Ramifications and Sufficient Causes

Peter Grünwald
CWI, P.O. Box 4079
1009 AB Amsterdam
The Netherlands
pdg@cwi.nl

Abstract

Most 'causal' approaches to reasoning about action have not addressed the basic question of causality: what has to be the case in the world in order for the assertion 'A causes B' to be valid? Pearl's recent causal theories based on *structural equations* do provide an answer to this question. In this paper, we extend Pearl's formalism so that the typical problems encountered in common sense reasoning about action can be represented in it. The resulting theory turns out to be a powerful tool for handling the ramification problem. It also provides new insights into actions with non-deterministic and/or 'disjunctive' effects and it may help bridge the conceptual gap between 'causal' and 'non-causal' approaches to common sense temporal reasoning.

1 Introduction

Causality-based approaches to reasoning about action and change have attracted a lot of attention over the last few years [2, 13, 14, 15, 16, 17, 25], particularly in the context of modeling ramifications and actions with non-deterministic effects. Indeed, it has been argued by some [25] that the explicit modeling of causal relationships is a necessary ingredient of any solution to the ramification problem. Still, witness the frequently heated debates on the issue taking place during workshops and conferences, it seems that many researchers remain skeptical about the power and applicability of causal approaches. A reason for this may be that hardly any of these approaches attempts to define causal relationships in non-causal terms. On the other hand, the *causal theories* of J. Pearl [19, 20, 21, 22] do provide such a definition. However, as will be explained in section 2, Pearl's theory as such cannot be directly applied to common-sense reasoning about action. In this paper, we extend Pearl's theory in a straightforward way so as to overcome this hindrance. We thus arrive at a causal theory that has the advantage that it can also be understood in non-causal terms, namely as a theory that gives a certain semantics to 'actions'. The theory turns out to provide a simple solution for many of the notorious problems of common-sense temporal reasoning . Moreover, we find that the resulting theory is remarkably similar to several existing causal approaches, most notably Thielscher's [25] and Lin's [13, 14]. In this paper, we concentrate on application of the theory to the *ramification* *problem* [24] and to the problem of representing actions with disjunctive effects. The theory however is also quite capable of handling causal chains of events, concurrent actions, incompletely specified action schedules and surprises. Discussion of those features will be left for another paper; in [8, 9, 10] we present preliminary results in that direction.

We will now first provide an informal introduction to Pearl's semantics of causation. We then introduce a simple, propositional extension of Pearl's theory. In section 4 we instantiate our theories to domains involving persistence. This allows us to handle domains with complicated ramifications and actions with disjunctive effects, as we show in sections 5 and 6. The paper ends with some concluding remarks.

2 An Informal Introduction to Pearl's Causal Theories

Pearl's causal theories are about domains of variables whose values may be influenced by *interventions*. Each variable X in the domain depends on a subset $S(X)$ of all the other variables. If an intervention takes place that sets the value of a variable X to some value, then some of the variables in $S(X)$ become independent of X while the rest of the variables in $S(X)$ keep their dependence. If we equate an *action* or *event* with a *set of interventions*, this induces a semantics for the concepts of 'action' and 'event'. This resulting semantics, which will be discussed in more detail below, has sometimes been called *the sufficient cause principle* [4, 5]. An action that sets the value of a variable X is then called a *sufficient cause* for X.

We give a simple example: Let us denote by $\mathit{Alive}(t)$ whether or not some turkey is alive at time t. Normally, i.e. if no interventions take place, we have that if it is given that the turkey is alive at time $t+1$, we can deduce that it will also be alive at time $t+2$ and that it already had been alive at time t. Now if an intervention takes place that sets $\mathit{Alive}(t+1)$ to FALSE (for example, if somebody shoots at the turkey), then the value of *Alive* at time $t+1$ becomes independent of the value of *Alive* at time t: we can no longer infer anything on whether or not the turkey was alive at time t (disappearance of a dependency). On the other hand, the dependency between $\mathit{Alive}(t+1)$ and $\mathit{Alive}(t+2)$ has not disappeared: we can still infer that the turkey will *not* be alive at time $t+2$.

One can represent theories of systems which are subject to interventions either by using *causal graphs* (as done in, e.g., [4, 5]) or by using sets of *structural equations*, as done in [21, 22]. Here, we opt for the latter possibility. Structural equations are equations with a directionality attached to them. A set of structural equations for our example would look as follows: $\{\mathit{Alive}(1) = \mathit{Alive}(0)\ ;\ \mathit{Alive}(2) = \mathit{Alive}(1)\ ; \ldots\}$. Our action semantics, i.e. the sufficient cause principle, can now be rephrased like this: if an intervention takes place that sets variable X to value x, then the structural equation with X on the *left* is replaced by the new equation $X = x$. In our example, if just one intervention takes place, and this intervention sets $\mathit{Alive}(2)$ to FALSE, then the equation $\mathit{Alive}(2) = \mathit{Alive}(1)$ gets replaced by $\mathit{Alive}(2) = \text{FALSE}$, while all other structural equations remain unchanged.

Why we need to extend Pearl's theories

In common-sense reasoning about action, we are typically interested in domains where the fact whether or not an intervention takes place may depend on the values of some of the variables in the domain; this is impossible to model in Pearl's [21, 22] basic theory. Pearl's theory also lacks the representational power needed to express *global* domain constraints. For example, we may have $Alive(t) \equiv \neg Dead(t)$ irrespective of any action that influences the value of *Alive*; this formula should always hold in our domains and should not be replaced when *Alive* is set to FALSE. In our work we extend Pearlian causal theories in a way that allows us to represent both kinds of domain knowledge mentioned above. We do this by simply adding to the set of structural equations a set of *global constraints*, consisting of formulas that should hold in all models of our theories irrespective of what interventions take place. These global constraints then are also allowed to mention interventions. To be able to express those, we introduce for each propositional variable X in our domain two additional propositional variables $Do(X, \text{TRUE})$ and $Do(X, \text{FALSE})$, representing interventions setting X either to TRUE or to FALSE.

