A Theory of First-Order Counterfactual Reasoning

Michael Thielscher

Dresden University of Technology
mit@inf.tu-dresden.de

Abstract. A new theory of evaluating counterfactual statements is presented based on the established predicate calculus formalism of the Fluent Calculus for reasoning about actions. The assertion of a counterfactual antecedent is axiomatized as the performance of a special action. An existing solution to the Ramification Problem in the Fluent Calculus, where indirect effects of actions are accounted for via causal propagation, allows to deduce the immediate consequences of a counterfactual condition. We show that our theory generalizes Pearl etal.'s characterization, based on causal models, of propositional counterfactuals.

1 Introduction

A counterfactual sentence is a conditional statement whose antecedent is known to be false as in, "Had the gun not been loaded, President Kennedy would have survived the assassination attempt." If counterfactuals are read as material implications, then they are trivially true due to the presupposed falsehood of the condition. Nonetheless it can be both interesting and important to use a more sophisticated way of deciding acceptability of a counterfactual query like, e.g., "Had the hit-and-run driver immediately called the ambulance, the injured would not have died." Counterfactuals can also teach us lessons which can be useful in the future when similar situations are encountered, as in "Had you put your weight on the downhill ski, you would not have fallen" [11]. Other applications of processing counterfactuals are fault diagnosis, determination of liability, and policy analysis [1].[1]

The first formal theory of reasoning about counterfactuals was developed in [8], where a counterfactual sentence was accepted iff its consequent holds in all hypothetical 'worlds' which are 'closest' to the factual one but satisfy the counterfactual antecedent. A first method for processing actual counterfactual queries on the basis of a concrete concept of worlds and closeness was proposed in [5]. Generally, the value of a theory of processing counterfactuals depends crucially on how well the expected consequences of a counterfactual condition are determined. It has been observed, among others, by the authors of [1, 4] that knowledge of causality is required to this end. Accordingly, their method is based

[1] It is worth mentioning that the importance of theories of counterfactual reasoning for the field of AI is documented by the fact that Judea Pearl receives this year's *IJCAI award for research excellence* also for his pioneering work on causality.

on so-called *causal models*. A distinguishing feature of this approach is that it deals with probabilities of statements and the way these probabilities change under counterfactual conditions. On the other hand, causal models are essentially propositional. This does not allow for processing counterfactuals which involve disjunctions or quantifications as in, “Had you worn a safety helmet, or had you taken any other route, you would not have been hurt by a roof tile.” Further restrictions of the causal models approach to counterfactual reasoning are entailed by the requirement that the value of each dependent variable is uniquely determined by the exogenous variables (see Section 4 on the implications of this).

The Fluent Calculus is a general strategy for axiomatizing knowledge of actions and effects using full classical predicate logic [14]. Based on a Situation Calculus-style branching time structure, the plain Fluent Calculus allows for processing a particular kind of counterfactual statements, namely, where the counterfactual antecedent asserts a sequence of actions different from the one that has actually taken place as in, “Had the assassin shot at the vice president, the president would have survived the shot:” Consider the generic predicate $Holds(f, s)$ denoting that fluent[2] f holds in situation s, and the generic function $Do(a, s)$ denoting the situation reached by performing action a in situation s. Then it is a simple exercise to axiomatize, by means of the Fluent Calculus, knowledge of the effect of $Shoot(p)$, denoting the action of shooting p, in such a way that the following is entailed:[3]

$$\begin{array}{l} Holds(Alive(President), S_0) \wedge Holds(Alive(Vice), S_0) \\ \wedge\, S_1 = Do(Shoot(President), S_0) \wedge S_1' = Do(Shoot(Vice), S_0) \\ \supset\, \neg Holds(Alive(President), S_1) \,\wedge\, Holds(Alive(President), S_1') \end{array}$$

The obvious reason for this to work without further consideration is that the two statements $\neg Holds(Alive(President), S_1)$ and $Holds(Alive(President), S_1')$ do not mutually contradict due to the differing situation argument.

Counterfactual assertions about situations instead of action sequences cannot be processed in such a straightforward manner. If to an axiom like,

$$\begin{array}{l} Holds(Loaded(Gun), S_0) \wedge \\ \neg Holds(Alive(President), Do(Shoot\text{-}with(Gun, President), S_0)) \end{array} \tag{1}$$

the counterfactual condition $\neg Holds(Loaded(Gun), S_0)$ is added, then a plain inconsistency is produced. Our theory for processing counterfactual statements with the Fluent Calculus solves this problem by associating a new situation term with a counterfactual antecedent that modifies facts about a situation. The step from an actual one to a situation thus modified is modeled by performing an action which has the very modification as effect. Suppose, e.g., the

[2] A *fluent* represents an atomic property of the world which is situation-dependent, that is, whose truth value may be changed by actions.

[3] A word on the notation: Predicate and function symbols, including constants, start with a capital letter whereas variables are in lower case, sometimes with sub- or superscripts. Free variables in formulas are assumed universally quantified.

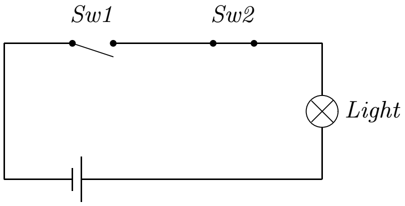

Fig. 1. An electric circuit consisting of a battery, two switches, and a light bulb, which is on if and only if both switches are closed.

