The Sufficient Cause Principle and Reasoning about Action

P.D. Grünwald

Information Systems (INS)

INS-R9709 November 30, 1997

Report INS-R9709
ISSN 1386-3681

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

The Sufficient Cause Principle and Reasoning about Action

Peter Grünwald
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
pdg@cwi.nl

ABSTRACT
Most 'causal' approaches to reasoning about action have not addressed the basic question of causality directly: what has to be the case in a domain in order for the assertion 'A causes B' to be valid with respect to the domain? 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 comes in both a propositional and a first-order version. The propositional version turns out to be capable of handling many complicated instances of the ramification problem, including domains that contain cyclic causal relationships. It also provides new insights into actions with non-deterministic and/or 'disjunctive' effects. The first-order version additionally handles domains with 'dependent' fluents, incompletely specified action schedules, causal chains of events and 'surprises'. We provide an in-depth comparison between our theory and three other recent approaches: those of McCain & Turner, Baral & Gelfond and Lin. For the former two, we show that they can be reinterpreted as approximations of our extension of Pearl's theory: we prove theorems stating that for large classes of reasoning domains, they permit the same inferences as our theory does, and we give examples of reasoning domains that fall outside these classes, for which our approach clearly works better. For the approach of Lin, we show informally that, for those reasoning domains on which both his and our approach are defined, the two approaches are nearly equivalent. In this way we are able to establish a connection between Lin's approach and Pearl's semantics of causation. Since all concepts in our theory which involve causality are defined in completely non-causal terms, we hope that our work may help bridge the conceptual gap between 'causal' and 'non-causal' approaches to common sense temporal reasoning.

1991 Computing Reviews Classification System: I.2.3,I.2.4
Keywords and Phrases: common sense reasoning, temporal reasoning, causality, nonmonotonic reasoning.
Note: work carried out within the theme INS, under SION-project 612-16-507 'MDL-Neurocomputing'.

1. Introduction

Recently there has been a strong revival of causality-based approaches to common-sense reasoning about action [23, 25, 24, 21, 22, 38, 5]. Causal approaches have proven to be particularly useful in the context of modeling ramifications and actions with non-deterministic effects. Indeed, it has been argued by some [38] 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 has addressed the basic issue of causality directly: what has to be the case in the world in order for the assertion 'A causes B' to be valid? In other words, it is not attempted to define causal relationships in non-causal terms. In stead, causal domain knowledge is formalized explicitly, typically as rules of the form 'A causes B' – and the approaches rely on the assumption that their semantics will yield the right inferences from such assertions.

On the other hand, J. Pearl has provided an empirical semantics for causation that has met with considerable success in statistical applications [29, 30, 31, 32]. However, as will be explained below, Pearl's theory as such cannot be directly applied to common-sense reasoning about action and change.

In this paper, we extend Pearl's theory in a straightforward way so as to overcome this hindrance. In this way we 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 [38] and Lin's [21, 22]. For the causal approaches of McCain and Turner [23, 25] and Baral and Gelfond [3, 5], we show that they can be reinterpreted as approximations of our extension of Pearl's theory: we provide theorems stating that for large classes of reasoning domains, they permit the same inferences as our theory does, and we give examples of reasoning domains that fall outside these classes, for which, we claim, our approach works better.

Structure of this report In section 2 we 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 the ramification problem, as we show in section 5. Section 6 discusses how our propositional causal theories deal with nondeterminism. In section 7 we extend our theory to the first-order case and show how this allows us to handle 'dependent' fluents [11], concurrent events, causal chains of events and surprises. Section 8 then compares our theory to the approaches of McCain & Turner, Lin and Baral & Gelfond. We end with some conclusions. We provide an extensive appendix, which spells out in detail how our causal theory arises as a straightforward extension of Pearl's. Our other appendices contain proofs of the theorems presented throughout the paper.

2. An Informal Introduction to Causal Theories

2.1 Actions Remove Some Dependencies, but keep others Intact!

Causal theories are about domains of variables whose values may be influenced by *interventions*. Each variable X in the domain depends in some deterministic manner on a subset $S(X)$ of all the other variables; here 'deterministic' means that the value of X is completely determined by the values of the variables in $S(X)$. Now 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* with a *set of interventions*, this induces a semantics for the concept of 'action'. This resulting semantics, which will be discussed in more detail below, has sometimes been called *the sufficient cause principle* [8, 7]. 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 $Alive(t)$ 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 $Alive(t+1)$ 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 $Alive(t+1)$ and $Alive(t+2)$ has not disappeared: we can still infer that the turkey will *not* be alive at time $t+2$.

Hence if we have a causal theory for a domain, and we know the interventions that take place in the domain, we end up with a new, updated causal theory. We now discuss two equivalent ways of defining how such a new theory may be arrived at.