3 Propositional Causal Theories

Our propositional causal theories are a simple extension of Pearl's causal theories. In the long version of this paper, we will show exactly how they are related to Pearl's causal theories as defined in [21, 22]. The interested reader may check that the definitions below are straightforward extensions of definition 1 in [21] and [22].

Let $\mathbf{B} = \{\text{TRUE}, \text{FALSE}\}$. A *truth value* is an element of $\mathbf{B}$. Let $\mathbf{X}$ be a set of propositional variables, i.e. variables taking on truth values. We assume a standard propositional language for $\mathbf{X}$ containing the additional symbols TRUE and FALSE. A *valuation* or *interpretation* of $\mathbf{X}$ is an assignment of truth values to the variables in $\mathbf{X}$. A valuation $\mathcal{M}$ is a *model* for a set of propositional formulas Γ (written as $\mathcal{M} \models \Gamma$) if Γ is true in $\mathcal{M}$; here truth of propositional formulas with respect to interpretations is defined as usual.

Definition 1 *A propositional causal theory T is a 4-tuple $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS} \rangle$ where*

1. $\mathbf{V} = \{X_1, \ldots, X_n\}$ *is a set of observed propositional variables*
2. $\mathbf{U} = \{U_1, \ldots, U_m\}$ *is a set of unobserved propositional variables*
3. EQ *is a set of* structural equations*, i.e. propositional formulas of the form*

$$X_i \equiv \Phi_i(X_1, \ldots, X_n, U_1, \ldots, U_m) \tag{1}$$

   where Φ_i is a formula involving zero or more of the variables $X_1, \ldots, X_n, U_1, \ldots, U_m$.

4. CONS*, the set of* constraints*, is a finite set of propositional formulas over the variables $\mathbf{V} \cup \mathbf{U} \cup \mathcal{A}(\mathbf{V})$. Here $\mathcal{A}(\mathbf{V})$ is defined as the set of propositional variables*

$$\{Do(X_i, \text{TRUE}), Do(X_i, \text{FALSE}) \mid X_i \in \mathbf{V}\}$$

Notice that expressions of the form '$Do(X_i, b)$' simply stand for propositional variables here. The set $\mathbf{U}$ can be used to model external influences we are ignorant about or that we consider unlikely; we will only need it in section 6.

Example 1 Let $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS} \rangle$ be a causal theory with $\mathbf{V} = \{Alive(0), Alive(1)\}$, $\mathbf{U} = \emptyset$, $\text{EQ} = \{Alive(1) \equiv Alive(0)\}$, $\text{CONS} = \{Alive(0), \neg Do(Alive(1), \text{FALSE}), \neg Do(Alive(1), \text{TRUE})\}$. The equations in EQ express that the value of *Alive* persists (from time 0 to 1) if there are no interventions; the formulas in CONS ensure that $Alive(0)$ should hold in all models of T and that there is no intervention at time 1.

We now introduce the notion of models for causal theories. Condition (1) below simply ensures that all constraints hold in all models; condition (2) implements the 'sufficient cause principle'.

Definition 2 *A model for a propositional causal theory* $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS} \rangle$ *is a valuation* $\mathcal{M}$ *for the variables in* $\mathbf{V} \cup \mathbf{U} \cup \mathcal{A}(\mathbf{V})$ *such that*

1. $\mathcal{M} \models \text{CONS}$
2. *The restriction of* $\mathcal{M}$ *to the variables in* $\mathbf{V} \cup \mathbf{U}$ *is a model for the set of equations* EQ'*, i.e.* $\mathcal{M} \models \text{EQ}'$*. Here* EQ' *is obtained from* EQ *as follows:*

   For all $X_i \in \mathbf{V}, b \in \mathbf{B}$ *such that* $\mathcal{M} \models Do(X_i, b)$*, we* delete *(if present) from* EQ *the equation* $X_i \equiv \Phi_i(\ldots)$ *and we* add *the equation* $X_i \equiv b$.

If $\mathcal{M}$ is a model for a *causal* theory T, we write $\mathcal{M} \models_{\mathbf{c}} T$.

Example 1, continued Let T be as in example 1 above. Condition (1) of definition 2 ensures that all models $\mathcal{M}$ for T have $\mathcal{M} \models Alive(0)$ and that for all b, $\mathcal{M} \models \neg Do(Alive(1), b)$. By condition (2), $\text{EQ}' = \text{EQ}$ and so we also have $Alive(1)$ in all models for T: *Alive* persists. Now consider a theory T' identical to T except that CONS is now equal to $\text{CONS} = \{Alive(0), Do(Alive(1), \text{FALSE})\}$. By condition 1 of definition 2 we now have $\mathcal{M} \models Alive(0) \wedge Do(Alive(1), \text{FALSE})$ for all causal models $\mathcal{M}$. By condition 2 of that definition, we have $\text{EQ}' = \{Alive(1) \equiv \text{FALSE}\}$ and thus all models for T' must have $\neg Alive(1)$.

