Causality, Constraints and the Indirect Effects of Actions

Hector Geffner*
Dpto. de Computación
Universidad Simón Bolívar
Aptdo. 89000, Caracas 1080-A
Venezuela

Abstract

The goal of this paper is the formulation of a well-founded solution to the ramification problem: the problem of determining the indirect effects of an action. The solution relies on a model of causal reasoning based on two simple ideas: that the values that a variable may take are a *function* of the values of its causes, and that the *form* of that function is determined by the rules acting upon the variable. The model is related to models of causal reasoning based on Bayesian Networks and Structural Equation Models but does not appeal to probabilities or rankings of any sort. It also combines (propositional) action rules and causal rules in a clear manner, and provides a simple way for determining when 'derived information' can be assumed to persist.

1 Introduction

The goal of this paper is the formulation of a well-founded solution to the ramification problem: the problem of determining the indirect effects of an action. The solution follows recent approaches that give causality a prominent role [McCain & Turner, 1997; Lin, 1995] but relies on a novel model of causal reasoning that is based on two simple ideas; namely that the values that a variable may take are a *function* of the values of its causes, and that the *form* of that function is determined by the causal rules acting upon the variable. The model is closely related to models of causal reasoning based on Bayesian Networks (e.g., [Goldszmidt & Pearl, 1992; Darwiche & Pearl, 1994; Geffner, 1996]) and Structural Equation Models [Simon, 1953; Druzdzel & Simon, 1993; Pearl, 1996]) but does not appeal to probabilities or rankings of any sort. It also combines (propositional) action rules and causal rules in a clear manner, provides a simple way for determining when 'derived information' can be assumed to persist [Myers & Smith, 1988; Giunchiglia & Lifschitz, 1995], and provides another perspective on the distinction between 'qualification' and 'ramification' constraints [Lin & Reiter, 1994].

The paper starts with the description of the causal model in Section 2, and proceeds in Section 3 with the model of actions and time. The two models are integrated in Section 4 and two extensions are considered in Section 5. A brief summary closes the paper in Section 6.

2 Causality

The causal model provides a semantics for collections of *causal rules* defined over a simple language made up of two sets of symbols: *variables* and *values.* All atoms are of the from $\mathsf{x} = x$, where x is a variable and x is a value. The language is the closure of all such atoms under the standard propositional connectives.

Each variable x has also a *domain* D_{x} of values. For example, the variable $\mathsf{temperature}$ may take values from the domain $D_{temp} = \{high, medium, low\}$, $\mathsf{distance}$ may take values from $D_{dist} = \{1, 2, 3, \ldots, 10\}$, rain may take values from the boolean domain $D_{rain} = \{\mathbf{t}, \mathbf{f}\}$, etc.

A *state* s is an interpretation that assigns to each variable x a value x from its domain. We call x the denotation of x in s and refer to it as x^s. Atoms $\mathsf{x} = x$ are true in a state s when $\mathsf{x}^s = x$, conjunctions $A \wedge B$ are true in s when both A and B are true in s, disjunctions $A \vee B$ are true in s when A or B are true, etc. A state thus provides an interpretation for *all* formulas in the language, and is a *model* of a set of formulas when it satisfies (i.e., makes true) each formula in the set.

Causal rules are of the form $F \rightarrow A$ where F is a formula and A is an atom. Causal rules express *mechanisms* by which the truth of F *normally* causes the truth of A [Simon, 1953; Druzdzel & Simon, 1993; Pearl, 1996]. Causal rules in which the truth of F *always* causes the truth of A are written as $F \supset A$. We call the former, *default* causal rules and the latter *strict* causal rules.

For example, using the expressions x and $\neg x$ to abbreviate atoms of the form $\mathsf{x} = \mathbf{t}$ and $\mathsf{x} = \mathbf{f}$ for *boolean* variables x, a situation in which a car may fail to start due a dead battery can be modeled as:

$$turn\_key \;\; \rightarrow \;\; car\_starts$$

*Mailing address from US and Europe: Hector Geffner, Bamco CCS 144-00, P.O.BOX 02-5322, Miami Florida 33102-5322, USA. E-mail: hector@usb.ve.

$$\begin{aligned} dead\_battery &\supset \neg car\_starts \\ lights &\rightarrow dead\_battery \end{aligned}$$

where the last rule says that a battery dies when the lights are left on for a long time.