Causal Graphs and Structural Equations We can depict the dependencies between the different variables of a given causal theory as a (possibly cyclic) *causal graph* [8, 7]. Figure 1 shows the causal graph for the example above. The semantics of actions can now be characterized in terms of these causal graphs: if an action takes place that sets the variable X to some value, then that variable becomes independent of all its non-descendants in the graph. Hence the intervention changes the theory into a new, updated one that is characterized by the causal graph in which all *incoming* arcs into the

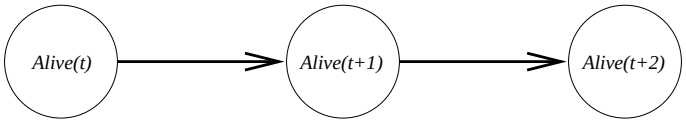

Figure 1: A Very Simple Causal Graph.

node corresponding to X have been removed while all other arcs remain. The graph does not give any information about the exact functional relationships between the variables in the domain, but only about which variables become independent in the presence of an intervention; we can now either regard causal theories as causal graphs together with a description of the functional relations between parents and children in the graph, or, as has been done by Pearl in his more recent papers [31, 32] we can avoid using graphs by writing the functional relations in the form of *structural equations*: these are equations with a directionality attached to them. A set of structural equations for our example would look as follows: $\{Alive(1) = Alive(0)\ ;\ Alive(2) = Alive(1)\ ; \ldots\}$. 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 $X = \ldots$ is replaced by the new equation $X = x$. In our example, if just one intervention takes place, and this intervention sets $Alive(2)$ to FALSE, then the equation $Alive(2) = Alive(1)$ gets replaced by $Alive(2) =$ FALSE, while all other structural equations remain unchanged.

In Pearl's terminology, each structural equation can be seen as a *micro theory* [8]. It describes a single 'mechanism' whose input and output are given by, respectively, the right-hand and the left-hand side of the equation. All these mechanisms are coupled to form a complete theory. But if an intervention takes place, some of the mechanisms get *decoupled* form the rest and the output they normally provide is replaced by the result of the intervention. In other words, in Pearl's theory interventions are defined as *surgeries* on the set of mechanisms: an intervention keeps some of the mechanisms intact, while 'turning off' others, thereby changing the behaviour of the system of mechanisms as a whole.

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 [31, 32] 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. It is therefore not a structural equation. All constraints between variables in Pearl's basic theory must be given in the form of structural equations, so it is impossible to model such a global constraint. In our work we extend Pearlian causal theories in a way that allows us to represent both kinds of domain knowledge mentioned here. 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 straightforward extension of Pearl's causal theories. In appendix 1.1 we show exactly how they are related to Pearl's causal theories as defined in [31, 32].

Preliminaries 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 which are abbreviations of $X \vee \neg X$

and $X \wedge \neg X$ respectively; here X is any variable in $\mathbf{X}$. 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. Until section 7,'$\models$' will stand for standard propositional entailment, later it may also stand for standard first-order entailment. '$\subset$' always denotes *proper* inclusion.

Causal Theories and their Models Here is our initial definition of causal theories:

Definition 3.1 *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 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 equation in EQ expresses 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 3.2 *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 and $\mathcal{M}$ 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 $X_i \equiv b$.*

Note that when we say that $\mathcal{M}$ is a model for T, where T is a causal theory, we mean that $\mathcal{M}$ satisfies the two conditions of definition 3.2. However, when we say $\mathcal{M}$ is a model for Γ or when we write $\mathcal{M} \models \Gamma$, where Γ is a set of propositional formulas, then we mean that the formula Γ is true in $\mathcal{M}$ in the standard sense of propositional logic.

Example 1, continued Let T be as in example 1 above. Condition (1) of definition 3.2 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), $\text{EQ}' = \text{EQ}$ for all models 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 3.2 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 $\text{EQ}' = \{Alive(1) \equiv \text{FALSE}\}$ for all these models and thus all models for T' must have $\neg Alive(1)$.

4. Causal Theories Involving Persistence

In the remainder of this paper we consider domains in which two basic kinds of entities exist: *fluents*, which are properties of the world that, if no interventions take place, do not change over time, and *events*, which will stand for the 'triggers' of interventions. We use the term 'event' rather than 'action' here since our 'events' do not necessarily have to be performed by some agent – the difference however is not crucial in what follows. Causal theories for our domains of interest will thus be defined with respect to a set of fluents $\mathbf{F}$ and a set of events $\mathbf{E}$; throughout this paper we assume these sets to be finite. Specifically, from now on 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 section and the next, we are only interested in the behaviour of our domains directly before and directly after some actions happen. We therefore 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 actions we are interested in happen at the same time, namely directly after the initial point in time. Hence we do not have to index actions by time points and can simply let $\mathbf{V_E} = \mathbf{E}$.

We define a fluent literal to be an expression of the form F or $\neg F$ where F is some fluent in $\mathbf{F}$. The *initial state* of a model $\mathcal{M}$ for a causal theory for sets $\mathbf{E}$ and $\mathbf{F}$ is 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 now extend our turkey domain with a new fluent *Dark*. *Dark*(t) will denote that it is dark at time t; shooting at the turkey will have its intended effect 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

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

There are potentially four different initial and final states for T, corresponding to the four possible interpretations of *Alive* and *Dark*. 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 3.2 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 clearly exactly those in which the values of both *Alive* and *Dark* persist. Thus $\mathbf{M}$ contains four models, each of which has the same initial and final state. Hence in models in which no interventions take place, we have persistence, which is in accord with intuition.

4.1 We need more...

Let us now look at the set of all models $\mathbf{M}'$ for T in which the *Shoot*-event *does* take place. By axiom (1) all $\mathcal{M} \in \mathbf{M}'$ have $\mathcal{M} \models Do(Alive(1), \text{FALSE})$. The set EQ$'$ for these models must therefore contain $Alive(1) \equiv$ FALSE so each $\mathcal{M} \in \mathbf{M}'$ also has $\mathcal{M} \models \neg Alive(1)$. But notice that there is also model $\mathcal{M}' \in \mathbf{M}'$ with

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

since both conditions (1) and (2) of def. 3.2 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 those triggered by the *Shoot*-event ; (2) *Shoot* does not affect *Dark*, so the value of *Dark* should remain unchanged after the *Shoot*-action.