4 Causal Theories Involving Persistence

We now instantiate our causal theories to domains consisting of *fluents*, which are properties of the world that are subject to persistence, and *events*, which will stand for the 'triggers' of interventions. Causal theories for these domains will thus be defined with respect to a finite set of fluents $\mathbf{F}$ and a finite set of events $\mathbf{E}$. Specifically, we assume that for any causal theory for the sets $\mathbf{E}$ and $\mathbf{F}$, the set of observables $\mathbf{V}$ can be partitioned into two subsets: $\mathbf{V_E}$, the set of *event – time pairs*, and $\mathbf{V_F}$, the set of *fluent – time pairs*. In this short paper, we restrict ourselves to *2-state causal theories* which represent domains using only two points in time: the initial point in time $t = 0$ and the final point in time $t = 1$. Given the set $\mathbf{F} = \{F_1, \ldots, F_n\}$, $\mathbf{V_F}$ can now be written as $\mathbf{V_F} = \{F_1(0), \ldots, F_n(0), F_1(1), \ldots, F_n(1)\}$. Here $F_i(t)$ denotes fluent i at time t. We assume that all events we are interested in happen at the same time, namely directly after the initial point in time. Hence we do not have to index events by time points and can simply let $\mathbf{V_E} = \mathbf{E}$. The *initial state* of a model $\mathcal{M}$ for a causal theory for sets $\mathbf{E}$ and $\mathbf{F}$ is defined to be the set $\{F \mid \mathcal{M} \models F(0)\} \cup \{\neg F \mid \mathcal{M} \models \neg F(0)\}$. The *final state* of a model $\mathcal{M}$ is the set $\{F | \mathcal{M} \models F(1)\} \cup \{\neg F | \mathcal{M} \models \neg F(1)\}$.

Example 2 We enrich our turkey domain with a new fluent *Dark* and an event *Shoot*. $Dark(t)$ will denote that it is dark at time t; a *Shoot*-event will end the turkey's life, but only if it is not dark (so that one can aim properly). For this, let T be a causal theory such that $\mathbf{V_E} = \{Shoot\}$, $\mathbf{V_F} = \{Alive(0), Dark(0), Alive(1), Dark(1)\}$; $\mathbf{U} = \emptyset$. EQ contains two equations: $\{Alive(1) \equiv Alive(0)\ ;\ Dark(1) \equiv Dark(0)\}$. CONS contains the single axiom

$$[\ \neg Dark(0) \wedge Shoot\ ]\ \supset Do(Alive(1), \text{FALSE}) \tag{2}$$

Let us look at the set of valuations $\mathbf{M}$ in which no *Shoot*-event and no interventions take place; i.e. we have $\neg Shoot$ and $\neg Do(X, b)$ for any $X \in \mathbf{V}, b \in \mathbf{B}$. Clearly, for any member $\mathcal{M}$ of $\mathbf{M}$ we have $\mathcal{M} \models$ CONS; hence condition (1) of definition 2 is satisfied. Since, according to condition (2) of that definition, the updated set of equations EQ$'$ is equal to EQ, the valuations in $\mathbf{M}$ which also satisfy this condition are exactly those in which the values of both *Alive* and *Dark* persist: in the models in which no interventions take place, we have persistence, which is in accord with intuition.

We need more!

Let us now look at the models for T in which interventions do occur. One can see there is a model $\mathcal{M}$ among them with

$$\mathcal{M} \models Dark(0) \wedge \neg Dark(1) \wedge Do(Dark(1), \text{FALSE}),$$

since both conditions (1) and (2) of def. 2 are satisfied for $\mathcal{M}$. This is not what we would intuitively expect! Intuitively, we would reason as follows: 1) there are no interventions except possibly those triggered by the *Shoot*-event ; 2) *Shoot* does not affect *Dark*, so the value of *Dark* should remain unchanged after the *Shoot*-event.

What we have yet forgotten to model are the implicit assumptions behind 1) and 2), namely that 1) *there are no external interventions* and that 2) *actions affect no more things in the world than those we explicitly state they affect.* How should we incorporate these assumptions in our theories? To answer this question, we must first know under exactly what conditions a fluent is affected, or equivalently, under what conditions an intervention takes place that sets it to some value. We note that this always happens in a specific *context*: for example, $Alive(1)$ is set to FALSE only in a context in which *Shoot* and $\neg Dark(0)$ is true. In general, such a context should contain *every fact in the world that can possibly influence whether or not an intervention takes place.* This means that we may identify contexts with specific interpretations of the variables in $\mathbf{V}$ and $\mathbf{U}$. Now suppose we are in one specific context; let us call it C. If there is a model $\mathcal{M}_1$ with this context and with $\mathcal{M} \models Do(X_i, b)$ while there is also a model $\mathcal{M}_2$ with the same context C and with $\mathcal{M} \models \neg Do(X_i, b)$, then we can be sure that nothing in our axioms states that in context C the fluent X_i is set to value b. By assumptions 1) and 2) above, we should now prefer model $\mathcal{M}_2$. Apparently, we should partition our models into classes sharing the same context, i.e. the same interpretation of $\mathbf{V}$ and $\mathbf{U}$, and then within each such class pick the models with a minimal (in the subset sense) interpretation of the *Do*-variables, i.e. the variables in $\mathcal{A}(\mathbf{V})$. Here the set $\mathcal{A}(\mathbf{V})$ is defined as in def. 1. But notice that we must do this *before* we replace the structural equation set EQ by EQ$'$ (step 2 of def. 2), since this replacement will only work if the *Do*-propositions already have the right interpretations in each model!