The *semantic problem* in causal reasoning is the problem of characterizing in a meaningful way the consequences that follow from a set C of causal rules possibly extended with a set O of observations (atoms). While the *models* of the resulting *causal theories* $T = \langle C, O \rangle$ are the states s that satisfy the strict rules in C and the formulas in O, not all such models can be considered. Defeasible causal rules in C must play a role as well yet it hasn't been clear what such role actually is.

For instance, the simplest idea of focusing on the models that violate *minimal* sets of defaults does not work.[1] Such account correctly predicts *dead_battery* from the observation *lights* but incorrectly retracts the prediction in the presence of the additional but irrelevant piece of information *turn_key*. This problem is avoided by approaches in which causal rules are regarded as one-directional rules of inference.[2] However, those approaches do not explain *why* reasoning 'forward' is often the right thing to do, nor when reasoning 'forward' may fail (as when the three atoms above are all true). An interpretation of causal theories that sheds light on this question is [Goldszmidt & Pearl, 1992], where causal theories are seen as describing a family of Bayesian Networks [Pearl, 1988] that make explicit the independences in the theory. Our model is related to [Goldszmidt & Pearl, 1992] but is simpler and more transparent as it also appeals to notions of independence but does not rely on either Bayesian Networks or rankings of any sort.

2.1 The Model

The model is based on two ideas that are very simple and appear in one way or the other in several accounts of causal reasoning (e.g. [Goldszmidt & Pearl, 1992; Darwiche & Pearl, 1994; Geffner, 1996]):

1. the values that a variable x may take in a given state s is a *function* of the value of the *causes* of x in s
2. the *form* of that function is determined by a standard interpretation of the rules that express causal influences over x

In order to formalize these two ideas, we express the set of *causal influences* acting on a variable x in a theory $T = \langle C, O \rangle$ by T_x; i.e. T_x contains all the rules in C of the form $F \supset \mathsf{x} = x$ or $F \rightarrow \mathsf{x} = x$. We call T_x the (causal) *microtheory* for x.[3] The microtheories T_x will be used to determine the possible values of x in a given state s. We assume that no variable is its own cause; i.e., no rule involves the same variable in both its body and its head.

We write $s/[\mathsf{x} = x]$ for $x \in D_\mathsf{x}$ to denote the state that makes $\mathsf{x} = x$ true and is otherwise identical to s. We also say that a model $s/[\mathsf{x} = x]$ of the microtheory T_x is a *minimal model* of T_x when no other model $s/[\mathsf{x} = x']$ of T_x violates a strictly smaller set of rules from T_x. The ideas above can then be expressed as follows:

1. for each variable x there is a function $\phi_\mathsf{x}(s) \subseteq D_\mathsf{x}$, such that $x \in \phi_\mathsf{x}(s)$ means that x is a possible value of x given the *causes* of x that are true in s
2. the function $\phi_\mathsf{x}(s)$ is such that $x \in \phi_\mathsf{x}(s)$ iff $s/[\mathsf{x} = x]$ is a *minimal model* of the microtheory T_x

If *min* is a function that takes a collection of states and returns the ones that violate a minimal set of defaults from T_x but no strict rules, then the definition of the functions ϕ_x above can be expressed equivalently as:

$$\phi_\mathsf{x}(s) \stackrel{\text{def}}{=} min^{-1}_{x \in D_\mathsf{x}}\ s/[\mathsf{x} = x] \qquad (1)$$

Default rules of different priorities can be accommodated in this definition by doing the mimization lexicographically (stronger rules first).

The functions ϕ_x, that we call *causal functions,* determine the values that a variable x can take given the value of its causes. We say that a state s is a *causal interpretation* relative to a set of rules when the *actual* value of every variable x in s is compatible with the value of its causes; i.e., when $\mathsf{x}^s \in \phi_\mathsf{x}(s)$. We also say that s is a *causal model* of a causal theory $T = \langle C, O \rangle$ when s is a causal interpretation relative to C that satisfies the observations O. We assume throughout that causal theories are consistent in the sense that $\phi_\mathsf{x}(s)$ is never empty.

The *consequences* of a causal theory T are the formulas that hold in all its (causal) models.

