths. Here, however, the strength of the causal dependence is not specified numerically, but to be inferred from world knowledge. The first scenario was designed as a causal *chain* structure with three variables. As discussed in Section 2.2 above, we hypothesize that when changing the value of the parent of the antecedent variable is not sufficient as an explanation, the search for an explanation will continue recursively to the grandparents, and so on. We also assume that this spreading of the impact terminates at some point up the chain. In this experiment we examined this latter assumption, asking whether there was any evidence of a difference in treatment between causal parents and causal grandparents. The second scenario is again a *collider*, similar in structure to the one in Experiment 1.

2.3.2.1 Method. 36 Northwestern undergraduate students participated in this experiment. The methods and materials were similar to Experiment 1.

Scenario 4 (Chain)

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.

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

(B) If the boat had not sunk, would the captain have decided to search for the child? ($C \,\Box\!\!\rightarrow A$)

Scenario 5 (Collider)

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.

(A) If the boat had not sunk, would they have found the child? ($C \mathrel{\Box\!\!\rightarrow} B$)

(B) If the boat had not sunk, would they have found the cook? ($C \mathrel{\Box\!\!\rightarrow} A$)

2.3.2.2 Results. In the first scenario, the mean for $C \mathrel{\Box\!\!\rightarrow} A$ was 6.20 and the mean for $C \mathrel{\Box\!\!\rightarrow} B$ was 4.5, which came out to be the significantly different ($F(1, 35) = 12.10$, $p < .05$). In the second scenario, the mean for $C \mathrel{\Box\!\!\rightarrow} A$ was 6.5 while the mean for $C \mathrel{\Box\!\!\rightarrow} B$ was 7.0. Within-subject test revealed that the difference between the two questions was significant ($F(1, 35) = 7.11$, $p < .05$).

2.3.2.3 Discussion. As before, a theory based on the *do* operator would predict 'Yes' (10) in all cases. Both scenarios mentioned explicitly that a few additional pounds would be enough to sink the lifeboat, therefore the child in Scenario 4 and both the cook and the child in Scenario 5 were potential causes for the sinking of the boat. The first scenario was a chain $\mathbb{A} \rightarrow \mathbb{B} \rightarrow \mathbb{C}$, where the variables corresponded to the captain's decision whether to search, the finding of the missing child, and the sinking boat, respectively. There was a significant difference between the treatment of $\mathbb{C}$'s parent $\mathbb{B}$ (which was more likely to be affected, suggesting that the link was left intact) and that of $\mathbb{C}$'s grandparent $\mathbb{A}$ (which more likely to be unaffected, suggesting that the link was cut). This constitutes evidence that while the counterfactual premise required a revision of the parent, this effect did not spread further up the causal chain.

In Scenario 5, the results show a clear preference for changing the value of $\mathbb{A}$ (the cook was found) rather than $\mathbb{B}$ (the child was found). This corroborates our finding from Experiment 1 that participants were more likely to leave the link from the causal parent with the highest explanatory power intact, thus adjusting the value of that parent to provide an explanation.

3 Fact mutability

In the previous section, we discussed cases in which the decision as to how a counterfactual hypothesis affects the parents of the manipulated variable depends on the explanatory power of those parents. This is a relational notion, in the sense that a parent's explanatory power is not an intrinsic property, but determined relative to $\mathbb{A}$. In this section, we turn to another factor affecting the interpretation of counterfactuals, one which is a property of the parent itself, or rather of the parent's having its actual value. The basic idea is that even among beliefs held with equal certainty, some are given up more easily than others. Preferences of this kind have been explored in various contexts. We start this section by giving a brief outline of two major approaches, Norm Theory (Kahneman and Miller 1986) and Epistemic Entrenchment (Gärdenfors 1988). We then discuss how the notion is incorporated into an account of counterfactual reasoning, and give experimental evidence that speakers' judgments are indeed sensitive to it.

3.1 Norm theory

The central idea behind Kahneman and Miller 's (1986) Norm Theory is that the experience of a stimulus (or an event) triggers the retrieval of other relevant experiences constituting a 'norm', which is then used as a basis for assessing the 'normality' of the triggering stimulus. Though based on memory traces, the norms are constructed ad-hoc and, as such, are sensitive to the context in which the stimulus appears. Therefore they cannot be equated to pre-existing categories or concepts. In some cases there may be a significant discrepancy between the norm and the triggering event. For example, when attending a friend's wedding, one will construct a norm aggregating different wedding experiences from the past. If the wedding in question is on board an airplane, it is likely that the constructed norm to which the experience is compared is very different from the actual experience of an airborne wedding. Such activation of a norm that is different from the stimulus leads to the construction of alternatives to reality and triggers counterfactual thinking.