'action' *Modify*(*Not-Loaded*(x)) has the sole effect of x becoming unloaded, then a suitable axiomatization in the Fluent Calculus yielded,

$$\begin{array}{l}(1) \wedge S_0' = \mathit{Do}(\mathit{Modify}(\mathit{Not\text{-}Loaded}(\mathit{Gun})), S_0) \\ \supset \mathit{Holds}(\mathit{Alive}(\mathit{President}), \mathit{Do}(\mathit{Shoot\text{-}with}(\mathit{Gun}, \mathit{President}), S_0'))\end{array} \tag{2}$$

which corresponds to the counterfactual statement above, "Had the gun not been loaded, the president would have survived the assassination attempt."

A counterfactual antecedent may have consequences which are more immediate, that is, which do not refer to another action like in (2). Consider, for example, the simple electric circuit depicted in Fig. 1, taken from [9]. A counterfactual statement whose truth is obvious from the wiring (and the assumption that battery, light bulb, and the wires are not broken) is, "Were *Sw1* closed in the situation depicted, light would be on." This conclusion is rooted in this so-called *state constraint*:

$$\mathit{Holds}(\mathit{On}(\mathit{Light}), s) \equiv \mathit{Holds}(\mathit{Closed}(\mathit{Sw1}), s) \wedge \mathit{Holds}(\mathit{Closed}(\mathit{Sw2}), s) \tag{3}$$

However, it is less obvious how to formalize the grounds by which is rejected the counterfactual statement, "Were *Sw1* closed in the situation depicted, *Sw2* would be open," suggested by the implication, $\mathit{Holds}(\mathit{Closed}(\mathit{Sw1}), s) \wedge \neg \mathit{Holds}(\mathit{On}(\mathit{Light}), s) \supset \neg \mathit{Holds}(\mathit{Closed}(\mathit{Sw2}), s)$, which is a logical consequence of axiom (3). This question is closely related to the Ramification Problem [6], that is, the problem of determining the indirect effects of actions. If someone closes *Sw1* in our circuit, does the light come on as an indirect effect or does *Sw2* jump open to restore state consistency?

A standard extension of the Fluent Calculus addresses the Ramification Problem, and in particular meets the challenge illustrated with the circuit, by means of directed causal relations where indirect effects are obtained via causal propagation [12]. This solution, together with our proposal of modeling counterfactual antecedents as actions, furnishes a ready method for accommodating the immediate consequences caused by a counterfactual assertion.

Our proposal for evaluating counterfactuals with the Fluent Calculus and the approach of [1, 4] thus have in common the property of being grounded on causality. On the other hand, two features are not shared even if we confine ourselves to the restricted case of propositional counterfactuals. First, our theory is not restricted by the so-called *reversibility* property [4]. Second, a rather unique feature of our account of counterfactuals is that a counterfactual antecedent may be rejected as being unacceptable in the current state of affairs. Both non-reversibility and the possibility of rejecting counterfactual conditions will be discussed in detail below. We will also show formally how, despite these differences, the approach of [1, 4], if restricted to causal models without probabilities, is embedded in our proposal.

2 The Fluent Calculus

In what follows we give a concise introduction to the axiomatization strategy of the Fluent Calculus. The reader who is unfamiliar with it might want to consult the (electronically available) gentle introduction [13]. A distinguished purpose of the Fluent Calculus, which roots in the logic programming formalism of [7], is to address not only the representational but also the inferential aspect [2] of the classical Frame Problem. We use a many-sorted second order language with equality, which includes sorts for fluents, actions, situations, and states. Fluents are reified propositions. That is to say, we use terms like $\mathit{On}(\mathit{Sw1})$ to denote fluents, where On is a unary function symbol. States are fluents connected via the binary function symbol "$\circ$", written in infix notation, which is assumed to be both associative and commutative and to admit a unit element, denoted by $\emptyset$. Associativity allows us to omit parentheses in nested applications of $\circ$. A function $\mathit{State}(s)$ relates a situation s to the state of the world in that situation, as in the following partial description of the initial state in our circuit example:

$$\exists z\,[\,\mathit{State}(S_0) = \mathit{Closed}(\mathit{Sw2}) \circ z \,\wedge\, \forall z'.\; z \neq \mathit{Closed}(\mathit{Sw1}) \circ z'\,] \tag{4}$$

Put in words, of state $\mathit{State}(S_0)$ it is known that $\mathit{Closed}(\mathit{Sw2})$ is true and possibly some other fluents z hold, too—with the restriction that z does not include $\mathit{Closed}(\mathit{Sw1})$, of which we know that it is false.

Fundamental for any Fluent Calculus axiomatization is the axiom set *EUNA* (the *extended unique names-assumption*) [12]. This set comprises the axioms AC1 (i.e., associativity, commutativity, and unit element) and axioms which entail inequality of two state terms whenever these are not AC1-unifiable. In addition, we have the following foundational axiom, where f is of sort fluent,

$$\forall s, f, z.\; \mathit{State}(s) \neq f \circ f \circ z \tag{5}$$

by which double occurrences of fluents are prohibited in any state which is associated with a situation. Finally, we need the *Holds* predicate introduced in Section 1, though it is not part of the signature but a mere abbreviation of an equality sentence: $\mathit{Holds}(f, s) \overset{\mathrm{def}}{=} \exists z.\, \mathit{State}(s) = f \circ z$.