Example 1 The domain theory considered above involves four boolean variables: turn_key, lights, dead_battery and cart_starts. The microtheories for the first two variables are empty and hence from (1):

$$\phi_{\mathsf{turn\_key}}(s) = \phi_{\mathsf{lights}}(s) = \{\mathbf{t}, \mathbf{f}\}$$

The microtheories for car_starts and dead_battery, on the other hand, are given by the first two rules and third rule respectively, and hence:

$$\phi_{\mathsf{dead\_battery}}(s) = \begin{cases} \{\mathbf{t}\} & \text{if } s \models lights \\ \{\mathbf{t}, \mathbf{f}\} & \text{otherwise} \end{cases}$$

$$\phi_{\mathsf{car\_starts}}(s) = \begin{cases} \{\mathbf{t}\} & \text{if } s \models turn\_key \land \neg dead\_battery \\ \{\mathbf{f}\} & \text{if } s \models dead\_battery \\ \{\mathbf{t}, \mathbf{f}\} & \text{otherwise} \end{cases}$$

Clearly the state s that makes:

$$\mathsf{turn\_key}^s = \mathbf{t},\ \mathsf{lights}^s = \mathbf{t},\ \mathsf{dead\_battery}^s = \mathbf{t},\ \mathsf{car\_start}^s = \mathbf{f}$$

[1] Strict and defeasible rules $F \rightarrow A$ are violated in a state s when s satisfies F but does not satisfy A.

[2] This directionality can be defined procedurally as in inheritance hierarchies and logic programs or by means of a formalism like Reiter's default logic.

[3] These microtheories are similar to the ones considered in [Darwiche & Pearl, 1994] even if their use here is different. Indeed, many of our microtheories are 'inconsistent' according to the criterion in [Darwiche & Pearl, 1994].

is a *causal interpretation*, while the state s' that corresponds to the 'unintended' model:

$\mathsf{turn\_key}^{s'} = \mathbf{t}$, $\mathsf{lights}^{s'} = \mathbf{t}$, $\mathsf{dead\_battery}^{s'} = \mathbf{f}$, $\mathsf{car\_start}^{s'} = \mathbf{t}$

is not as $\phi_{\mathsf{dead\_battery}}(s') = \{\mathbf{t}\}$ and thus $\mathsf{dead\_battery}^{s'} \notin \phi_{\mathsf{dead\_battery}}(s')$. Indeed, the state s is the *only* causal interpretation that satisfies the observations $O = \{turn\_key, lights\}$, and thus s is the only causal model of the theory $T = \langle C, O\rangle$.

2.2 Related Models

Theories of Action. The model above is a refinement of the minimal model semantics in which a single global minimization over the whole set of rules is replaced by multiple minimizations over the local microtheories T_{x}. Moreover, the local minimizations are done over a a *fixed* state s in which only the value of the variable x is allowed to change. Something similar occurs in the recent approaches to the frame problem. It is well known that if default rules are added for fluents and persistence violations are minimized, the resulting semantics is too weak [Hanks & McDermott, 1987]. The most recent approaches to the frame problem thus replace the global minimization by local minimizations in which the current state and actions are fixed, and the resulting state is allowed to change [Gelfond & Lifschitz, 1993; Reiter, 1991]. The result is that the action and persistence rules get compiled into state transition functions that determine the states that can follow a given state s and action a, very much like causal rules above get compiled into the causal functions $\phi_{\mathsf{x}}(s)$ that determine the states of the variable x that can 'follow' a given state of its causes.

Bayesian Networks. The definition of the causal interpretations s above involves a fixed point condition: s has to be such that the value x^s of each variable x belongs to $\phi_{\mathsf{x}}(s)$. This fixed point condition can be simplified when the causal rules in the theory do not involve cycles. Indeed, let us define the *causes* of x as any *minimal* set of variables $causes(\mathsf{x})$ that shield x from all other variables; i.e., $[causes(\mathsf{x})]^s = [causes(\mathsf{x})]^{s'}$ must imply $\phi_{\mathsf{x}}(s) = \phi_{\mathsf{x}}(s')$. Then it is easy to show that the function $\phi_{\mathsf{x}}(s)$ can be replaced by an equivalent function $\psi_{\mathsf{x}}(causes^s(\mathsf{x}))$ where the global state s is replaced by the local state of x's causes. Then if there are no causal cycles and thus the variables can be ordered into a sequence $\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n$ such no variable is preceded by one of its causes, the causal interpretations s can be computed *iteratively* by setting the value of each variable x_i, from $i = 1$ to n, to any value $x_i \in \psi_{\mathsf{x}_i}(causes^s(\mathsf{x}_i))$. Indeed, the model that results in this case is exactly a Bayesian Network in which probabilities are replaced by κ measures [Goldszmidt & Pearl, 1992] that are either 0 (plausible) or ∞ (impossible).