The required minimization can be conveniently implemented[1] using *circumscription*, [12, 18]: we circumscribe all *Do*-variables at the same time while keeping everything else (i.e. the variables in $\mathbf{V} \cup \mathbf{U}$) fixed:

$$\overline{\text{CONS}} = \mathit{Circum}(\text{CONS}; \{Do(X,b) | X \in \mathbf{V}, b \in \mathbf{B}\}) = \mathit{Circum}(\text{CONS}; \mathcal{A}(\mathbf{V}))$$

For details about circumscription, see for example [12]. Suffice it to say here that it will indeed give us what we want – to see this, we need a few definitions: Let $\mathbf{X}$ be a set of variables and $\mathbf{Y}$ and $\mathbf{Z}$ be subsets of $\mathbf{X}$ such that $\mathbf{Y} \cap \mathbf{Z} = \emptyset$. For any interpretation $\mathcal{M}$ for the variables in $\mathbf{X}$ we write $\mathcal{M}\rfloor_{\mathbf{Z}}$ to denote the restriction of $\mathcal{M}$ to $\mathbf{Z}$. Let Γ be a set of formulas over $\mathbf{X}$. We call a model $\mathcal{M}$ for Γ *minimal* for $\mathbf{Y}$ in context $\mathbf{Z}$ if there is no other model $\mathcal{M}'$ for Γ such that $\mathcal{M}'\rfloor_{\mathbf{Z}} = \mathcal{M}\rfloor_{\mathbf{Z}}$ while $\{Y \in \mathbf{Y} \mid \mathcal{M}' \models Y\} \subset \{Y \in \mathbf{Y} \mid \mathcal{M} \models Y\}$. Here '$\subset$' denotes proper inclusion. Minimizing *Do* then amounts to selecting the models that are minimal for $\mathcal{A}(\mathbf{V})$ in context $\mathbf{U} \cup \mathbf{V}$. This is exactly what the circumscription achieves:

Proposition 1 *For any interpretation $\mathcal{M}$ for the variables in $\mathbf{V} \cup \mathbf{U} \cup \mathcal{A}(\mathbf{V})$:*

$$\mathcal{M} \models \overline{\text{CONS}} \Leftrightarrow \mathcal{M} \textit{ is minimal for } \mathcal{A}(\mathbf{V}) \textit{ in context } \mathbf{V} \cup \mathbf{U}.$$

The proposition is a simple corollary of proposition 1 in [12] – we omit the proof. It is important to realize that the circumscription can never completely rule out any interpretations of $\mathbf{V}$ and $\mathbf{U}$: if $\mathcal{M} \models \text{CONS}$, then there must be an $\mathcal{M}' \models \overline{\text{CONS}}$ such that $\mathcal{M}\rfloor_{\mathbf{V}\cup\mathbf{U}} = \mathcal{M}'\rfloor_{\mathbf{V}\cup\mathbf{U}}$.

From now on we assume that all our causal theories are such that *Do* still has to be minimized. This brings us to the final definition of 2-point causal theories and their models:

Definition 3 *A 2-point causal theory for the set of fluents* $\mathbf{F}$ *and the set of events* $\mathbf{E}$ *is a propositional causal theory* $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS} \rangle$ *such that*

- $\mathbf{V} = \mathbf{V_F} \cup \mathbf{V_E}$ *with* $\mathbf{V_F} = \{F_i(0), F_i(1) \mid F_i \in \mathbf{F}\}$ *and* $\mathbf{V_E} = \mathbf{E}$.
- $\text{EQ} = \{F_i(1) \equiv F_i(0) \mid F_i \in \mathbf{F}\}$

Notice that the following definition only differs from definition 2 in its first condition.

Definition 4 *A model for a 2-point causal theory* $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS} \rangle$ *is a valuation* $\mathcal{M}$ *for the variables in* $\mathbf{V} \cup \mathbf{U} \cup \mathcal{A}(\mathbf{V})$ *such that*

1. $\mathcal{M} \models \mathit{Circum}(\text{CONS}; \mathcal{A}(\mathbf{V}))$
2. *The restriction of* $\mathcal{M}$ *to the variables in* $\mathbf{V} \cup \mathbf{U}$ *is a model for the set of equations* EQ'*, where* EQ' *is obtained from* EQ *as follows:*

   For all $X_i \in X, b \in \mathbf{B}$ *such that* $\mathcal{M} \models Do(X_i, b)$*, we* delete *(if present) from* EQ *the equation* $X_i \equiv \Phi_i(\ldots)$ *and we* add *the equation* $\Phi_i \equiv b$.

We are now ready to harvest.

[1]Circumscription has originally been defined for first-order logic; its definition [12] also applies to the propositional case though, by regarding each propositional variable as a 0-ary predicate.

5 Ramifications

The *ramification problem* [11, 13, 15, 17, 24, 25] can be defined as the problem of how to correctly represent the indirect effects of actions in a concise manner. Our *Do*-variables enable us to express any propositional combination of observations (e.g. $\neg Alive(1)$) and interventions (e.g. $Do(Alive(1), \text{FALSE})$). We will see that this is exactly what we need to express complicated domain constraints involving ramifications; specifically, we will often use axioms of the form $Do(F_1, b) \supset Do(F_2, b')$. Such axioms can be interpreted as saying that *any intervention in the domain that sets the value of F_1 to b also sets the value of F_2 to b'*. We start by showing how our approach handles two well-known examples:

Example 3 [The Walking Turkey] [15, 11, 25] Let us assume that if our turkey is alive, it may also possibly be walking. If we shoot the turkey while it is walking, it will stop doing so. Hence $\neg Walking$ is an indirect effect of *Shoot*, or equivalently a *ramification* of $\neg Alive$. However, if the turkey is not alive, and we try to make it walking, for example, by *Entice*-ing it, then it should *not* start walking: *Alive* can never be an indirect effect of *Entice*. This is modeled using a 2-point causal theory T_{WT} for the sets $\mathbf{E} = \{Shoot, Entice\}$ and $\mathbf{F} = \{Alive, Walking\}$; $\mathbf{U} = \emptyset$; $\text{EQ} = \{Alive(1) = Alive(0) \ ; \ Walking(1) = Walking(0)\}$. CONS consists of the following four axioms:

$$Shoot \supset Do(Alive(1), \text{FALSE}) \tag{3}$$

$$Entice \supset Do(Walking(1), \text{TRUE}) \tag{4}$$

$$\neg Alive(t) \supset \neg Walking(t) \tag{5}$$

$$Do(Alive(t), \text{FALSE}) \supset Do(Walking(t), \text{FALSE}) \tag{6}$$

Here axioms of the form $\phi(t)$ should be read as $\phi(0) \wedge \phi(1)$. Let us denote by $\mathbf{M}$ the set of models for T_{WT} in which *Shoot* takes place while *Entice* does not; by axioms (3) and (6) we have $Do(Alive(1), \text{FALSE})$ and $Do(Walking(1), \text{FALSE})$ in all such models. By condition (1) of definition 4, we also have $\neg Do(X_i, b)$ for all other (X_i, b). Hence, the updated set of structural equations EQ' in condition (2) of that definition becomes:

$$Alive(1) \equiv \text{FALSE} \quad ; \quad Walking(1) \equiv \text{FALSE}$$

So all models in $\mathbf{M}$ have as their final state $\{\neg Alive, \neg Walking\}$. It is easy to see that such models indeed exist, by simply checking that all axioms in CONS and EQ' hold in them and by noting that the minimization of *Do* never rules out any specific interpretations of $\mathbf{V} \cup \mathbf{U}$. We see in the same way that $\mathbf{M}$ contains models for three different initial states: $\{Alive, Walking\}$, $\{Alive, \neg Walking\}$ and $\{\neg Alive, \neg Walking\}$. So the case in which the domain constraints should entail a ramification of the *Shoot*-event works fine. Now for the models in which no *Shoot* but an *Entice* event takes place. All such models have $Do(Walking(1), \text{TRUE})$. By the minimization of *Do* in condition (1) in def. 4 they also have $\neg Do(Alive(1), \text{TRUE})$. Hence the set EQ' in condition (2) of that definition becomes

$$Alive(1) \equiv Alive(0) \quad ; \quad Walking(1) \equiv \text{TRUE}$$

By axiom (5) all models must then also have $\mathcal{M} \models Alive(1)$, and thus also (since EQ' still contains the persistence equation for *Alive*), $\mathcal{M} \models Alive(0)$. This means that there are no models for T_{WT} with both $Alive(0)$ and *Entice*: *Entice* has *Alive* as an *implicit qualification*.

Example 4 The Suitcase & Switches Problem The following example is equivalent to Lin's [13] 'suitcase problem'. The version we give here is adapted from [25]. Imagine a light that is connected to two switches; the light is only on if both of the switches are in the *on*-position: any event that puts a switch in the *on*-position in a context in which the other switch is on, will have as a ramification that the light goes on. However, if the event takes place in a context in which the other switch is not on, the light will not be affected. We formalize this using event and fluent sets $\mathbf{E} = \emptyset$ and $\mathbf{F} = \{Swi_1, Swi_2, Light\}$. Swi_i denotes that switch i is in the *on*-position. Let T_{SW} be a causal theory for $\mathbf{E}$ and $\mathbf{F}$ with $\mathbf{U} = \emptyset$, EQ as usual and CONS as follows:

$$\begin{aligned}
[\ Swi_1(t) \wedge Do(Swi_2(t), \text{TRUE})\ ] \quad &\supset \quad Do(Light(t), \text{TRUE}) && (7)\\
[\ Swi_2(t) \wedge Do(Swi_1(t), \text{TRUE})\ ] \quad &\supset \quad Do(Light(t), \text{TRUE}) && (8)\\
Do(Swi_1(t), \text{FALSE}) \quad &\supset \quad Do(Light(t), \text{FALSE}) && (9)\\
Do(Swi_2(t), \text{FALSE}) \quad &\supset \quad Do(Light(t), \text{FALSE}) && (10)\\
(Swi_1(t) \wedge Swi_2(t)) \quad &\equiv \quad Light(t) && (11)
\end{aligned}$$

Here axiom (11) describes a domain constraint that must hold in all models of the domain; the other axioms describe the 'dynamics' of the domain. Let us see whether we get the expected models for T_{SW}: suppose first we turn on the first switch while the second one is off; i.e. we add the axiom $\neg Swi_2(0) \wedge Do(Swi_1(1), \text{TRUE})$ to CONS. By the minimization of Do, we get $\neg Do(Swi_2(1), b)$ for all b in all models. Hence EQ' contains $Swi_2(1) \equiv Swi_2(0)$, so all models for T_{SW} must have $\neg Swi_2(1)$. But in all models with $\neg Swi_2(1)$, we already have (again by the minimization of Do) $\neg Do(Light(1), \text{TRUE})$ and $\neg Light(1)$. It follows that all models will have $\neg Light$ in their final state; it is easy to see that such models indeed exist. One can show in a similar manner that one obtains the intended models in all other possible scenarios.

5.1 A Glimpse at other approaches

We have seen in the introduction that, in Pearl's terminology, the proposition $Do(X_i, b)$ can be read as 'there is a sufficient cause for X_i to have the value b'. Indeed, usage of '*Do*' in axioms often corresponds to the colloquial use of the word 'causes'. It turns out that if the word *Do* is replaced by the word `causes`, then our axioms start to look very similar to those used in various other approaches. We leave detailed comparisons for the longer version of this paper and give only a suggestive example for now. In Thielscher's work, the equivalent of axiom (6) looks as follows ([25], page 330):

$$\neg Alive \ \texttt{causes} \ \neg Walking$$

In Lin's work [13, 14], we would get:

$$Caused(Alive, \text{FALSE}, s) \supset Caused(Walking, \text{FALSE}, s)$$

In McCain & Turner's approach [15, 16, 17], it would be formalized as:

$$\neg Alive \Rightarrow \neg Walking \tag{12}$$

where '⇒' is to be read as 'if $\neg Alive$, then the fact that $\neg Walking$ is caused' [16]. McCain & Turner's semantics interprets (12) as something like

$$\neg Alive(t) \supset Do(Walking(t), \text{FALSE}) \tag{13}$$

In the turkey example, replacing (6) by (13) does not make any difference: it is easy to show that one obtains exactly the same models in both cases. However, as we will see below, in general, $(X \equiv b_1) \supset Do(Y, b_2)$ is *not* equivalent to $Do(X, b_1) \supset Do(Y, b_2)$.