So-called state update axioms specify the entire relation between the states at two consecutive situations. Regarding our circuit example, let the only direct effect of an action called $\mathit{Toggle}(x)$ be that switch x changes its position from open to closed or vice versa. Ignoring indirect effects for the moment, this is a suitable pair of state update axioms:

$$\begin{array}{r} \neg \mathit{Holds}(\mathit{Closed}(x), s) \supset \mathit{State}(\mathit{Do}(\mathit{Toggle}(x), s)) = \mathit{State}(s) \circ \mathit{Closed}(x) \\ \mathit{Holds}(\mathit{Closed}(x), s) \supset \mathit{State}(\mathit{Do}(\mathit{Toggle}(x), s)) \circ \mathit{Closed}(x) = \mathit{State}(s) \end{array}$$

That is, if $\mathit{Toggle}(x)$ is performed in a situation s where x is not closed, then the new state equals the old state plus $\mathit{Closed}(x)$. Conversely, if x happens to be closed, then the new state plus $\mathit{Closed}(x)$ equals the old state. In other words, in the first case $\mathit{Closed}(x)$ is the only positive direct effect, while it is the only negative direct effect in the second case.

A crucial extension of the basic Fluent Calculus introduced so far copes with the Ramification Problem. Recall, for instance, state constraint (3). It gives rise, among others, to the indirect effect that light turns on if $\mathit{Sw1}$ gets closed whenever $\mathit{Sw2}$ is already closed. Such indirect effects are accounted for by the successive application of directed causal relations [12]. An example for such a relation, which holds for the circuit, is $\mathit{Closed}(\mathit{Sw1})\ \underline{\text{causes}}\ \mathit{On}(\mathit{Light}),\ \underline{\text{if}}\ \mathit{Closed}(\mathit{Sw2})$ while the following should *not* be formalized and added to the axiomatization: $\mathit{Closed}(\mathit{Sw1})\ \underline{\text{causes}}\ \neg\mathit{Closed}(\mathit{Sw2}),\ \underline{\text{if}}\ \neg\mathit{On}(\mathit{Light})$.

It cannot be gathered from a mere state constraint which of its logical consequences correspond to 'real' indirect effects. Yet with the aid of additional domain knowledge about potential causal influence it is possible to automatically extract suitable causal relationships from state constraints [12]:[4]

> Consider a given binary relation $\mathcal{I}$ among the underlying fluents.[5] For all state constraints C, all prime implicates $L_1 \vee \ldots \vee L_m$ of C, all $i = 1, \ldots, m$, and for all $j = 1, \ldots, m,\ j \neq i$: If $(\mathit{atom}(L_i), \mathit{atom}(L_j)) \in \mathcal{I}$,[6] then this is a valid causal relationship:
>
> $$\neg L_i\ \underline{\text{causes}}\ L_j\ ,\ \underline{\text{if}} \bigwedge_{\substack{k=1,\ldots,m \\ k\neq i;\, k\neq j}} \neg L_k$$

For example, $\mathcal{I} = \{(\mathit{Closed}(\mathit{Sw1}), \mathit{On}(\mathit{Light})), (\mathit{Closed}(\mathit{Sw2}), \mathit{On}(\mathit{Light}))\}$ is the appropriate influence relation for our electric circuit. If $\mathcal{I}$ is used for the generation of causal relationships from state constraint (3), then this is the result of the above algorithm:

$$\begin{array}{rl} \mathit{Closed}(\mathit{Sw1})\ \underline{\text{causes}}\ \mathit{On}(\mathit{Light}), & \underline{\text{if}}\ \mathit{Closed}(\mathit{Sw2}) \\ \mathit{Closed}(\mathit{Sw2})\ \underline{\text{causes}}\ \mathit{On}(\mathit{Light}), & \underline{\text{if}}\ \mathit{Closed}(\mathit{Sw1}) \\ \neg\mathit{Closed}(\mathit{Sw1})\ \underline{\text{causes}}\ \neg\mathit{On}(\mathit{Light}) & \\ \neg\mathit{Closed}(\mathit{Sw2})\ \underline{\text{causes}}\ \neg\mathit{On}(\mathit{Light}) & \end{array} \tag{6}$$

[4] The following procedure assumes state constraints to have a format where each occurrence of $\mathit{Holds}(\varphi, s)$ is replaced by the simple atomic expression φ. For the sake of simplicity, we confine ourselves to constraints with the universally quantified s being the only variable. A generalization can be found in [12].

[5] If $(F, G) \in \mathcal{I}$, then this indicates that fluent F may have direct causal influence on fluent G.

[6] By $\mathit{atom}(L)$ we denote the atom of a literal L.

The axiomatization of each single causal relationship in the Fluent Calculus is based on a predicate $\mathit{Causes}(z, e, z', e')$, which shall be true if, according to the causal relationship, in the current state z the occurred effects e give rise to an additional, indirect effect resulting in the updated state z' and the updated current effects e'. Let $\mathcal{R}$ be a set of causal relationships, then by $\mathit{\Pi}[\mathcal{R}]$ we denote the corresponding set of Fluent Calculus axioms defining Causes in this way.