Structural Equation Models. In the presence of cycles, the resulting models cannot be interpreted as Bayesian Networks but as Structural Equation Models [Simon, 1953; Druzdzel & Simon, 1993]. Structural Equation Models are used widely in Economics and the Social Sciences and have the form of a set of simultaneous equations of the form:

$$x_i = f_i(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, \epsilon_i) \qquad (2)$$

where x_i stands for the value of variable x_i, ϵ_i is an error variable, and f_i is a deterministic function. Such equations are called *structural* as their *form* is important: equations that are *mathematically* equivalent are no necessarily *structurally* equivalent.[4] Structural equations represent mechanisms that propagate external perturbations on the right hand side variables to changes in the left hand side variable but not the other way around. For example, an structural equation like $spending = income/2 + \epsilon$ predicts that doubling the income may double the level of spending but does *not* predict that doubling the level of spending will double the income (see also [Pearl, 1996]).

The equation that defines our causal interpretations can be expressed in the form:

$$x_i^s \in \phi_{\mathsf{x}_i}(x_1^s, x_2^s, \ldots, x_{i-1}^s, x_{i+1}^s, \ldots, x_n^s) \qquad (3)$$

that is very close to (2) except that the function ϕ_{x_i}, unlike the function f_i above, does not have to pick a *single* value for the variables x_i.

Coherent Models. Some approaches to causal reasoning replace the notion of *minimal models* by the notion of *coherent models* [Geffner, 1992; McCain & Turner, 1997]). Roughly, a model is *coherent* when all default violations are *causally explained.* In our language, we can consider the violation of a default rule $F \to A$ *explained* in a model s, when there is *conflicting* defeasible or strict rule $F' \to A'$ that is 'active' in s, i.e., whose body and head both hold in s. It is very easy to show then that causal and coherent models coincide. In this sense, coherence considerations which otherwise appear somewhat ad-hoc, can be justified by the ideas that define the causal models.

3 Time and Actions

Many causal relations such as the action of toggling a switch cannot be modeled without a notion of time or context. In this section we focus on such causal relations and consider the interpretation of *temporal* causal rules of the form $F \longrightarrow A$ where F and A are formulas and atoms as before. Strict temporal rules are written as $F \Longrightarrow A$. The difference between a (non-temporal) causal rule $F \to A$ and a temporal (causal) rule $F \longrightarrow A$ is that the former says that A is normally true in the state s when F is true in s, while the latter says that A is normally true in the state *following* s when F is true in s. The action of toggling a switch can be modeled by

[4] Note that the same holds for 'qualification' and 'ramification' constraints [Lin & Reiter, 1994]: they may be logically equivalent yet, for the same structural reasons, they behave in different ways. See Section 5.2.

the *temporal* rules of the form (the notation is similar to [Gelfond & Lifschitz, 1993]):

$$\begin{array}{lcl} \mathit{toggle} \wedge \mathsf{switch} = \mathit{on} & \longrightarrow & \mathsf{switch} = \mathit{off} \\ \mathit{toggle} \wedge \mathsf{switch} = \mathit{off} & \longrightarrow & \mathsf{switch} = \mathit{on} \end{array}$$

The semantics of temporal rules involves sequences of states $s_0, s_1, s_2, \ldots$, called *trajectories,* where s_i refers to the state of the world at time i. As before, the *temporal interpretations* will be the trajectories that interpret all variables x in a way that is compatible with the causal function ϕ^t_{x} determined by the causal microtheory T^t_{x} (we use the index t to express that we are dealing now with *temporal* rules only).

The microtheories T^t_{x} are defined exactly as before as the collection of rules that express causal influences on x; namely the rules of the form $F \longrightarrow \mathsf{x} = x$ and $F \Longrightarrow \mathsf{x} = x$.

These microtheories are interpreted by means of *extended states* of the form $s//[\mathsf{x}] = \mathsf{x}$, where s represents a state at some time point and $\mathsf{x} = x$ represents the value of x at the *next* time point. More precisely, an *extended state* $s//[\mathsf{x} = x]$ satisfies a *temporal rule* $F \longrightarrow \mathsf{x} = x'$ if F^s is false or $x = x'$ is true.