What we have yet forgotten to model are the implicit assumptions behind (1) and (2), namely that (a) *there are no external interventions* and that (b) *actions affect no more things in the world*

than those we explicitly state they affect. Notice that, unlike assumption (a) and the assumption of persistence, assumption (b) is *not* an assumption about the *physics* of our reasoning domains, i.e. it is not an assumption about how the world works. Rather it is an assumption about *what we really mean when we specify our domain knowledge in a certain way.* It is thus what Amsterdam calls a 'narrative convention' [1].

We could avoid having to live with assumption (b) by explicitly stating for each fluent the complete set of actions by which it can be affected. Indeed, something similar is done in several other approaches [7, 8, 35]. However, one of the main problems we are interested in solving is the ramification problem – often, we will deal with actions that can have a very large number of *indirect* effects and it would be very unwieldy to specify all of those explicitly; see section 5.

So let us look for another solution. When exactly does an 'intervention', take place, or equivalently, under what conditions is a variable *set* to a value? Variables are always set to a value 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. 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, and then within each such class pick the model which has $\neg Do(X_i, b)$ for as many X_i and b as possible.

Now what does it mean that two models 'have the same context'? Clearly, the context should contain at least *every fact in the world that can possibly influence whether or not an intervention takes place.* Since we assume (assumption 1) that there are no external events, we may safely say that two models in which exactly the same events take place and the same fluents hold at the same time share the same context. Any two such models interpret all the variables in $\mathbf{V}$ and $\mathbf{U}$ the same; hence we should partition our models into equivalence classes where each class corresponds to one particular interpretation of the variables in $\mathbf{V} \cup \mathbf{U}$ and contains all models with that particular interpretation. For each equivalence class we should then pick the models which have a minimal interpretation (see below) of *Do*. But notice that we must do all this *before* we replace the structural equation set EQ by EQ′ (step 2 of def. 3.2), since this replacement will only work if the *Do*-propositions already have the right interpretations in each model.

We thus have to make precise the notion of 'models that are minimal within a context'. First we need some notation: let $\mathbf{X}$ be any set of propositional variables; let $\mathbf{Y}$ be some subset of $\mathbf{X}$ and let $\mathcal{M}$ be any interpretation of the variables in $\mathbf{X}$. We write $\mathcal{M}\rfloor_{\mathbf{Y}}$ to denote the restriction of $\mathcal{M}$ to $\mathbf{Y}$, i.e. the set of assignments that $\mathcal{M}$ attaches to the variables in $\mathbf{Y}$.

Now let again $\mathbf{X}$ be any set of propositional variables; let $\mathbf{Y}$ and $\mathbf{Z}$ be subsets of $\mathbf{X}$ with $\mathbf{Y} \cap \mathbf{Z} = \emptyset$. Let $\mathcal{M}$ and $\mathcal{C}$ be two interpretations of the variables in $\mathbf{X}$ and let Γ be a set of formulas over $\mathbf{X}$.

Definition 4.1 *We call $\mathcal{M}$ a minimal model for Γ of the variables $\mathbf{Y}$ within context $\mathcal{C}\rfloor_{\mathbf{Z}}$ iff:*

1. $\mathcal{M} \models \Gamma$ *and* $\mathcal{M}\rfloor_{\mathbf{Z}} = \mathcal{C}\rfloor_{\mathbf{Z}}$.
2. *there is no* $\mathcal{M}'$ *with* $\mathcal{M}' \models \Gamma$, $\mathcal{M}'\rfloor_{\mathbf{Z}} = \mathcal{C}\rfloor_{\mathbf{Z}}$ *and*

$$\{Y \in \mathbf{Y} \mid \mathcal{M}' \models Y\} \subset \{Y \in \mathbf{Y} \mid \mathcal{M} \models Y\}$$

where '$\subset$' denotes proper inclusion.

We will thus be looking for the models $\mathcal{M}$ for CONS with a minimal interpretation of $\mathcal{A}(\mathbf{V})$ in context $\mathcal{M}\rfloor_{\mathbf{V}\cup\mathbf{U}}$. It is important to realize that the minimization can never completely rule out any interpretations of $\mathbf{V}$ and $\mathbf{U}$: if $\mathcal{M} \models$ CONS, then there *must* exist some minimal model $\mathcal{M}'$ for CONS of $\mathcal{A}(V)$ with the same context, i.e. with $\mathcal{M}'\rfloor_{\mathbf{V}\cup\mathbf{U}} = \mathcal{M}\rfloor_{\mathbf{V}\cup\mathbf{U}}$

We call causal theories in which *Do* still has to be minimized *partially specified* (since it is only specified what fluents are affected by events). In the remainder of this paper, we will assume that

all our causal theories are actually partially specified ones. This brings us to the final definition of 2-point causal theories and their models:

Definition 4.2 *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 3.2 in its first condition.

Definition 4.3 *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. $\mathcal{M}$ *is a minimal model for* CONS *of the variables* $\mathcal{A}(\mathbf{V})$ *within context* $\mathcal{M}\rfloor_{\mathbf{V}\cup\mathbf{U}}$.
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 *and* $\mathcal{M}$ *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* $X_i \equiv \Phi_i(\dots)$ *and we* add *the equation* $\Phi_i \equiv b$.