In order to account for possible indirect effects, the state update axioms from above are refined as follows:

$$\begin{array}{l} \neg \mathit{Holds}(\mathit{Closed}(x), s) \supset \\ \quad [\, z = \mathit{State}(s) \circ \mathit{Closed}(x) \supset \mathit{Ramify}(z, \mathit{Closed}(x), \mathit{State}(\mathit{Do}(\mathit{Toggle}(x), s)))\,] \end{array}$$

$$\begin{array}{l} \mathit{Holds}(\mathit{Closed}(x), s) \supset \\ \quad [\, z \circ \mathit{Closed}(x) = \mathit{State}(s) \supset \mathit{Ramify}(z, \,-\mathit{Closed}(x), \mathit{State}(\mathit{Do}(\mathit{Toggle}(x), s)))\,] \end{array}$$

where the term $-F$ represents the occurrence of a negative effect and where $\mathit{Ramify}(z, e, z^*)$ means that state z^* is reachable from z, e by the successive application of (zero or more) causal relationships. Following [12], Ramify is defined via a standard second-order axiom characterizing the reflexive and transitive closure of Causes.

To summarize, let $\mathit{\Sigma}_{\mathit{Circuit}}$ be the union of the two state update axioms just mentioned, state constraint (3), $\mathit{\Pi}[(6)]$, the second-order definition of Ramify, foundational axiom (5), and the appropriate axioms EUNA. This Fluent Calculus theory we will use in the next two sections to illustrate various features of our approach to counterfactual reasoning.

3 Axiomatizing Counterfactuals

We now propose a theory for evaluating counterfactual queries whose antecedent changes what is known about one or more situations. Our theory is implicitly defined by an axiomatization strategy—based on the Fluent Calculus—for counterfactual statements. Consider, for example, the atomic counterfactual condition, for some x and s, "If $\mathit{Closed}(x)$ were true in situation s, ...". By making this assertion one wishes to talk about a situation which is like s except for $\mathit{Closed}(x)$ being true *and* except for all consequences caused by that modification. Generally, our theory allows to process counterfactual statements of the form, "If $\mathit{\Phi}$, then $\mathit{\Psi}$," where the antecedent $\mathit{\Phi}$ expresses modifications—of one or more situations—which can be modeled as actions, and $\mathit{\Psi}$ is a statement about what holds in these (and possibly other) situations. A unified treatment of modifications according to $\mathit{\Phi}$ is provided by the following generic state update axiom, which defines the action $\mathit{Modify}(p, n)$ where p and n are finite collections of fluents which shall become true and false, resp., as requested by the counterfactual antecedent. All further consequences of this update are obtained as indirect effects via ramification. Hence, $\mathit{Modify}(p,n)$ is suitably defined by,[7]

$$\begin{array}{l} \mathit{Poss}(\mathit{Modify}(p,n),s) \supset \\ \quad [\, z \circ n = \mathit{State}(s) \circ p \supset \mathit{Ramify}(z, p \circ -n, \mathit{State}(\mathit{Do}(\mathit{Modify}(p,n),s)))\,] \end{array}$$

where the generic predicate $\mathit{Poss}(a,s)$ means that action a is possible in situation s. The state update axiom is accompanied by this action precondition axiom:[8] $\mathit{Poss}(\mathit{Modify}(p,n),s) \supset \overline{\mathit{Holds}}(p,s) \wedge \mathit{Holds}(n,s)$.

To enhance readability, we introduce the following notation: The expression

$$s \lhd f_1 \wedge \ldots \wedge f_m \wedge \neg f_{m+1} \wedge \ldots \wedge \neg f_n$$

denotes the situation $\mathit{Do}(\mathit{Modify}(f_1 \circ \ldots \circ f_m, f_{m+1} \circ \ldots \circ f_n), s)$. For example, the term $S_0 \lhd \mathit{Closed}(\mathit{Sw1})$ shall denote $\mathit{Do}(\mathit{Modify}(\mathit{Closed}(\mathit{Sw1}), \emptyset), S_0)$.[9]

The axiomatization of a counterfactual conclusion Ψ refers to the modified situation(s) produced by the counterfactual antecedent, as in the following proposition, which asserts the correctness of, "If *Sw1* were closed in the situation depicted in Fig. 1, then light would be on:"

Proposition 1. *The formulas* $\Sigma_{\mathit{Circuit}} \cup \{(4)\}$ *entail,*

$$\mathit{Poss}(S_0 \lhd \mathit{Closed}(\mathit{Sw1})) \supset \mathit{Holds}(\mathit{On}(\mathit{Light}), S_0 \lhd \mathit{Closed}(\mathit{Sw1}))$$

A counterfactual statement may also involve the performance of actions in the hypothetical situation(s) as in, "If *Sw1* were closed in the situation depicted in Fig. 1, then light would be off after toggling *Sw1*:"

Proposition 2. *The formulas* $\Sigma_{\mathit{Circuit}} \cup \{(4)\}$ *entail,*

$$\begin{array}{c} \mathit{Poss}(S_0 \lhd \mathit{Closed}(\mathit{Sw1})) \supset \\ \neg \mathit{Holds}(\mathit{On}(\mathit{Light}), \mathit{Do}(\mathit{Toggle}(\mathit{Sw1}), S_0 \lhd \mathit{Closed}(\mathit{Sw1}))) \end{array}$$

As opposed to existing, propositional accounts of counterfactual reasoning, our theory allows evaluating counterfactual antecedents which exploit the full expressive power of first-order logic and, for instance, include disjunctions of modifications and modifications of more than one situation as in, "If either *Sw2* would have been open in the initial situation as depicted in Fig. 1, or if *Sw1* were open now that we have toggled it, then light would be off now:"