The function $\phi^t_{\mathsf{x}}(s)$ is defined as the set of values $x \in D_{\mathsf{x}}$ for which the extended state $s//[\mathsf{x} = x]$ satisfies all strict rules in T^t_{x} and violates a minimal set of default rules from T^t_{x}. In analogy to (1):

$$\phi^t_{\mathsf{x}}(s) \stackrel{\text{def}}{=} min^{-1}_{x \in D_{\mathsf{x}}}\ s//[\mathsf{x} = x] \qquad (4)$$

Now $x \in \phi^t_{\mathsf{x}}(s)$ no longer means that x is a possible value of x in the state s, but a possible value of x in the states *following* s. This suggests that the *temporal interpretations* could be defined as the trajectories s_0, s_1, s_2, ... in which the condition $\mathsf{x}^{s_{i+1}} \in \phi_{\mathsf{x}}(s_i)$ holds for all variables x and all time points $i \geq 0$.

This actually captures the semantics of the rules but misses something important: the *law of inertia* that expresses the tendency of certain variables, usually called *fluents*, to retain their values.

The persistence of fluents can be captured by *extending* the causal functions $\phi^t_{\mathsf{x}}(s)$ into stronger functions $\phi^*_{\mathsf{x}}(s)$ in which the value x^s of fluents is allowed to persist when the rules in T_{x} provide no reason for a change.

More precisely we assume that the value of x persists when the rules in T_{x} *do no discriminate among the values of* x; i.e. when $\phi^t_{\mathsf{x}}(s) = D_{\mathsf{x}}$. The extended or *persistent causal function* ϕ^*_{x} for *fluent* variables x is thus defined as follows:

$$\phi^*_{\mathsf{x}}(s) \stackrel{\text{def}}{=} \begin{cases} \phi_{\mathsf{x}}(s) & \text{if } \phi^t_{\mathsf{x}}(s) \subset D_{\mathsf{x}} \\ \{\mathsf{x}^s\} & \text{otherwise} \end{cases} \qquad (5)$$

For non-fluent variables, $\phi^*_{\mathsf{x}}(s) \stackrel{\text{def}}{=} \phi_{\mathsf{x}}(s)$.

With the functions $\phi^*_{\mathsf{x}}(s)$, the *temporal interpretations* can be defined in complete analogy to the *causal interpretations* as:

Definition 1 *A state trajectory* s_0, s_1, ... *is a* temporal interpretation *relative to a given set of temporal rules, if* $\mathsf{x}^{s_{i+1}} \in \phi^*_{\mathsf{x}}(s_i)$ *for each variable* x *and each time point* $i \geq 0$.

3.1 Temporal Theories

The *models* of a *temporal theory* $T = \langle D, A, O \rangle$ comprised of a set D of *temporal rules,* a set A of *actions,* and a set O of *observations,* are the temporal interpretations (relative to D) that are *compatible with both the actions A and the observations O.*

Actions and observations in temporal theories are encoded by means of *temporal atoms* of the form $\mathsf{x}[i] = x$, where x is an *action* or *fluent variable* respectively, x belongs to D_{x}, and i is a time index (i.e., a non-negative integer). For boolean variables $\mathsf{x}[i] = \mathbf{t}$ and $\mathsf{x}[i] = \mathbf{f}$ are abbreviated as $x[i]$ or $\neg x[i]$. For simplicity, action variables are assumed boolean.

A trajectory s_0, s_1, ... is *compatible with the observations O* when $\mathsf{x}^{s_i} = x$ if $\mathsf{x}[i] = x$ is in O, and is *compatible with the actions A* when for all action variables x, $\mathsf{x}^{s_i} = \mathbf{t}$ iff $x[i] \in A$.[5]

The *consequences* of a temporal theory are the temporal atoms $\mathsf{x}[i] = x$ that are true in all models, where $\mathsf{x}[i] = x$ is true in a model s_0, s_1, ..., when $\mathsf{x}^{s_i} = x$.

The model that results from these definitions is similar to several other models of action. Here we draw a simple comparison with the model $\mathcal{A}$ of Gelfond and Lifschitz [Gelfond & Lifschitz, 1993]. For simplicity we consider domain descriptions $D_{\mathcal{A}}$ comprised only of actions rules 'A **causes** B **if** C' and initial conditions '**initially** L'.[6] Such description can be expressed as a set D of temporal rules of the form $A \wedge C \longrightarrow B$ and a set O of observations $L[0]$. Then, it is easy to show that:

Theorem 1 *A* value proposition *L* **after** *A_0, ... A_n is entailed by $D_{\mathcal{A}}$ iff $L[n+1]$ follows from the temporal theory $T = \langle D, A, O \rangle$, where $A = \{A_0[0], A_1[1], \ldots, A_n[n]\}$*