If $\mathcal{M}$ *is a model for a 2-point causal theory* T, *we write* $\mathcal{M} \models_{\mathbf{c}} T$.

We are now ready to harvest.

5. Ramifications

The *ramification problem* [34] has been given a lot of attention recently [21, 15, 23, 25, 38]. It occurs in domains where an action or event has only a small number of *direct* effects while it can have a very large number of *indirect* effects. For example, if a turkey is shot at, it will stop being alive, but it will also stop walking, breathing, eating, pecking etc. Theories in which all such indirect effects of an action are stated explicitly would quickly become unmanageably large. The ramification problem can be defined as the problem of how to correctly represent the indirect effects of actions in a concise manner. This cannot be simply done by adding constraints between fluents to our theory; for example, suppose the turkey from our previous examples may either be walking or not, denoted by the fluent *Walking*. As pointed out by several authors [23], if the turkey is alive and walking when being shot at, it will stop walking: $\neg$*Walking* is a *ramification* of the effect of *Shoot*. But if the turkey is not alive and somebody tries to make it walking (for example, by performing the action *Entice*), it will not suddenly become alive! *Alive* is *not* a ramification of *Entice*; rather, $\neg$*Alive* is a *qualification* of *Entice*. Clearly, a sentence of the form $\neg Alive(t) \supset \neg Walking(t)$ is not enough to express this knowledge, since it treats $Alive(t)$ and $\neg Walking(t)$ in an equivalent manner and therefore cannot represent the different 'dynamics' of *Alive* and $\neg$*Walking*.

Causal theories can handle domains like the above by making use of *Do* in domain constraints. Using *Do*, we are able to express the difference between an observation (e.g. $\neg Alive(1)$) and an intervention (e.g. $Do(Alive(1), \text{FALSE})$). In the remainder of this section we will see that this is exactly what we need to express complicated domain constraints; specifically, we will often use axioms of the form $Do(F_1, b) \supset Do(F_2, b')$; such an axiom can be interpreted as saying that *any event in the domain that sets the value of* F_1 *to* b *also sets the value of* F_2 *to* b'.

In the remainder of this section we give a first demonstration of the utility of 2-point causal theories by showing how they correctly handle several types of ramification constraints that are known to pose problems for many existing approaches; the only other approach we are aware of that handles all of them correctly is Thielscher's [38]. In Thielscher's paper one also finds an excellent analysis of why these examples cause so many difficulties.

Example 3 [The Walking Turkey] Let us formalize the example above. Consider the 2-point causal theory T_{WT} for the event and fluent sets $\mathbf{E} = \{Shoot, Entice\}$ and $\mathbf{F} = \{Alive, Walking\}$. $\mathbf{U} = \emptyset$. EQ $= \{Alive(1) \equiv Alive(0)\ ;\ Walking(1) \equiv Walking(0)\}$. CONS consists of the following four axioms:

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

$$Entice \supset Do(Walking(1), \text{TRUE}) \tag{2}$$

$$\neg Alive(t) \supset \neg Walking(t) \tag{3}$$

$$Do(Alive(t), \text{FALSE}) \supset Do(Walking(t), \text{FALSE}) \tag{4}$$

Here axioms of the form $\phi(t)$ should be read as $\phi(0) \wedge \phi(1)$. We see that axiom (3) denotes a *static* domain constraint ('there can be no state in which a turkey is both walking and not alive') while axiom (4) denotes its corresponding 'dynamics'.

Let us denote by $\mathbf{M}$ the set of models for T_{WT} in which *Shoot* takes place while *Entice* does not; by axioms (1) and (4) we have $Do(Alive(1), \text{FALSE})$ and $Do(Walking(1), \text{FALSE})$ in all such models. By condition (1) of definition 4.3, we also have $\neg Do(X_i, b)$ for all other (X_i, b). Hence, the updated set of structural equations EQ$'$ for the models in $\mathbf{M}$ (definition 4.2) becomes:

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

Hence all models in $\mathbf{M}$ have as their final state $\{\neg Alive, \neg Walking\}$. We also see (by checking whether all the axioms in CONS and EQ$'$ hold) 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 definition 4.3 they also have $\neg Do(Alive(1), \text{TRUE})$. Hence the set EQ$'$ in condition (2) of that definition becomes

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

By axiom (3) all models $\mathcal{M}$ must then also have $\mathcal{M} \models Alive(1)$, and, since $\mathcal{M} \models$ EQ$'$ also $\mathcal{M} \models Alive(0)$. This means that there are no models for T_{WT} with both $\neg Alive(0)$ and *Entice*: *Entice* has *Alive* as an *implicit qualification*.

Example 4 [the Clever Turkey] The following example was introduced by Thielscher [38]. It is interesting for two reasons: first, as argued extensively by Thielscher [38], it cannot be handled by McCain & Turner's theory of ramifications [23]. Secondly, the example shows that it may be useful to include axioms of the form

$$X \supset \neg Do(Y, b)$$

in CONS to express qualifications on interventions. Axioms corresponding to this form do not exist in Thielscher's [38] formalism, and indeed, his formalization of the domain looks rather different from ours.