Proposition 3. *The formulas* $\Sigma_{\mathit{Circuit}} \cup \{(4)\}$ *entail,*

$$\begin{array}{l} \mathit{Poss}(S_0 \lhd \neg \mathit{Closed}(\mathit{Sw2})) \wedge \mathit{Poss}(\mathit{Do}(\mathit{Toggle}(\mathit{Sw1}), S_0) \lhd \neg \mathit{Closed}(\mathit{Sw1})) \\ \wedge [\, S_1 = \mathit{Do}(\mathit{Toggle}(\mathit{Sw1}), S_0 \lhd \neg \mathit{Closed}(\mathit{Sw2})) \\ \quad \vee S_1 = \mathit{Do}(\mathit{Toggle}(\mathit{Sw1}), S_0) \lhd \neg \mathit{Closed}(\mathit{Sw1})\,] \supset \neg \mathit{Holds}(\mathit{On}(\mathit{Light}), S_1) \end{array}$$

[7] Below, $-(f_1 \circ \ldots \circ f_m)$ means $-f_1 \circ \ldots \circ -f_m$.

[8] Below, $\overline{\mathit{Holds}}(f_1 \circ \ldots \circ f_m, s)$ means $\neg \mathit{Holds}(f_1, s) \wedge \ldots \wedge \neg \mathit{Holds}(f_m, s)$. The usefulness of preconditions for the *Modify* action will become clear in Section 5, where we consider the rejection of counterfactual antecedents.

[9] With a slight abuse of notation, $\mathit{Poss}(\mathit{Modify}(f_1 \circ \ldots \circ f_m, f_{m+1} \circ \ldots \circ f_n), s)$ shall similarly be written as $\mathit{Poss}(s \lhd f_1 \wedge \ldots \wedge f_m \wedge \neg f_{m+1} \wedge \ldots \wedge \neg f_n)$.

4 Non-Reversibility

The part of our theory where the immediate consequences of a counterfactual antecedent are determined via causal propagation, and the approach of [1, 4] based on causal models, have in common the notion of causality. Nonetheless and even if we cut down the expressiveness of our theory to propositional counterfactuals, there are some important properties which are not shared, one of which is *reversibility* [4]. Informally, reversibility means that if counterfactually asserting that a fluent F has a value x results in a value y for some fluent G, and on the other hand asserting G to have value y results in F achieving value x, then F and G will have the respective values x and y anyway, that is, without any counterfactual assertion.

No comparable property is implied by a set of causal relationships in our approach. This allows to process counterfactuals of the following kind, which cannot be dealt with in the approach of [1, 4]. Suppose the two switches in our main example are tightly mechanically coupled so that it cannot be the case that one is open and the other one is closed [12]. Then both of these counterfactuals are obviously true: "If $\mathit{Sw1}$ were in a different position than it actually is, $\mathit{Sw2}$, too, would assume a different position," and "If $\mathit{Sw2}$ were in a different position than it actually is, $\mathit{Sw1}$, too, would assume a different position." Yet this does not imply, contrary to what reversibility would amount to, that both $\mathit{Sw1}$ and $\mathit{Sw2}$ actually do occupy different positions than they do.

In order to evaluate these counterfactuals, let $\Sigma'_{\mathit{Circuit}}$ be $\Sigma_{\mathit{Circuit}}$ augmented by the state constraint $\mathit{Holds}(\mathit{Closed}(\mathit{Sw1}), s) \equiv \mathit{Holds}(\mathit{Closed}(\mathit{Sw2}), s)$, along with the causal relationships (or rather the Fluent Calculus axiomatization thereof) which are determined by the constraint if the influence relation is extended by $(\mathit{Closed}(\mathit{Sw1}), \mathit{Closed}(\mathit{Sw2}))$ and $(\mathit{Closed}(\mathit{Sw2}), \mathit{Closed}(\mathit{Sw1}))$.

Proposition 4. *The formulas* $\Sigma'_{\mathit{Circuit}}$ *entail,*

$$\begin{array}{l} \mathit{Poss}(S_0 \lhd \mathit{Closed}(\mathit{Sw1})) \land S'_0 = S_0 \lhd \mathit{Closed}(\mathit{Sw1}) \\ \lor\ \mathit{Poss}(S_0 \lhd \neg \mathit{Closed}(\mathit{Sw1})) \land S'_0 = S_0 \lhd \neg \mathit{Closed}(\mathit{Sw1}) \\ \supset\ [\, \mathit{Holds}(\mathit{Closed}(\mathit{Sw2}), S'_0) \equiv \neg \mathit{Holds}(\mathit{Closed}(\mathit{Sw2}), S_0)\,] \end{array}$$

To the same class as this example belong counterfactuals which talk about properties that by their very definition are mutually dependent as in, "If the president were alive, he would not be dead—and vice versa." The reversibility property of the causal models approach to counterfactual reasoning prohibits processing this kind of counterfactual statements.

5 Rejecting Counterfactual Antecedents

Causal propagation of indirect effects is in general not guaranteed to produce a unique result, nor to produce any result at all [12]. In the context of the Ramification Problem, the lack of a resulting state is known as an instance of the Qualification Problem: Rather than giving rise to indirect effects of an action, a state constraint implies an implicit precondition.[10] For our theory of counterfactuals this property of causal relationships implies the rather unique feature that counterfactual antecedents may be rejected if desired.