Example 2: The YSP [Hanks & McDermott, 1987] can be encoded by means of two temporal rules of the form:

$$\begin{array}{rcl} \mathit{load} & \longrightarrow & \mathit{loaded} \\ \mathit{shoot} \wedge \mathit{loaded} & \longrightarrow & \neg \mathit{alive} \end{array}$$

All variables are boolean, and *load* and *shoot* stand for actions. Given the observation $\mathit{alive}[0]$ and the actions $\mathit{load}[1]$ and $\mathit{shoot}[3]$, $\neg \mathit{alive}[4]$ follows. This is because $\phi^*_{\mathsf{loaded}}(s) = \{\mathbf{t}\}$ iff $\mathsf{load}^s = \mathbf{t}$ or $\mathsf{loaded}^s = \mathbf{t}$, while $\phi^*_{\mathsf{alive}}(s) = \{\mathbf{f}\}$ iff $\mathsf{shoot}^s = \mathsf{loaded}^s = \mathbf{t}$ or $\mathsf{alive}^s = \mathbf{f}$. Then, since in all models s_0, s_1, s_2, ..., load and shoot are true in the states s_1 and s_3 and false in all other states, we get that loaded^{s_1} must be true, loaded^{s_2} and loaded^{s_3} must also be true, and thus, alive^{s_4} must be false.

[5] The double implication for actions captures that actions not included in A must be false. A similar 'closed world assumption' for actions is implicit in approaches such as [Gelfond & Lifschitz, 1993; Reiter, 1991] in which actions are not represented as part of the state of the system.

[6] We assume that the domain description is consistent in the sense that no pair of rules associated with the same action have antecedents that are jointly satisfiable and consequents that are jointly unsatisfiable.

4 Causal and Temporal Rules Combined

We consider now the combination of temporal (causal) rules and (non-temporal) causal rules. Namely the *extended* temporal theories that we consider now are of the form $T = \langle D, C, A, O \rangle$ where $\langle D, A, O \rangle$ stands for a temporal theory as before, and C stands for a set of (non-temporal) causal rules. Temporal and causal rules play the role of *effect axioms* and *ramification constraints* [Lin & Reiter, 1994] respectively.

For example, the scenario in [Kartha & Lifschitz, 1994] that involves two switches and a light that is 'on' when the switches are in the same position can be modeled by a theory with *temporal rules:*

$$\begin{aligned}
&toggle1 \wedge sw1 \longrightarrow \neg sw1\\
&toggle1 \wedge \neg sw1 \longrightarrow sw1\\
&toggle2 \wedge sw2 \longrightarrow \neg sw2\\
&toggle2 \wedge \neg sw2 \longrightarrow sw2
\end{aligned}$$

and *causal rules:*

$$\begin{aligned}
&(sw1 \equiv sw2) \rightarrow on\\
&\neg(sw1 \equiv sw2) \rightarrow \neg on
\end{aligned}$$

The semantics of temporal theories extended in this way must combine the causal and temporal rules into a single causal function ϕ^*_{x} for each variable x. For simplicity, we follow other formalisms and assume that *all temporal rules are strict,* and *all variables are boolean.* These two assumptions decouple the causal rules from the temporal rules, allowing us to define the persistent causal function ϕ^*_{x} for fluent variables x (that now takes two sucessive states s_p and s_c as arguments) as:[7]

$$\phi^*_{\mathsf{x}}(s_p, s_c) \stackrel{\text{def}}{=} \begin{cases} \phi^t_{\mathsf{x}}(s_p) & \text{if } \phi^t_{\mathsf{x}}(s_p) \subset D_{\mathsf{x}} \\ \phi^c_{\mathsf{x}}(s_c) & \text{if } \phi^c_{\mathsf{x}}(s_c) \subset D_{\mathsf{x}} \\ \{\mathsf{x}^{s_p}\} & \text{otherwise} \end{cases} \tag{6}$$

This definition assumes that fluent variables persist when neither the temporal or causal rules tell us otherwise. The restrictions above can be removed by simply defining a causal function $\phi_{\mathsf{x}}(s_p, s_c)$ in terms of a single microtheory T_{x} that includes both causal and temporal rules.