Imagine we want to hunt a turkey using a trap-door rather than a gun. The turkey can be either at the trap or not (denoted by fluent *At_Trap*) and the trap can be either open or closed (denoted by *Trap_Open*). If the turkey for some reason or another finds itself in an open trap, it dies. If we *Entice* the turkey, it walks to the trap. However, the turkey is clever enough to never walk into an open trap.

We formalize this scenario in the 2-point causal theory T_{CT} for sets $\mathbf{E} = \{Entice\}, \mathbf{F} = \{At\_Trap, Alive, Trap\_Open\}$; $U = \emptyset$; EQ is as required by definition 4.2 and CONS consists of:

$$Entice \supset Do(At\_Trap(1), \text{TRUE}) \tag{5}$$

$$[\ At\_Trap(t) \wedge Do(Trap\_Open(t), \text{TRUE})\ ] \supset Do(Alive(t), \text{FALSE}) \tag{6}$$

$$[\ At\_Trap(t) \wedge Trap\_Open(t)\ ] \supset \neg Alive(t) \tag{7}$$

$$Trap\_Open(0) \supset \neg Do(At\_Trap(1), \text{TRUE}) \tag{8}$$

Consider the initial state

$$\{\neg \mathit{At\_Trap}, \mathit{Alive}, \mathit{Trap\_Open}\}$$

So the trap is open and the turkey is not at the trap. Suppose we want to entice the turkey; hence we add the axiom *Entice* to CONS; let the resulting theory be $T_{\mathrm{CT},2}$. By axiom (5) the turkey will move while by axiom (8) the turkey will not move: there are no models for $T_{\mathrm{CT},2}$ with the above initial state. But now look at the initial state

$$\{\neg \mathit{At\_Trap}, \mathit{Alive}, \neg \mathit{Trap\_Open}\} \tag{9}$$

and we want to entice the turkey and then open the trap. We thus add the following axiom to CONS and let the resulting theory be $T_{\mathrm{CT},3}$:

$$\mathit{Entice} \wedge \mathit{Do}(\mathit{Trap\_Open}(1), \mathrm{TRUE}) \tag{10}$$

Notice that in initial state (9) there is nothing which contradicts $T_{\mathrm{CT},3}$; by axiom (5) the set EQ′ in definition 4.3 will contain: $\mathit{At\_Trap}(1) \equiv \mathrm{TRUE}$. This means that in all models for $T_{\mathrm{CT},3}$ we have $\mathit{At\_Trap}(1)$; since we also have $\mathit{Do}(\mathit{Trap\_Open}(1),\mathrm{TRUE})$ in all these models, axiom (6) applies, and we further have $\mathit{Do}(\mathit{Alive}(1), \mathrm{FALSE})$. Hence EQ′ further contains $\mathit{Alive}(1) \equiv \mathrm{FALSE}$, and any model for $T_{\mathrm{CT},3}$ will have final state $\{\mathit{At\_Trap}, \neg \mathit{Alive}, \mathit{Trap\_Open}\}$. It is easy to check that there indeed exists an interpretation with such a final state that is a model for $T_{\mathrm{CT},3}$.

Example 5 [The Suitcase & Switches Problem] The following example is due to Lifschitz' [19]. It is equivalent to Lin's [21] 'suitcase problem'. 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. The main importance of this example is that some previous approaches cannot express the domain properly: they cannot rule out models in which, as a result of turning on one switch in a context in which the other one is on already, the other switch may jump into the *off*-position while light remains out; see Lin [21] for details. We formalize the domain using event and fluent sets $\mathbf{E} = \emptyset$ and $\mathbf{F} = \{\mathit{Swi}_1, \mathit{Swi}_2, \mathit{Light}\}$. 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:

$$[\ \mathit{Swi}_1(t) \wedge \mathit{Do}(\mathit{Swi}_2(t), \mathrm{TRUE})\ ] \supset \mathit{Do}(\mathit{Light}(t), \mathrm{TRUE}) \tag{11}$$

$$[\ \mathit{Swi}_2(t) \wedge \mathit{Do}(\mathit{Swi}_1(t), \mathrm{TRUE})\ ] \supset \mathit{Do}(\mathit{Light}(t), \mathrm{TRUE}) \tag{12}$$

$$\mathit{Do}(\mathit{Swi}_1(t), \mathrm{FALSE}) \supset \mathit{Do}(\mathit{Light}(t), \mathrm{FALSE}) \tag{13}$$

$$\mathit{Do}(\mathit{Swi}_2(t), \mathrm{FALSE}) \supset \mathit{Do}(\mathit{Light}(t), \mathrm{FALSE}) \tag{14}$$

$$(\mathit{Swi}_1(t) \wedge \mathit{Swi}_2(t)) \equiv \mathit{Light}(t) \tag{15}$$

Here axiom (15) 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 $\neg \mathit{Swi}_2(0) \wedge \mathit{Do}(\mathit{Swi}_1(1), \mathrm{TRUE})$ to CONS. By the minimization of *Do*, we get $\neg \mathit{Do}(\mathit{Swi}_2(1), b)$ for all b in all models. Hence EQ′ contains $\mathit{Swi}_2(1) \equiv \mathit{Swi}_2(0)$, so all models for T_{SW} must have $\neg \mathit{Swi}_2(1)$. But in all models with $\neg \mathit{Swi}_2(1)$, we already have (again by the minimization of *Do*) $\neg \mathit{Do}(\mathit{Light}(1), \mathrm{TRUE})$ and $\neg \mathit{Light}(1)$. It follows that all models will have $\neg \mathit{Light}$ 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 First Glimpse at other approaches