To explain how and when norms resemble or differ from target stimuli, Norm Theory appeals to the idea of *mutability* of attributes. The constructed norm shares the immutable attributes with the stimulus, for example the presence of a bride, but may differ with regard to the mutable ones, such as the place of the wedding. The mutability of the features that describe an event determines both the availability and the content of counterfactual thinking. The presence of mutable attributes facilitates reasoning about alternatives to reality; and when engaged in a mental search for 'what would normally be the case', alternatives which differ from reality only with respect to highly mutable attributes are likely to be considered first. We believe this is the case whenever the difference in mutable attributes provides a satisfactory explanation for why reality differs from the norm. If one of the guests of the airborne wedding faints, for example, we could easily construct an alternative wedding in a church at which the guest does not faint. The difference in location offers a satisfactory explanation for the fainting, thus alternatives which differ in less mutable respects (i.e., the presence or absence of a best man) are not considered.

On the question of which psychological factors determine mutability, Kahneman and Miller, borrowing from Appley (1971) and Helson (1964), appeal to some of the factors influencing the settings of adaptation level. In addition, Kahneman and Tversky (1982) report that actions are more easily undone than inactions. Furthermore, actions of a focal agent, and particularly the self, are more easily undone than the actions of a background agent (Kahneman and Miller 1986), and exceptions are more mutable than routines.

3.2 Epistemic entrenchment

Another influential theoretical approach to modeling the relative ease with which beliefs are retained or given up is Gärdenfors's (1988) notion of *Epistemic Entrenchment*. Recall from Section 2.1 that the construction of a hypothetical belief state K_A^* from K involves a *contraction* $K_{\bar{A}}^-$, giving up the belief in A's falsehood. The result of this operation is a maximal subset of K which is consistent with A. There is in general no unique such subset. The role of epistemic entrenchment is to guide the choice between them: The sentences with the lowest entrenchment are given up first (p. 87). Importantly, the point of appealing to this measure in addition to probabilistic ones is that two beliefs may differ in epistemic entrenchment even if both are held with certainty (see also Levi 1983).

Gärdenfors goes on to explore the logical consequences of various postulates for the formal implementation of epistemic entrenchment. Rather than reproducing this work here, we only add a couple of comments. First, while Gärdenfors assumes that epistemic entrenchment is at bottom a *property* of sentences, not a relation between them, formally its main role is to impose an order of *relative* entrenchment between sentences. Since his main interest is in the logical properties of this order, he is able to largely sidestep the question of how epistemic entrenchment itself is to be measured or what factors might contribute to it. Secondly, epistemic entrenchment is fixed only relative to a given epistemic state, thus two epistemic states with the very same beliefs and probabilities may differ in this regard. Consequently, Gärdenfors notes, if epistemic entrenchment is at work in driving contraction functions, it may lead to diverging opinions about explanations (p. 188) and causality (p. 198) relative to otherwise equivalent belief states. Since both notions are crucial in the interpretation of counterfactuals (or so we argue), so is epistemic entrenchment.

3.3 Application to counterfactual reasoning

While Norm Theory is mostly concerned with spontaneous responses to external stimuli and the interaction between memory and perception, Gärdenfors is interested in the dynamics of belief change, citing changing scientific theories, logical databases and legal codes as applications. Despite the differences in motivation and intended application, however, both seem to highlight different aspects of the same phenomenon: that beliefs which are held with the same firmness may nevertheless be treated differently in counterfactual reasoning.

Specifically, we hypothesize that fact mutability plays an important role in guiding people's choices between alternative explanations, thus ultimately determining the fate of the manipulated variable's causal parents. As in Section 2.2, we assume that an explanation is needed when the counterfactual premise $\mathbb{A}= A$ is surprising or unexpected, given the values of $\mathbb{A}$'s parents. While the two theories treat surprise differently, both emphasize the discrepancy between a (hypothetical) update and the agent's expectations. Norm theory does not specify a formal mechanism for assessing the relative goodness of alternative explanations. Gärdenfors's causal explanation specifies such a mechanism, but does not incorporate the full range of psychological factors that have been found to play a role. We hypothesize that highly mutable parents make for more plausible, thus 'better ' explanations, then less mutable ones.[21] In terms of interventions in causal networks, the links to the effect from highly mutable causes remain intact, while the other links are cut.

It remains an open empirical question how exactly mutability interacts with explanatory power in ranking alternative explanations. It is possible that one only comes into play when the other does not yield a clear preference, or that both enter the calculation equally. What our experiments do show is that factors which influence mutability also play a role in the interpretation of counterfactuals.