Provided with the functions ϕ^*_{x}, and assuming that actions are exogenous (and hence that action variables do not occur in the head of rules), the resulting *temporal interpretations* can be defined as follows:

Definition 2 *A trajectory s_0, s_1, ... is a* temporal interpretation *relative to a set of temporal rules and causal rules if for each variable x, $\mathsf{x}^{s_0} \in \phi^c_{\mathsf{x}}(s_0)$, and $\mathsf{x}^{s_{i+1}} \in \phi^*_{\mathsf{x}}(s_i, s_{i+1})$ for each $i \geq 0$.*

The *models* of an *extended* temporal theory $T = \langle D, C, A, O \rangle$ are the temporal interpretations relative to the rules in D and C that are compatible with the actions A and the observations O.

[7] The conditions are to be evaluated top down.

Example 3. Using the rules above, we prove that if $sw1$ and $sw2$ are initially true and false respectively, and the action $toggle2$ is done, then $sw1$ stays true, $sw2$ becomes true, and hence on becomes true as well (see [Lin, 1995] for a similar example). Clearly, $\phi^*_{\mathsf{sw1}}(s_i, s_{i+1}) = \{\mathbf{t}\}$ iff either $toggle1 \wedge \neg sw1$ or $\neg toggle1 \wedge sw1$ is true in s_i. Samething similar holds for $\phi^*_{\mathsf{sw2}}(s_i, s_{i+1})$. Then since $toggle2[1]$ must be the only true action in the trajectory, $\mathsf{sw1}$ must remain true at all points, $\mathsf{sw2}$ must become true at time 2, and thus on must too become true at time 2.

5 Extensions

5.1 Non-persistent Causation

Rieger and Grinberg [Rieger & Grinberg, 1977] develop a language for representing causal mechanisms that, among other features, includes two different types of cause-effect relations: *continuous relations* in which the cause has to remain present in order to sustain the effect (e.g., $wet \supset hard\_to\_grasp$) and *one-shot relations* in which the effect remains true even after the cause that triggered the effect has vanished (e.g., $in\_water \supset wet$). In our model we have not made this distinction, yet this distinction is often needed.

Consider for instance a scenario from [Myers & Smith, 1988] cited in [Giunchiglia & Lifschitz, 1995]: if an object is put on the table it becomes safe for the baby, yet if removed from the table it won't necessarily remain safe. Such a scenario can be modeled by the temporal rules:

$$\begin{aligned}
put &\longrightarrow on_table\\
remove &\longrightarrow \neg on_table
\end{aligned}$$

and the causal rule:

$$on\_table \rightarrow safe$$

When the action *put* is performed, *on_table* and $safe$ become true, but if later *remove* is done, *on_table* becomes false but $safe$ remains true. That's clearly anomalous. The problem, using the terminology of Rieger and Grinberg, is that we are modeling all causal rules as *one-shot* rules, yet the causal rule $on\_table \rightarrow safe$ is a *continuous* rule; i.e., $safe$ is true *as long as* *on_table* is true. We call such rules, *non-persistent* rules.

Assuming now that all (non-temporal) causal rules are non-persistent, and for simplicity, that all rules are *strict,* the definition (6) of the function ϕ^*_{x} can be changed to:

$$\phi^*_{\mathsf{x}}(s_p, s_c) \stackrel{\text{def}}{=} \begin{cases} \phi^t_{\mathsf{x}}(s_p) \cap \phi^c_{\mathsf{x}}(s_c) & \text{if } \phi^t_{\mathsf{x}}(s_p) \cap \phi^c_{\mathsf{x}}(s_c) \subset D_{\mathsf{x}} \\ D_{\mathsf{x}} & \text{if } \phi^c_{\mathsf{x}}(s_p) \subset D_{\mathsf{x}} \\ \{\mathsf{x}^{s_p}\} & \text{otherwise} \end{cases} \tag{7}$$

where a variable x is *not* assumed to persist in states where it could have been caused by a non-persistent causal rule.[8]

[8] More accurate definitions are possible but the one above is simple and fits naturally into the format of the previous definition (6).

In the example above, the new definition means that $safe$ won't be assumed to persist in states s in which on_table holds. Yet if on_table persists so will be $safe$ due of the rule $on\_table \rightarrow safe$.

5.2 Action Constraints

Strict *causal* rules of the form $F \supset A$ stand for a *mechanism* by which interventions to make F true make A true as well. Other type of expressions have the *same logical form* but do not have this property For example, the constraint $eat \supset have\_food$ is always true yet we don't expect $have\_food$ to become true by performing the action eat. In other words, $eat \supset have\_food$ does not stand for any mechanism. Something similar occurs with the so-called integrity constraints in databases like $high\_salary \supset is\_a\_boss$. Indeed, while causal rules express conditions that are self-enforcing in the model (i.e., they are assumed to be enforced by a mechanism), constraints like $eat \supset have\_food$ and $high\_salary \supset is\_a\_boss$ express conditions that must be enforced by the agent. For that reason we call the latter constraints *action constraints* (they correspond to the qualification constraints in [Lin & Reiter, 1994]).