We have seen in the introduction that, in Pearl's terminology, the proposition $\mathit{Do}(X_i, b)$ 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 until section 8 and give only a suggestive example for the time being. In Thielscher's work, the equivalent of axiom (4) looks as follows ([38], page 330):

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

In Lin's work [21, 22], we would get:

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

In McCain & Turner's approach [23, 24, 25], it would be formalized as:

$$\neg \mathit{Alive} \Rightarrow \neg \mathit{Walking} \tag{16}$$

where '$\Rightarrow$' is to be read as 'if $\neg \mathit{Alive}$, then the fact that $\neg \mathit{Walking}$ is caused' [24]. McCain & Turner's semantics interprets (16) as something like

$$\neg \mathit{Alive}(t) \supset \mathit{Do}(\mathit{Walking}(t), \text{FALSE}) \tag{17}$$

In the turkey example, replacing (4) by (17) 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, $(X \equiv b_1) \supset \mathit{Do}(Y, b_2)$ is *not* equivalent to $\mathit{Do}(X, b_1) \supset \mathit{Do}(Y, b_2)$.

5.2 Causal Cycles

We have just seen that $\mathit{Do}(X, \text{TRUE}) \supset \mathit{Do}(Y, \text{TRUE})$ may often be read as 'X causes Y'. As pointed out by Sandewall and Gustafsson & Doherty [15, 33], 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 6 Let $T_{\text{SW},2}$ be as T_{SW} but with the following axioms added to CONS:

$$\mathit{Do}(\mathit{Light}(t), \text{FALSE}) \supset [\ \mathit{Swi}_1(t) \supset \mathit{Do}(\mathit{Swi}_2(t), \text{FALSE})\ ] \tag{18}$$

$$\mathit{Do}(\mathit{Light}(t), \text{FALSE}) \supset [\ \mathit{Swi}_2(t) \supset \mathit{Do}(\mathit{Swi}_1(t), \text{FALSE})\ ] \tag{19}$$

$$\mathit{Do}(\mathit{Light}(t), \text{TRUE}) \supset \mathit{Do}(\mathit{Swi}_1(t), \text{TRUE}) \tag{20}$$

$$\mathit{Do}(\mathit{Light}(t), \text{TRUE}) \supset \mathit{Do}(\mathit{Swi}_2(t), \text{TRUE}) \tag{21}$$

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 $\mathcal{M} \models \text{CONS}$, there is an interpretation $\mathcal{M}'$ that interprets all variables in $\mathbf{V} \cup \mathbf{U}$ the same way but which has $\mathcal{M}' \models \neg \mathit{Do}(X_i, b)$ for all X_i and b. So the minimization of *Do* in the definition of causal models will make sure that all models have $\neg \mathit{Do}(X_i, b)$ for all X_i and b. Hence $\text{EQ}' = \text{EQ}$ for all models 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:

$$\mathit{Do}(\mathit{Light}(1), \text{TRUE}) \wedge \neg \mathit{Swi}_1(0) \wedge \neg \mathit{Swi}_2(0) \tag{22}$$

Axioms (20) and (21) make sure that we have $\mathit{Do}(F(1), \text{TRUE})$ for $F \in \{\mathit{Swi}_1, \mathit{Swi}_2, \mathit{Light}\}$ in all models; so EQ' becomes $F(1) \equiv \text{TRUE}$ for all these F in all models, and all models get final state $\{\mathit{Swi}_1, \mathit{Swi}_2, \mathit{Light}\}$; it is easy to see that models with such a final state indeed exist, so we get exactly the result we wanted. What happens if we turn off the light will be discussed in example 8.

The example above leads to a problem for some of the earlier causal approaches (i.e. those of Gustafsson & Doherty and McCain & Turner; see [15] 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 (11) and (12) would have been replaced by the single axiom

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

while axioms (20) and (21) would have become:

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

Axioms (18) and (19) would have changed similarly. But this would allow us to conclude that each model with final state $S = \{Light\ ;\ Swi_1\ ;\ Swi_2\}$ 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 for models with final state S the set EQ$'$ in condition (2) of definition 4.3 would become $F \equiv \text{TRUE}$ for all these F and all three persistence relations would be 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 (11)-(15) and (18)-(21)) this cannot happen.

Thielscher ([38], 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' [15].

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 6 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 [25, 38]. A special case of this question concerns the more well-known issue of representing actions with non-deterministic effects [3, 25, 34]. 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 6 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. 3.1). We illustrate their use by considering a domain[1] which clearly asks for eliminating all structural equations involved in the disjunction.

Example 7 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:

$$\neg Touches\_table(0) \wedge \neg Touches\_paper(0) \tag{1}$$

$$\begin{aligned} Drop \supset [\ & (\ Do(Touches\_table(1), \text{TRUE}) \wedge U_1\ ) \vee \\ & (\ Do(Touches\_paper(1), \text{TRUE}) \wedge U_2\ )\ ] \end{aligned} \tag{2}$$

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 (2). By the minimization of Do in definition 4.3 and by axiom (2), we get the following interpretation for each class of models:

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

(3)

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 8 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[2] 6. Consider a scenario in the context of that example 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 (18) and $\neg Do(Swi_1(1), \text{FALSE})$ by the minimization of Do. For these models EQ$'$ contains $Swi_2(1) \equiv \text{FALSE}$ 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.