3.4 Experiments on fact mutability

[21] It is tempting to think of fact mutability as a property of variables, rather than the events of those variables having certain values. It is not clear that the former would be correct, however, in view of some of the factors found relevant in Norm Theory: If the difference between action and inaction comes down to a difference between the values of a variable, then mutability cannot be a property of that variable itself. Thus our talk here of 'mutable parents' should be understood as referring to the values of those parents.

As outlined above, Norm Theory makes predictions about a variety of factors affecting mutability. These go beyond Gärdenfors's probability-based notion of causal explanatory power to include fine details about the way an event is presented. Thus actions are predicted to be more mutable than inactions, and actions by the focal agent (the person from whose perspective the event is depicted in our scenarios) than those by a background agent. In this section, we examine whether these two factors do indeed play a role in people's choice between alternative explanations (Experiment 3) and in answering counterfactual questions (Experiment 4).

3.4.1 Experiment 3

The third experiment is concerned with the role of mutability in choosing between alternative explanations. We present two scenarios, each describing a sequence of events, and each followed by an alternative outcome with two possible explanations for why the outcome might have been different. In terms of causal relations, both scenarios describe *collider* topologies. In Scenario 6, one of the alternative outcomes reverses an omission by the agent, whereas the other reverses an action. In Scenario 7, the two outcomes differ in reversing an action by a background agent vs. one by the focal agent. We predict that when asked to choose between the two explanations, participants will pick the second one in each scenario (action and focal agent) more often than the first one (inaction and background agent), since the former refer to factors with a higher mutability rate.

3.4.1.1 Method. 12 Northwestern undergraduate students were presented with the following two scenarios. They were asked to choose either option (A) or option (B) as the best explanation for a hypothetical different outcome.

Scenario 6

You need 99 out 100 to get an A in a test. All the questions are in the form of True/False questions. You have no problem answering the first 98 questions, and you are sure that you've answered them all correctly. But the last two questions happen to be very difficult. First you circle True on both questions, but just before you hand in your exam, you change your answer to question 99, by circling False. When you get the results of your test, you realize that your answer to both question 99 and 100 were wrong, resulting in you not getting an A.

> Suppose you had gotten an A. What would be the best explanation for that?
>
> (A) You got an A because you did not stick to your answer on question 100.
>
> (B) You got an A because you did not change your answer on question 99.

Scenario 7

You play a tennis match in your university's tennis tournament. You lose on the tiebreaker when your opponent serves an exceptionally good serve which you return with a backhand. You have a weak backhand and the ball ends up in the net.

Suppose you had won the tie-breaker. What would be the best explanation for that?

(A) You won the tie-breaker because your opponent did not serve an exceptionally good serve.

(B) You won the tie-breaker because you returned the ball with your forehand.

3.4.1.2 Results. In both scenarios, participants chose the second option as the best explanation more often than the first option (binomial test, Scenario 6: $p < 0.01$, Scenario 7: $p < 0.05$).

3.4.1.3 Discussion. In the above scenarios, option (B) refers to the more mutable cause compared to option (A). In both cases, participants chose the more mutable cause as the best explanation for the alternative outcome. This provides support for the connection between mutability of facts and the goodness of explanations referring to those facts.

3.4.2 Experiment 4

The previous experiment established a connection between the mutability of facts and explanatory goodness. In this experiment, we use the same scenarios to investigate the difference in how participants answer counterfactual questions about facts with different mutability rates. Given that actions are more easily reversed than inactions in searching for explanations, we predict that the former will be affected in counterfactual reasoning more often than the latter. Similarly, given that actions of the main agent are more easily reversed than those of a background agent, we predict that the former will be affected more often than the latter.

3.4.2.1 Method. 69 Northwestern undergraduate students were presented with the following scenarios, and after each scenario they were asked to evaluate the likelihood of two counterfactual statements. The questions were presented on a computer screen, and participants were asked to rate the likelihood of each question from 0 to 10, 0 being ‘definitely no’ and 10 being ‘definitely yes’. Each participant randomly received either the first or the second question of each of the scenarios.

Scenario 8

You need 99 out 100 to get an A in a test. All the questions are in the form of true or false questions. You have no problem answering the first 98 questions, and you are sure that you’ve answered them all correctly. But the last two questions happen to be very difficult. First you circle True on both questions, but just before you hand in your exam, you change your answer to question 99, by circling False. When you get the results of your test, you realize that your to answer to both question 99 and 100 were wrong, resulting in you not getting an A.