In order to accommodate action constraints in our model, let us say that a theory $T = \langle D, C, A, O\rangle$ complies with a set AC of *action constraints* if *every* state s_i of *every* model $s_0, s_1, \ldots$ of T satisfies those constraints. If so, we say that the theory $T' = \langle D, C, AC, A, O\rangle$ that results from incorporating those constraints is *consistent*. Otherwise we say that the theory is *inconsistent*.

Note that we do not prune the models that violate the constraints. When there are such models we simply declare the theory *inconsistent*. This is to comply with the so-called 'causality principle' in system theory [Padulo & Arbib, 1974] by which (exogenous) actions should not have an effect on the past. This principle holds for consistent theories but does not hold for inconsistent ones.

6 Summary

We have developed an approach to the ramification problem in terms of a simple and well-founded model of causal reasoning that is related to recent theories actions, Bayesian Networks and Structural Equation Models (see Section 2.2). A number of refinements and extensions that would be desirable are a more accurate model of non-persistent causation, constraints introducing *interactions* among mechanisms (e.g., a rigid bar among two moving blocks), and a more expressive language.

References

[Darwiche & Pearl, 1994] Darwiche, A., and Pearl, J. 1994. Symbolic causal networks. In *Proceedings AAAI-94*, 238–244. Seattle, WA: MIT Press.

[Druzdzel & Simon, 1993] Druzdzel, M., and Simon, H. 1993. Causality in bayesian belief networks. In *Proceedings UAI-93*, 3–11.

[Geffner, 1992] Geffner, H. 1992. *Default Reasoning: Causal and Conditional Theories.* MIT Press.

[Geffner, 1996] Geffner, H. 1996. A formal framework for causal modeling and argumentation. In *Proceedings FAPR'96*, 208–222. Springer Verlag.

[Gelfond & Lifschitz, 1993] Gelfond, M., and Lifschitz, V. 1993. Representing action and change by logic programs. *J. of Logic Programming* 17:301–322.

[Giunchiglia & Lifschitz, 1995] Giunchiglia, E., and Lifschitz, V. 1995. Dependent fluents. In *Proceedings IJCAI-95*, 1964–1969.

[Goldszmidt & Pearl, 1992] Goldszmidt, M., and Pearl, J. 1992. Rank-based systems In *Proceedings KR'92*, 661–672.

[Hanks & McDermott, 1987] Hanks, S., and McDermott, D. 1987. Non-monotonic logics and temporal projection. *AIJ* 33.

[Kartha & Lifschitz, 1994] Kartha, G. N., and Lifschitz, V. 1994. Actions with indirect effects (preliminary report). In *Proceedings KR-94*, 341–350.

[Lin & Reiter, 1994] Lin, F., and Reiter, R. 1994. State constraints revisited. *J. of Logic and Computation* 4(5):655–678.

[Lin, 1995] Lin, F. 1995. Embracing causality in specifying the indirect effects of actions. In *Proceedings IJCAI-95*, 1985–1991.

[McCain & Turner, 1997] McCain, N., and Turner, H. 1997. Causal theories of action and change. In *Proceedings AAAI-97.*

[Myers & Smith, 1988] Myers, K., and Smith, D. 1988. The persistence of derived information. In *Proceedings AAAI-88*, 496–500.

[Padulo & Arbib, 1974] Padulo, L., and Arbib, M. 1974. *System Theory.* Hemisphere Publishing Co.

[Pearl, 1988] Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann.

[Pearl, 1996] Pearl, J. 1996. Structural and probabilistic causality. *Motivation and Learning.*

[Reiter, 1991] Reiter, R. 1991. The frame problem in the situation calculus In Lifschitz, V., ed., *Artificial Intelligence and Mathematical Theory of Computation.* Academic Press.

[Rieger & Grinberg, 1977] Rieger, C., and Grinberg, M. 1977. The declarative representation and procedural simulation of causality in physical systems. In *Proceedings IJCAI-77*, 250–256.

[Simon, 1953] Simon, H. 1953. Causal ordering and identifiablity. In Hood, W., and Koopmans, T., eds., *Studies of the Econometric Method.* J. Wiley.