Consider, for example, this counterfactual sentence due to [11]: "If there had been another car coming over the hill when you passed the car, there would have been a head-on collision." But suppose you as the driver knew that you are on a on-way street, then it seems most appropriate to reject the counterfactual assertion by saying, "But there could not have been another car coming because we are on a one-way." This answer is indeed obtained in our approach by a straightforward formalization of the underlying scenario. Consider, to this end, the state constraint, $\mathit{Holds}(\mathit{Oncoming\text{-}car}, s) \wedge \mathit{Holds}(\mathit{Passing}, s) \supset \mathit{Holds}(\mathit{Collision}, s)$. Both fluents $\mathit{Oncoming\text{-}car}$ and $\mathit{Passing}$ may influence $\mathit{Collision}$. Hence, these two causal relationships are determined by the constraint:

$$\begin{array}{r}\mathit{Oncoming\text{-}car}\ \underline{\text{causes}}\ \mathit{Collision},\ \underline{\text{if}}\ \mathit{Passing}\\ \mathit{Passing}\ \underline{\text{causes}}\ \mathit{Collision},\ \underline{\text{if}}\ \mathit{Oncoming\text{-}car}\end{array}$$

Next we add the knowledge that on a one-way road there are no oncoming cars, formalized by $\mathit{Holds}(\mathit{One\text{-}way}, s) \supset \neg\mathit{Holds}(\mathit{Oncoming\text{-}car}, s)$. Changing the status of a road may causally affect the flow of oncoming traffic but not the other way round, which means that the only causal relationship triggered by the new constraint is, $\mathit{One\text{-}way}\ \underline{\text{causes}}\ \neg\mathit{Oncoming\text{-}car}$. Let $\Sigma_{\mathit{Collide}}$ denote the complete Fluent Calculus axiomatization of this scenario along the line of $\Sigma_{\mathit{Circuit}}$ in Section 2, then we have the following result:

Proposition 5. $\Sigma_{\mathit{Collide}} \cup \{\mathit{Holds}(\mathit{Passing}, S_0) \wedge \neg\mathit{Holds}(\mathit{Oncoming\text{-}car}, S_0) \wedge \neg\mathit{Holds}(\mathit{Collision}, S_0) \wedge \mathit{Holds}(\mathit{One\text{-}way}, S_0)\}$ *entails,*

$$\neg\mathit{Poss}(S_0 \lhd \mathit{Oncoming\text{-}car})$$

The reason is that no available causal relationship allows to restore consistency wrt. the state constraint, $\mathit{Holds}(\mathit{One\text{-}way}, s) \supset \neg\mathit{Holds}(\mathit{Oncoming\text{-}car}, s)$. Proposition 5 is to be interpreted as a rejection of the counterfactual antecedent, "If there had been another car coming over the hill, ..." as 'unrealistic' in the state of affairs. In this way, counterfactual antecedents are only accepted if a world can be constructed around them which is consistent with the state constraints and which does not require 'acausal' modifications.

In the approach of [1, 4], 'acausal' modifications are also not permitted, but the realization of counterfactual conditions involves annulling some of what corresponds to our state constraints, namely, those which normally determine the values of the fluents being altered by the counterfactual antecedent. Any antecedent is thus accepted.

[10] A standard example is the constraint which says that in certain cultures you cannot be married to two persons. This axiom gives rise to the (implicit) precondition that you cannot marry if you are already married. The constraint should not imply the indirect effect of automatically becoming divorced [10].

The possibility of a counterfactual being rejected, desired as it could be in general, may not always be accepted a reaction. Consider the counterfactual, “If light were on in the situation depicted in Fig. 1, the room would not be pitch dark.” As it stands, our axiomatization would reject the condition of this counterfactual on the grounds that the light could not possibly be on because the controlling switches are not in the right position. Insisting upon the counterfactual condition in question, and coming to the conclusion that the counterfactual statement holds, would require to deny the background knowledge of the relation between the switches and the light bulb. Making explicit the desire to deny this relation, the counterfactual statement can be evaluated without rejection if state constraint (3) is replaced by,

$$\begin{array}{c} \neg \mathit{Holds}(\mathit{Denied}(\mathit{Switch\text{-}Light\text{-}Relation}), s) \supset \\ [\, \mathit{Holds}(\mathit{On}(\mathit{Light}), s) \equiv \mathit{Holds}(\mathit{Closed}(\mathit{Sw1}), s) \wedge \mathit{Holds}(\mathit{Closed}(\mathit{Sw2}), s) \,] \end{array}$$

The generic fluent $\mathit{Denied}(x)$ shall be used in general whenever there is desire to weaken a state constraint in this fashion. Situations which do not result from counterfactual reasoning are supposed to not deny any underlying relation among fluents. This is expressed by these three axioms:

$$\begin{array}{l} \mathit{Factual}(S_0) \\ \mathit{Factual}(s) \wedge \forall p, n.\, a \neq \mathit{Modify}(p, n) \supset \mathit{Factual}(\mathit{Do}(a, s)) \\ \mathit{Factual}(s) \supset \neg \mathit{Holds}(\mathit{Denied}(x), s) \end{array}$$

Let $\Sigma^*_{\mathit{Circuit}}$ be $\Sigma_{\mathit{Circuit}}$ thus modified. Then the above counterfactual antecedent is acceptable if the denial of the dependence of the light is made explicit:

Proposition 6. $\Sigma^*_{\mathit{Circuit}} \cup \{(4)\}$ *is consistent with,*

$$\mathit{Poss}(S_0 \lhd \mathit{On}(\mathit{Light}) \wedge \mathit{Denied}(\mathit{Switch\text{-}Light\text{-}Relation}))$$

6 Axiomatizing Causal Model-Counterfactuals

In concentrating on the crucial connection between reasoning about counterfactuals and causal reasoning, the approach of [1, 4] based on causal models has a strong relation to the proposal of the present paper. Despite the conceptual difference between probabilistic, propositional causal models and the second-order Fluent Calculus with the full expressive power of logic, and despite the further differences discussed in the preceding two sections, counterfactual reasoning in causal models, in the non-probabilistic case, can be embedded into our theory. For the sake of simplicity and clarity, we assume all variables in causal models to be binary. The following definitions follow [4].

Definition 7. A causal model is a triple $M = \langle \mathcal{U}, \mathcal{V}, \mathcal{F} \rangle$ where $\mathcal{U}$ and $\mathcal{V} = \{V_1, \ldots, V_n\}$ are disjoint sets of propositional variables (exogenous and endogenous, resp.), and $\mathcal{F}$ is a set of propositional formulas $\{F_1, \ldots, F_n\}$ such that

(i) F_i contains atoms from $\mathcal{U}$ and $\mathcal{V} \setminus \{V_i\}$. (The set of variables from $\mathcal{V}$ that occur in F_i is denoted by PA_i (the *parents* of V_i).)

(ii) For each interpretation for the variables in $\mathcal{U}$ there is a unique model of $(V_1 \equiv F_1) \wedge \ldots \wedge (V_n \equiv F_n)$.

As an example, consider the causal model M_{Circuit} consisting of $\mathcal{U} = \{U_1, U_2\}$; $\mathcal{V} = \{\mathit{Sw1}, \mathit{Sw2}, \mathit{Light}\}$; and $F_{\mathit{Sw1}} = U_1$, $F_{\mathit{Sw2}} = U_2$, and $F_{\mathit{Light}} = \mathit{Sw1} \wedge \mathit{Sw2}$, which models the electric circuit of Fig. 1 using two additional, exogenous variables U_1 and U_2 that determine the positions of the two switches.

Definition 8. Let $M = \langle \mathcal{U}, \mathcal{V}, \mathcal{F} \rangle$ be a causal model, $\mathcal{X} \subseteq \mathcal{V}$, and $\iota_{\mathcal{X}}$ a particular interpretation for the variables in $\mathcal{X}$. A *submodel* of M is the causal model $M_{\iota_{\mathcal{X}}} = \langle \mathcal{U}, \mathcal{V}, \mathcal{F}_{\iota_{\mathcal{X}}} \rangle$ with $\mathcal{F}_{\iota_{\mathcal{X}}} = \{F_i \in \mathcal{F} : V_i \notin \mathcal{X}\} \cup \{X \equiv \iota_{\mathcal{X}}(X) : X \in \mathcal{X}\}$, provided $M_{\iota_{\mathcal{X}}}$ is a causal model according to Def. 7.

For $Y \in \mathcal{V}$, let $Y_{\iota_{\mathcal{X}}}(\iota_{\mathcal{U}})$ denote the truth-value for Y in the (unique) model of $F_{\iota_{\mathcal{X}}}$ with interpretation $\iota_{\mathcal{U}}$ for $\mathcal{U}$. Then $Y_{\iota_{\mathcal{X}}}(\iota_{\mathcal{U}})$ is the evaluation of the counterfactual sentence, "If $\mathcal{X}$ had been $\iota_{\mathcal{X}}$, then Y," in the setting $\iota_{\mathcal{U}}$.

E.g., a submodel of M_{Circuit} is given by, $\mathcal{F}_{\{\mathit{Sw1}=\mathit{True}\}} = \{F_{\mathit{Sw1}} = \mathit{True}, F_{\mathit{Sw2}} = U_2, F_{\mathit{Light}} = \mathit{Sw1} \wedge \mathit{Sw2}\}$. Consider $\iota_{\mathcal{U}} = \{U_1 = \mathit{False}, U_2 = \mathit{True}\}$, which characterizes the situation depicted in Fig. 1. Then $\mathit{Light}_{\{\mathit{Sw1}=\mathit{True}\}}(\iota_{\mathcal{U}}) = \mathit{True}$, which confirms the counterfactual, "If *Sw1* were closed, light would be on."

We will now present a correct Fluent Calculus axiomatization of causal models and the evaluation of counterfactuals. The (propositional) fluents are the variables of the model. The definitions $\{F_1, \ldots, F_n\}$ of the endogenous variables are directly translated into state constraints, each of which can possibly be denied, that is, $\neg \mathit{Holds}(\mathit{Denied}(\mathit{definition\text{-}of\text{-}V_i}), s) \supset \mathit{HOLDS}(F_i, s)$.[11] The state constraints determine a collection of causal relationships on the basis of the influence relation $\mathcal{I} = \{(V, V_i) : V \in \mathit{PA}_i\}$. For a causal model M, let Σ_M denote the Fluent Calculus axiomatization which consists of the foundational axioms, including a suitable set *EUNA*, along with the state constraints and the axiomatizations of the causal relationships determined by M as just described.