(A) If you had received an A, would you have stuck with your answer on question 100? ($C \mathrel{\Box\!\!\rightarrow} A$)

(B) If you had received an A, would you have changed your answer on question 99? ($C \mathrel{\Box\!\!\rightarrow} B$)

Scenario 9

You play a tennis match in your university's tennis tournament. You lose on the tie-breaker when your opponent serves an exceptionally good serve which you return with a backhand. You have a weak backhand and the ball ends up in the net.

(A) If you had won the tie-breaker, would your opponent have served an exceptionally good serve? ($C \Box\!\!\rightarrow A$)

(B) If you had won the tie-breaker, would you have returned the ball with your forehand? ($C \Box\!\!\rightarrow \bar{B}$)

3.4.2.2 Results. In the first scenario, the mean for $C \Box\!\!\rightarrow A$ was 5.91 while the mean for $C \Box\!\!\rightarrow B$ was 4.12. The difference between the two questions was significant (*F*(1, 66) = 2.09, *p* < .05). In the second scenario, the mean for $C \Box\!\!\rightarrow A$ was 5.65 while the mean for $C \Box\!\!\rightarrow \bar{B}$ was 6.62; the difference was significant (*F*(1, 66) = 3.29, *p* < .05).

3.4.2.3 Discussion. The results suggest that subjects more often changed the value of $\mathbb{B}$ than $\mathbb{A}$. In terms of causal networks, subjects were less likely to cut the link $\mathbb{B} \rightarrow \mathbb{C}$ than $\mathbb{A} \rightarrow \mathbb{C}$. We believe that this is due to the fact that in both scenarios $\mathbb{A}$ was more mutable than $\mathbb{B}$, therefore changing $\mathbb{A}$ provided a better explanation for the hypothetical premise involving $\mathbb{C}$. We note again that the do operator would predict that in a collider topology the value of the parents should not be affected by an intervention on the effect, thus the answer to both questions should be 'Yes'. Thus our results provide evidence against the application of the do operator and for the hypothesis that parents are affected (i.e., links are left in place) selectively.

4 Conclusions

In this paper we explored instances in which the radical notion of intervention seems to be too strong; in some cases, too many variables are left untouched by $do(\mathbb{A}= A)$. Furthermore, we argued that although intervention may not necessarily remove all the causal dependencies between the antecedent of the counterfactual and its immediate causes, how people choose the location of intervention is related to different psychological factors. Two related factors were discussed which play a role in determining whether and which causal parents of manipulated variable are affected: Speakers' need for an explanation of the hypothesized truth of the antecedent of the counterfactual, and an intrinsic preference to keep certain facts constant over others, independently of the role played by those facts in the causal scenario.

We offered an informal theoretical model of selective intervention which integrates these various factors. This approach consists of a modification to Pearl's intervention operator. As before, $do(\mathbb{A}= A)$ involves forcing the variable $\mathbb{A}$ to have the value A. However, rather than cutting all causal links into $\mathbb{A}$ and thus blocking any consequences of the intervention for $\mathbb{A}$'s non-descendants, we proposed an intervention in which the links are cut selectively following an analysis of the possible causal explanations for the hypothesized event that $\mathbb{A}= A$ and the mutability rates of the parents of $\mathbb{A}$. This alternative approach was used to evaluate a series counterfactual conditionals showing that selective intervention does occur and causal explanations and mutability of facts seem to play a role in this selection. The two factors discussed in this paper may not be the only forces which play a role in the theory of selective intervention. As discussed earlier, we believe the way forward for understanding and modeling human counterfactual reasoning is in identifying the range of factors which influence it and capturing the interactions between those factors. More empirical research is needed before the factors are identified and the model can be fully developed.

Morteza Dehghani
Institute for Creative Technologies
University of Southern California

Rumen Iliev
Department of Psychology
Northwestern University

Stefan Kaufmann
Department of Linguistics
Northwestern University

Acknowledgments

This research was supported in part by a grant from an AFOSR MURI & AFOSR FA9550-09-1-0507. The third author thanks the Japan Society for the Promotion of Science (JSPS, Project ‘Inferential mechanisms and their linguistic manifestations’), the American Council of Learned Societies (ACLS, Project ‘Speaking of time and possibility’), and the Lichtenberg-Kolleg at Georg-August-Universität, Göttingen, for support of the theoretical work.

Address for correspondence: Morteza Dehghani, Institute for Creative Technologies, University of Southern California, 12015 Waterfront Drive, Playa Vista, CA 90094, USA.

email: morteza@ict.usc.edu

References

Adams, E. (1965). The logic of conditionals. *Inquiry*, 8:166–197.