[1] Example 7 is an adaptation of a scenario communicated to the author by M. Thielscher.

[2] 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 7.

Example 9 [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. Handling many time-points, Events and Surprises

In this section we extend our 2-point causal theories to handle arbitrarily many points in time. This means we will have to move to a first-order language. In the subsections to come, we first introduce 'basic' first-order causal theories. We then instantiate these theories in more and more complicated ways: in section 7.2, we start with simple instantiations for handling domains in which exactly one event happens at a time. Section 7.3 extends our theories to handle 'dependent fluents' which form yet another instance of the ramification problem. Then (section 7.4) we move over to domains in which more than one event may happen at the same time and we show how this allows us to formalize causal chains of events. Finally, we extend our domains to handle complete surprises (section 7.6).

7.1 First-Order Causal Theories

We want to extend propositional causal theories to handle countably many points in time. It seems that we would get an infinite number of structural equations '$F_i(t+1) = F_i(t)$', one for each fluent $F_i \in \mathbf{F}$ and one for each t. This suggests using predicate logic and universally quantifying our structural equations over time points. However, in that case Pearl's semantics gets undefined: it is not immediately clear how to perform a replacement of structural equations if the structural equations are quantified over. But it turns out that, by changing our formalism slightly, we can *mimic* the replacement of equations *within* causal theories. Once we have done this, the generalization to the first order case is completely straightforward. The general idea of this mimicking operation is as follows: any equation in EQ of the form $X_i \equiv \Phi$ will be replaced by three new axioms:

$$\begin{aligned} [\ \neg Do(X_i, \text{TRUE}) \wedge \neg Do(X_i, \text{FALSE})\ ] \ &\supset (X_i \equiv \Phi) \\ Do(X_i, \text{TRUE}) &\supset X_i \\ Do(X_i, \text{FALSE}) &\supset \neg X_i \end{aligned}$$

In appendix 1.2 we show that the new axioms embody exactly the same semantics as the replacement of structural equations that was used in definition 4.3. We also show there how this enables us to quantify over structural equations.

In the definitions below, the mimicking trick has been applied to the special case of causal theories where the observable variables can be of two different kinds: *events* and *fluents* - just the type of entities we have encountered before.

Preliminaries We use a many-sorted first-order language $\mathcal{L}$. A structure $\mathcal{M}$ for $\mathcal{L}$ consists of universes for all the sorts and interpretations for all function and predicate constants in $\mathcal{L}$. For a sort X indicated by the letter X, we write $|\mathcal{M}|_x$ to denote the universe of the sort X. For a function or predicate constant K, we write $\mathcal{M}[\![K]\!]$ to denote the interpretation of K in $\mathcal{M}$ (which is then a function or a set, respectively). A model of a set of sentences Γ is any structure $\mathcal{M}$ such that Γ is true in $\mathcal{M}$, where truth with respect to sentences is defined as usual. If $\mathcal{M}$ is a model of Γ, we write $\mathcal{M} \models \Gamma$. So from now on '$\models$' stands for first-order rather than propositional entailment.

The language $\mathcal{L}$ in turn depends on the sets $\mathbf{E}$ and $\mathbf{F}$. We therefore sometimes write $\mathcal{L}(\mathbf{E},\mathbf{F})$. $\mathcal{L}(\mathbf{E},\mathbf{F})$ contains three sorts: *Booleans* (variables of the sort will be denoted by b); *time points* (t) and *observables* (x). There are two 'subsorts' to observables: *fluents* (variables of the sort denoted by f) and *events* (e). We explain what we mean by 'subsort' below. The set $\mathbf{B}$ of Boolean constants contains two elements: $\mathbf{B} = \{\text{TRUE}, \text{FALSE}\}$. The set of time-point constants is $\mathtt{N}_0 = \{0, 1, 2, \dots\}$, i.e. the set of nonnegative integers. The set of fluent constants coincides with $\mathbf{F}$; the set of event constants is identified with $\mathbf{E}$. $\mathcal{L}(\mathbf{E},\mathbf{F})$ contains at least the following functions and predicates:

- *Ho*: $Ho(x,t)$ will denote that observable x *Ho*lds at time t.
- *Do*: $Do(x,b,t)$ will denote that the value of observable x at time t has been set to value b by some (unspecified) action.
- '$=$,$<$,$+$' which will receive their usual interpretation.

By a 'subsort' we mean the following: events and fluents are really two different sorts. Whenever a predicate is defined for the sort of observables, it is really defined both for the sort events and the sort fluents. Whenever in a formula we quantify over x, for example, we have the axiom $\forall x \phi(x)$, we implicitly quantify over elements of both constituent sorts (i.e. the axiom should be read as $\forall e \phi(e) \wedge \forall f \phi(f)$). In what follows, we implicitly assume all those formulas that are listed without quantifiers to be universally quantified.

General First-Order Causal Theories We are now ready to define first-order causal theories. In appendix 1.2, we show formally that the definitions below are indeed a straightforward extension of the propositional causal theories defined earlier.