Theorem 9. *Let $M = \langle \mathcal{U}, \mathcal{V}, \mathcal{F} \rangle$ be a causal model wit Fluent Calculus axiomatization Σ_M. Consider a subset $\mathcal{X} \subseteq \mathcal{V}$ along with a particular realization $\iota_{\mathcal{X}}$ such that $M_{\iota_{\mathcal{X}}}$ is a submodel, a variable $Y \in \mathcal{V}$, and a particular realization $\iota_{\mathcal{U}}$ for $\mathcal{U}$. Let $\Sigma = \Sigma_M \cup \{\bigwedge_{U \in \mathcal{U}} [\mathit{Holds}(U, S_0) \equiv \iota_{\mathcal{U}}(U)]\}$ and let $S_0' = S_0 \lhd \bigwedge_{X \in \mathcal{X}} [X \equiv \iota_{\mathcal{X}}(X)] \wedge \bigwedge_{X \in \mathcal{X}} \mathit{Denied}(\mathit{definition\text{-}of\text{-}X})$. Then,*

1. *$\Sigma \cup \{\mathit{Poss}(S_0')\}$ is consistent.*
2. *$\Sigma \models \mathit{Poss}(S_0') \supset \mathit{Holds}(Y, S_0')$ iff $Y_{\iota_{\mathcal{X}}}(\iota_{\mathcal{U}}) = \mathit{True}$.*
3. *$\Sigma \models \mathit{Poss}(S_0') \supset \neg \mathit{Holds}(Y, S_0')$ iff $Y_{\iota_{\mathcal{X}}}(\iota_{\mathcal{U}}) = \mathit{False}$.*

Proof (sketch). The fact that $M_{\iota_{\mathcal{X}}}$ admits a unique model $\iota_{\mathcal{V}}$ under $\iota_{\mathcal{U}}$ implies that there is a unique state which complies with $\iota_{\mathcal{U}}$ and $\iota_{\mathcal{X}}$ and which satisfies the state constraints. From the construction of the underlying causal relationships it follows that this state is reachable by ramification, which proves claim 1. The construction of the causal relationships also implies that no relationship can

[11] $\mathit{HOLDS}(F, s)$ is F but with each atom A replaced by $\mathit{Holds}(A, s)$.

be applied by which is modified any fluent representing a variable from $\mathcal{U}$ or from $\mathcal{X}$. Hence, the aforementioned state is the only one which can be consistently assigned to $State(S'_0)$. This proves claims 2 and 3, since this state agrees with $\iota_{\mathcal{V}}$ on all variables.

7 Discussion

The author of [5] argues against pushing too far the connection between counterfactual and causal reasoning, on two grounds. First, a counterfactual statement may stress that antecedent and conclusion are *not* causally linked, as in, "Even if I were free tonight, I still would not have dinner with you." This is perfectly compatible with our theory, by which the example counterfactual would be confirmed *because* of the lack of a causal connection. Second, a counterfactual statement may reverse the direction of causality to serve as explanation as in, "If John had Koplic spots, he would have measles." In order to accommodate such explanatory counterfactuals, which amounts to saying which of the *causes* of a denied supposition require modification, an extension of our theory is needed which allows to carefully add appropriate explanatory 'causal' relationships which only apply when performing a *Modify* action.

References

1. A. Balke and J. Pearl. Counterfactuals and policy analysis. In P. Besnard and S. Hanks, ed.'s, *Proc. of UAI*, pp. 11–18. Morgan Kaufmann, 1995.
2. W. Bibel. A deductive solution for plan generation. *New Gener. Comput.*, 4:115–132, 1986.
3. W. Bibel. Let's plan it deductively! *Artif. Intell.*, 103(1–2):183–208, 1998.
4. D. Galles and J. Pearl. An axiomatic characterization od causal counterfactuals. In *Foundations of Science*. Kluwer Academic, 1998.
5. M. L. Ginsberg. Counterfactuals. *Artif. Intell.*, 30:35–79, 1986.
6. M. L. Ginsberg and D. E. Smith. Reasoning about action I: A possible worlds approach. *Artif. Intell.*, 35:165–195, 1988.
7. S. Hölldobler and J. Schneeberger. A new deductive approach to planning. *New Gener. Comput.*, 8:225–244, 1990.
8. D. Lewis. *Counterfactuals*. Harvard University Press, 1973.
9. V. Lifschitz. Frames in the space of situations. *Artif. Intell.*, 46:365–376, 1990.
10. F. Lin and R. Reiter. State constraints revisited. *J. of Logic and Comput.*, 4(5):655–678, 1994.
11. J. McCarthy and T. Costello. Useful counterfactuals and approximate theories. In C. Ortiz, ed., *Prospectes for a Commonsense Theory of Causation*, AAAI Spring Symposia, pp. 44–51, Stanford, 1998. AAAI Press.
12. M. Thielscher. Ramification and causality. *Artif. Intell.*, 89(1–2):317–364, 1997.
13. M. Thielscher. Introduction to the Fluent Calculus. *Electr. Transact. on Artif. Intell.*, 1998. (Submitted). URL: `http://www.ep.liu.se./ea/cis/1998/014/`.
14. M. Thielscher. From Situation Calculus to Fluent Calculus: State update axioms as a solution to the inferential frame problem. *Artif. Intell.*, 1999. (To appear).

This article was processed using the LaTeX macro package with LLNCS style