Adams, E. (1975). *The Logic of Conditionals*. Reidel.

Appley, M. H. (1971). *Adaptation Level Theory*. Academic Press, New York.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford University Press.

Braithwaite, R. B., editor (1931). *The Foundations of Mathematics and other Logical Essays*. Routledge and Kegan Paul.

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.

Edgington, D. (1995). On conditionals. *Mind*, 104(414):235–329. Eells, E. and Skyrms, B., editors (1994). *Probabilities and Conditionals: Belief Revision andRational Decision*. Cambridge University Press.

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

Evans, J. S. and Over, D. E. (2004). *If*. Oxford University Press, Oxford, UK.

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

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

Glymour, C. (2001). *The Mind'sArrows: BayesNets andGraphical CausalModels in Psychology*. The MIT Press.

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.

Hagmayer, Y. and Sloman, S. A. (2009). Decision makers conceive of their choices as interventions. *Journal of Experimental Psychology: General*, 138:22–38.

Halpern, J. Y. and Pearl, J. (2005). Causes and explanations: A structural-model approach. *British Journal of Philosophy of Science*, 56:843–887, 887–911.

Hausman, D. M. (1998). *Causal Asymmetries*. Cambridge University Press.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. John Wiley & Sons., New York.

Helson, H. (1964). *Adaptation-level Theory: An Experimental and Systematic Approach to Behavior*. Harper & Row, New York.

Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, 39(4):632–657.

Jeffrey, R. C. (1964). If. *Journal of Philosophy*, 61:702–703.

Kahneman, D. andMiller, D. T. (1986). Normtheory: Comparing reality to its alternatives. *Psychological Review*, 93:136–153.

Kahneman, D. and Tversky, A. (1982). *The Psychology of Preferences*. W.H. Freeman.

Kanazawa,M., Kaufmann, S., and Peters, S. (2005). On the lumping semantics of counterfactuals. *Jounal of Semantics*, 22:129–151.

Kaufmann, S. (2004). Conditioning against the grain: Abduction and indicative conditionals. *Journal of Philosophical Logic*, 33(6):583–606.

Kaufmann, S. (2005). Conditional predictions: A probabilistic account. *Linguistics andPhilosophy*, 28(2):181–231.

Kaufmann, S. (2009). Conditionals right and left: Probabilities for thewhole family. *Journal of Philosophical Logic*, 38:1–53.

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.

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

Kratzer, A. (1989). An investigation of the lumps of thought. *Linguistics and Philosophy*, 12:607–653.

Kratzer, A. (2002). Facts: Particulars of information units? *Linguistics and Philosophy*, 25:655–670.

Levi, I. (1983). Truth, fallibility, and the growth of knowledge. In Cohen, R. S. and Wartofsky, M. W., editors, *Boston Studies in the Philosophy of Science*, volume 31, pages 153–174. Reidel, Dordrecht.

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

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, 85:297–315.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, 13:455–76.

Lewis, D. (1981). Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic*, 10:217–234.

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

Mårtensson, J. (1999). *Subjunctive Conditionals and Time: A Defense of the Classical Approach*. Number 10 in Acta Philosophica Gothoburgensia. Department of Philosophy, University of Göteborg.

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.

Oaksford, M. and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford University Press.

Over, D., Hadjichristidis, C., Evans, J., Handley, S., and Sloman, S. (2007). The probability of causal conditionals. *Cognitive Psychology*, 54:62–97.

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

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

Pollock, J. L. (1981). A refined theory of counterfactuals. *Journal of Philosophical Logic*, 10:239–266.

Ramsey, F. P. (1929). General propositions and causality. Printed in Braithwaite (1931), pages 237-255.

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

Rips, L. J. (2009). Two causal theories of counterfactual conditionals. *Cognitive Science*, 34 (2):175–221.

Roese, N. J. and Olson, J. M., editors (1995). *What Might Have Been: The Social Psychology of Counterfactual Thinking*. Lawrence Erlbaum, Mahwah, NJ.

Schulz, K. (2007). *Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals*. PhD thesis, University of Amsterdam.

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.

Stalnaker, R. (1970). Probablity and conditionals. *Philosophy of Science*, 37:64–80.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., and Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27:453–489.

Veltman, F. (1976). Prejudices, presuppositions and the theory of counterfactuals. In Groenendijk, J. and Stokhof, M., editors, *Papers in Formal Grammar*, volume 1, pages 248–281. University of Amsterdam, Amsterdam.

Veltman, F. (2005). Making counterfactual assumptions. *Journal of Semantics*, 22:159–180.

Waldmann,M. R. and Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31:216–227. 28