5.2 Causal Cycles

We have just seen that $Do(X, \text{TRUE}) \supset Do(Y, \text{TRUE})$ may often be read as 'X causes XY'. Several authors [11, 23] have pointed out that we do not always want to reason in the 'causal direction' that is implied by the above; for example, consider the switches domain (example 4) again: if we know that the light has been put off, we may want to conclude that (at least) one of the two switches has been put off too. We model this as follows:

Example 5 Let $T_{\text{sw},2}$ be as T_{sw} but with the following axioms added to CONS:

$$Do(Light(t), \text{FALSE}) \supset [\ Swi_1(t) \supset Do(Swi_2(t), \text{FALSE})\ ] \tag{14}$$
$$Do(Light(t), \text{FALSE}) \supset [\ Swi_2(t) \supset Do(Swi_1(t), \text{FALSE})\ ] \tag{15}$$
$$Do(Light(t), \text{TRUE}) \supset Do(Swi_1(t), \text{TRUE}) \tag{16}$$
$$Do(Light(t), \text{TRUE}) \supset Do(Swi_2(t), \text{TRUE}) \tag{17}$$

We show first that if we take CONS as above and do not add anything else, then we get persistence: one easily checks that for *any* interpretation $\mathcal{M}$ with $\mathcal{M} \models$ CONS, there is an interpretation $\mathcal{M}'$ that interprets all variables in $\mathbf{V} \cup \mathbf{U}$ the same way but which has $\mathcal{M}' \models \neg Do(X_i, b)$ for all X_i and b. So the circumscription of Do in the definition of causal models will make sure that all models have $\neg Do(X_i, b)$ for all X_i and b. Hence $\text{EQ}' = \text{EQ}$ and all fluents in our domain must persist.

Now suppose we turn on the light in an initial state with both switches *off*, i.e. we further add the following axiom to CONS:

$$Do(Light(1), \text{TRUE}) \land \neg Swi_1(0) \land \neg Swi_2(0) \tag{18}$$

Axioms (16) and (17) make sure that we have $Do(F(1), \text{TRUE})$ for $F \in \{Swi_1, Swi_2, Light\}$; so EQ' becomes $F(1) \equiv \text{TRUE}$ for all these F, and all models get final state $\{Swi_1, Swi_2, Light\}$; it is easy to see that models with such a final state indeed exist, so we get exactly the result we wanted. The case of turning off the light will be discussed in example 7.

The example above leads to a problem for some of the earlier causal approaches (i.e. those of Gustafsson & Doherty and McCain & Turner ; see [11] for the details). The reason is that in these approaches, constructions similar to $X \supset Do(Y, \text{TRUE})$ are used at places where one should use $Do(X, \text{TRUE}) \supset Do(Y, \text{TRUE})$; see the remark at the end of section 5.1. Had we chosen the first possibility, axioms (7) and (8) would have been replaced by

$$[\ Swi_1(t) \land Swi_2(t)\ ] \supset Do(Light(t), \text{TRUE})$$

while axioms (16) and (17) would have become:

$$Light(t) \supset [\ Do(Swi_1(t), \text{TRUE}) \wedge Do(Swi_2(t), \text{TRUE})]$$

But this would allow us to conclude that each model with final state $\{Light\ ;\ Swi_1\ ;\ Swi_2\}$ also has $Do(F(1), \text{TRUE})$ for F equal to any of the three fluents: there is an 'automatic' intervention that sets their value to TRUE. Hence the set EQ' in condition (2) of definition 4 becomes $F \equiv \text{TRUE}$ for all these F and all three persistence relations are broken. It follows that no matter what the initial state is, there is always a model with the final state above: there is a cycle involved in that the fact that the light is on implies an intervention that puts both switches on, while the fact that the switches are on implies an intervention that the light is put on. We have already seen that in our own formalization (the one using axioms (14)-(17)) this cannot happen.

Thielscher ([25], page 329) solves the problem in a way similar to ours, but since his equivalent of *Do* is called `causes`, the corresponding rules look somewhat strange: we obtain rules like '*Light* `causes` Swi_1', while, intuitively, turning on the light does not cause turning on the switch. Related observations have led some authors to claim that there is more to ramification than mere causal relationships – as stated in Gustafsson & Doherty, 'physical causality is simply one of several reasons one might set up dependencies between fluents' [11].

We think that our 'Pearlian' approach sheds some new light on this issue: we define causal relations (i.e. relations involving interventions) fully in non-causal terms; this is reflected in choosing the name *Do* rather than *Causes* for our interventions. Formulas of the form $Do(A, \text{TRUE}) \supset Do(B, \text{TRUE})$ may or may not correspond to the colloquial statement 'A causes B' or to any notion of 'physical' causality. The only thing we care about is how the world works at the level of detail at which we want to formalize it - and if at that level of detail, we have $Do(A, \text{TRUE}) \supset Do(B, \text{TRUE})$, then we may say that A causes B within the constraints of our domain – but if you do not like the word 'causes', you do not have to use it - the important thing about our theory is the semantics that it gives to interventions.

6 Disjunctive Effects and Nondeterminism

Example 5 raises the interesting question of how to formalize actions whose effect is a *disjunction* of fluents. Recently, some authors have, either explicitly or implicitly, begun to take up this question [17, 25]. A special case of this question concerns the more well-known issue of representing actions with non-deterministic effects [1, 17, 24]. If an event occurs that sets the value of a proposition X to TRUE, where $X \equiv Y \vee Z$, it is not immediately clear what should happen: should we exempt both Y and Z from persistence, or should we keep as much persistence as is logically consistent? In example 5 we implicitly chose the latter option: setting *Light* to FALSE in an initial situation with both switches on will always keep one switch in the *on*-position: as will be shown below, there will be no final states with both $\neg Swi_1$ and $\neg Swi_2$. But is this correct, i.e. is it what we intuitively expect? A small poll conducted by the present author reveals that people have differing intuitions about this: some think that the position of both switches should be exempted from persistence, while others prefer the 'inert models' where there are as few changes as possible. We think that this clash of intuitions is not so surprising in light of Pearl's theory: if we intervene to set the value of a variable X that can be regarded as a disjunction of variables Y and Z, then we simply have an *incomplete specification* of our problem: it is not clear what should happen to the structural equations for Y and Z – should they both be removed from the set EQ, or should we remove as few equations as possible? In our view, this may change from domain to domain. We will therefore *not* attempt to extend the semantics of *Do* to constructions of the form $Do(Y \vee Z, b)$, but rather always make explicit what should happen to the structural equations of Y and Z if an intervention takes place. We can do so in our causal theories by using the *assumption symbols* in $\mathbf{U}$ (see def. 1). We illustrate their use by considering a domain[2] which clearly asks for eliminating all structural equations involved in the disjunction.

Example 6 We drop a pencil on a table with a piece of paper on it. As a result, the pencil may either lie fully on the paper, or touch both part of the paper and part of the table's surface, or touch only the table's surface. We write $Touches\_table(t)$ ($Touches\_paper(t)$) iff the pencil touches the surface of the table (the paper) at time t. We can model this using a causal theory with $\mathbf{E} = \{Drop\}$, $\mathbf{F} = \{Touches\_table, Touches\_paper\}$, $\mathbf{U} = \{U_1, U_2\}$, EQ as usual and CONS as follows:

$$\neg Touches\_table(0) \wedge \neg Touches\_paper(0) \tag{19}$$

$$\begin{aligned} Drop \supset [& \ (Do(Touches\_table(1), \text{TRUE}) \wedge U_1) \vee \\ & \ (Do(Touches\_paper(1), \text{TRUE}) \wedge U_2)] \end{aligned} \tag{20}$$

In this case, among the models with *Drop*, there will be three subclasses, corresponding to the three extensions of (U_1, U_2) that are consistent with (20). By the minimization of *Do* in definition 4 and by axiom (20), we get the following interpretation for each class of models:

`class`	`corresponding interpretation of` *Do*	
$U_1 \wedge U_2$	$Do(Touches\_table(1), \text{TRUE}) \wedge Do(Touches\_paper(1), \text{TRUE})$	
$U_1 \wedge \neg U_2$	$Do(Touches\_table(1), \text{TRUE}) \wedge \neg Do(Touches\_paper(1), \text{TRUE})$	
$\neg U_1 \wedge U_2$	$\neg Do(Touches\_table(1), \text{TRUE}) \wedge Do(Touches\_paper(1), \text{TRUE})$	(21)

It follows that there will be both models where only the persistence of *Touches_table* or *Touches_paper* is broken and models in which the persistence of both is broken.

Example 7 Let us now consider a case in which we want 'as much persistence as possible'. Let us say there is an intervention which sets the value of variable X where $X \equiv Y \vee Z$. If we let $X \equiv \neg Light$, $Y \equiv \neg Swi_1$ and $Z \equiv \neg Swi_2$, then we can see that we have already implicitly treated this case in the extended switches domain, example 5[3]. Consider a scenario with an intervention setting *Light* to FALSE, i.e. we add the following axiom to the set CONS of $T_{\text{sw},2}$:

$$Do(Light(1), \text{FALSE}) \wedge Light(0)$$

We see that in the models with $Swi_1(1)$ and $\neg Swi_2(1)$ we have $Do(Swi_2(1), \text{FALSE})$ by axiom (14) and $\neg Do(Swi_1(1), \text{FALSE})$ by the minimization of *Do*. Then EQ$'$ contains $Swi_2(1) \equiv$

[2] Example 6 is an adaptation of a scenario communicated to the author by M. Thielscher.

[3] We would like to stress that we do not say that the switches domain *should* be formalized so as to keep 'as much persistence as possible' – we formalized it this way just for illustrative purposes, and we feel that one may equally well decide to formalize it along the lines of example 6.

FALSE and $Swi_1(1) \equiv Swi_1(0)$. It is now easy to see that we obtain a model with final state $\{\neg Light, Swi_1, \neg Swi_2\}$. Similarly, we can show there is a final state with $\neg Swi_1$ and Swi_2. But now suppose there is a model $\mathcal{M}'$ with final state containing $\neg Swi_1$ and $\neg Swi_2$. By the minimization of *Do*, we must have $\neg Do(Swi_1(1), \text{FALSE})$ and $\neg Do(Swi_2(1), \text{FALSE})$ in such a model. This means the set EQ$'$ will contain persistence relations for both Swi_1 and Swi_2. Since $Swi_1(0)$ and $Swi_2(0)$ must hold in all models, we must have $\mathcal{M}' \models Swi_1(1) \wedge Swi_2(1)$ too and we have a contradiction. So indeed only one of the switches will change position.

Example 8 [Exclusive Non-Deterministic Effects] The classic example of an action with a non-deterministic effect is the tossing of a coin: if we toss a coin, it will come up either heads (*Heads*) or tails, but we do not know which. However, we do know that whether or not we have *Heads* after tossing is independent of the fact whether or not we had *Heads* before tossing, thus there is no persistence. In light of the previous examples, this can clearly be modeled by a causal theory containing the following CONS:

$$Toss \supset [\ Do(Heads(1), \text{TRUE}) \equiv \neg Do(Heads(1), \text{FALSE})\ ]$$

In this way, in the models for our causal theory in which a *Toss*-event takes place, we either have an intervention that sets the value of *Heads* to TRUE or an intervention that sets its value to FALSE, but not both.

7 Outlook and Conclusion

We have shown that by extending Pearl's causal theories we arrive at a powerful tool for handling ramifications and actions with non-deterministic effects. We had to extend Pearl's theory at two places: by introducing a set of global constraints and by the minimization of *Do*. However, the basic idea behind Pearl's theory, i.e. the sufficient cause principle, remained unchanged. The main ingredient of our causal theories is the *Do*-operator, which allows us to express any propositional combinations of observations (elements of $\mathbf{V}$) and interventions (elements of $\mathcal{A}(\mathbf{V})$). Several existing approaches employ a 'causes'-construct which has a semantics similar to our *Do*, but usually, this construct cannot be *freely* combined with observations (see for example [2, 7, 15, 25]). It is this freedom that makes the representational power of our formalism go beyond most existing ones. We have already shown (but could not include it due to length constraints on this paper) that it can be put to good use in the 'turkey at the trap-door' scenario of Thielscher [25] and to handle 'dependent fluents' [7]; we will show this in the longer version of the paper. There we will also give a detailed comparison of our work to that of McCain and Turner [15, 16, 17] and Thielscher [25].

A remark is in order about other approaches that are based in Pearl's ideas; we mention the approaches of Darwiche and Pearl, Boutilier and Goldszmidt, and Geffner [3, 4, 5, 6]. Though all of these bear some similarities to our approach, none of them extends Pearl's theories in a way that allows free use of *Do* in domain axioms. Therefore they lead to formalisms that are rather different from ours. Nevertheless, we intend to provide detailed comparisons to these works in the near future.

An interesting aspect of our approach is that it may help bridge the conceptual gap between 'causal' and 'non-causal' approaches. Seen in this light, the remark at the end of section 5.2 is probably the most important part of the paper.

References

[1] C. Baral and M. Gelfond. Reasoning about actual and hypothetical occurrances of concurrent and non-deterministic actions. In *Proceedings First Workshop on Nonmonotonic Reasoning, Action and Change*, 1995. Available via ftp://ftp.newcastle.edu.au/pub/papers/nrac95/baral.final.ps.

[2] C. Baral, M. Gelfond, and A. Provetti. Representing actions: Laws, observations and hypotheses. *Journal of Logic Programming*, 12:1–44, 1997.

[3] Craig Boutilier and Moises Goldszmidt. The frame problem and bayesian network action representations. In *Proceedings of the Canadian Conference on Artificial Intelligence (CCAI-96)*, 1996.

[4] A. Darwiche and J. Pearl. Symbolic causal networks. In *Proceedings AAAI-94*, 1994.

[5] A. Darwiche and J. Pearl. Symbolic causal networks for reasoning about actions and plans. In *Working notes: AAAI Spring Symposium on Decision-Theoretic Planning*, 1994.

[6] H. Geffner. Causality, constraints and the indirect effects of actions. In *Proceedings IJCAI '97*, pages 555–560, Nagoya, Japan, 1997.

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

[8] P. D. Grünwald. Causation and nonmonotonic temporal reasoning. In G. Brewka, C. Habel, and B. Nebel, editors, *KI-97: Advances in Artifical Intelligence*, number 1303 in Springer Lecture Notes in Artifical Intelligence, pages 159–170, 1997.

[9] P. D. Grünwald. Nonmonotonic temporal reasoning as a search for explanations. In *Proceedings NRAC '97 (Second IJCAI Workshop on Nonmonotonic Reasoning, Action and Change)*, Nagoya, Japan, 1997.

[10] P.D. Grünwald. Causation, explanation and nonmonotonic temporal reasoning. Technical Report INS-R9701, CWI, jan 1997. Available at ftp://ftp.cwi.nl/pub/pdg/R9701.ps.Z.

[11] J. Gustafsson and P. Doherty. Embracding occlusion in specifying the indirect effects of actions. In *Proceedings of the Fifth International Conference on Principles of Knowledge Representation and Reasoning (KR' 96)*, 1996.

[12] V. Lifschitz. Computing circumscription. In *Proceedings IJCAI-85*, pages 121–127, 1985.

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

[14] F. Lin. Embracing causality in specifying the indeterminate effects of actions. In *Proceedings AAAI-96*, 1996.

[15] N. McCain and H. Turner. A causal theory of ramifications and qualifications. In *Proceedings IJCAI-95*, 1995.

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

[17] N. McCain and H. Turner. On relating causal theories to other formalisms. Unpublished Manuscript, 1997. Available at http://www.cs.utexas.edu/users/hudson/papers.html.

[18] J. McCarthy. Circumscription – a form of non-monotonic reasoning. *Artificial Intelligence*, 13(1,2):27–39,171–172, 1980.

[19] J. Pearl. Graphical models, causality, and intervention. *Statistical Science*, 8(3):266–273, 1993.

[20] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–709, 1995.

[21] J. Pearl. Causation, action, and counterfactuals. Technical Report R-223-U, University of California, Los Angeles, Computer Science Department, 1995. Presented at UNICOM Seminar, London, April 3-5, 1995.

[22] J. Pearl. Causation, action and counterfactuals. In Y. Shoham, editor, *Proceedings TARK-VI*, pages 51–73. Morgan Kaufmann, 1996.

[23] E. Sandewall. Assessments of ramification methods that use static domain constraints. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR '96)*, 1996.

[24] E.J. Sandewall and Y. Shoham. Nonmonotonic temporal reasoning. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 4. Oxford University Press, 1995.

[25] M. Thielscher. Ramification and causality. *Artificial Intelligence*, 89:317–364, 1997.