Definition 7.1 *A first-order causal theory for a language $\mathcal{L}(\mathbf{E},\mathbf{F})$ is a tuple $\langle \text{EQ}, \text{CONS} \rangle$ where*

1. EQ *is a set of sentences for $\mathcal{L}(\mathbf{E},\mathbf{F})$ containing the 'intervention axioms'*

$$\forall x,t. \qquad Do(x, \text{TRUE}, t) \supset Ho(x,t) \tag{1}$$
$$\forall x,t. \qquad Do(x, \text{FALSE}, t) \supset \neg Ho(x,t) \tag{2}$$

and, in addition, one or more axioms of the form

$$\forall\ x,t\ .\ [\ \Psi(x,t) \wedge \neg Do(x,\text{TRUE},t) \quad \wedge \quad \neg Do(x,\text{FALSE},t)\ ]\ \supset\ [\ Ho(x,t) \equiv\ \Phi(x,t)\ ] \tag{3}$$

Here x is a variable of sort events *or* fluents*; t is of sort time-points. We call the expression '$Ho(x,t) \equiv \Phi(x,t)$' a 'structural equation'. We call $\Psi(x,t)$ a 'precondition' for this structural equation.*

2. CONS *is a set of sentences for $\mathcal{L}(\mathbf{E},\mathbf{F})$ containing at least uniqueness-of-names (UNA) and domain closure (DC) axioms*[3] *for the sets $\mathbf{B}, \mathbf{E}$ and $\mathbf{F}$.*

We will refer to the uniqueness-of-names axioms in CONS as the 'UNA-axioms' and to the domain closure axioms as the 'DC'-axioms.

We now give the definition of models for causal theories. Remember that in the propositional case, we were looking for the minimal interpretations of the set of *Do*-variables within each context, i.e. each interpretation of all other propositional variables of the theory. We will now do exactly the same thing in a first-order setting: now, we want the minimal interpretations of the *Do-predicate* for each context. Now, a context is any interpretation of all other *predicates* of the theory.

[3] A uniqueness-of-names (UNA) axiom for a finite set of constants $\mathbf{X} = \{X_1, \dots, X_n\}$ is the formula $\bigwedge_{i \neq j} X_i \neq X_j$. A domain closure axiom for the set $\mathbf{X}$ is the axiom $\forall x\ x = X_1 \vee \dots \vee x = X_n$.

This 'minimization within a context' can be conveniently implemented using circumscription [18, 26]. For details about circumscription, we refer to [18]. In appendix 2 we give Lifschitz' characterization of circumscription in model-theoretic terms. From that definition, it can be seen immediately that circumscribing *Do* in CONS with all other functions and predicates *fixed* will give us exactly the models we want: if a predicate is kept fixed while circumscribing *Do*, each interpretation of the predicate will serve as a context within which *Do* is minimized. Also, just as in the propositional case, we have to minimize *Do* in CONS *before* we add the set of structural equations EQ to it – cf. the remark in section 4.1. This is reflected in the following definition. The expression $Circum(\text{CONS}; Do)$ stands for the *circumscription of Do in* CONS *with all other functions and predicates kept fixed.*

Definition 7.2 *a structure* $\mathcal{M}$ *for the language* $\mathcal{L}$ *is a* model *for the first-order causal theory* T *(written as* $\mathcal{M} \models_{\mathbf{c}} T$*) iff*

1. $\mathcal{M} \models \text{EQ} \wedge Circum(\text{CONS}; Do)$.
2. *Time-points in* $\mathcal{L}$ *are interpreted as the integers; '+' and '<' are interpreted accordingly.*

Note that for any causal theory $T = \langle \text{EQ}, \text{CONS} \rangle$ and any $\mathcal{M}$ we have that if $\mathcal{M} \models_{\mathbf{c}} T$, then also $\mathcal{M} \models \text{CONS}$. Since by definition 7.1 above CONS contains both UNA- and DC-axioms for **B**, **E** and **F**, we easily see that the following proposition holds:

Proposition 7.1 *we may assume without loss of generality that for any model* $\mathcal{M}$ *for a causal theory* T*, we have*

1. $|\mathcal{M}|_b = \mathbf{B}$ *and* $|\mathcal{M}|_e = \mathbf{E}$ *and* $|\mathcal{M}|_f = \mathbf{F}$.
2. $\mathcal{M}$ *interprets all elements in* **B**, **E** *and* **F** *as themselves.*

We now turn to the exact kind of rules of form (3) that we will need if we want to model persistence. Fluents are supposed to behave as before: if no intervention takes place, they persist. This is modeled by the following axiom in EQ which we call our *persistence axiom*:

$$\forall f, t\ .(t > 0) \supset$$
$$[\ \neg Do(f, \text{TRUE}, t) \ \wedge \ \neg Do(f, \text{FALSE}, t)\ ] \ \supset \ [\ Ho(f, t) \equiv \ Ho(f, t-1)\ ] \tag{4}$$

Note that this axiom is indeed of the form prescribed by the definition of first-order causal theories above. Concerning events, we can opt for either one of two possibilities, The first is reminiscent of the way actions are treated in the standard situation calculus [27]: between each two 'time points', exactly one action happens. We can formalize this by adding an extra axiom to CONS ('$\exists!$' stands for 'there exists exactly one'):

$$\forall t \exists! e.\ Ho(e, t) \tag{5}$$

The formula $Ho(e, t)$ is to be interpreted as saying that event e takes place between time t and time $t + 1$. The second, more complicated possibility is to allow for multiple actions happening at the same time; this is similar to what happens in the works of Morgenstern & Stein and Baral, Gelfond & Provetti [5, 36]. We first discuss the former possibility, delaying treatment of the latter until section 7.4.

7.2 Handling Events like in the Situation Calculus

We first extend our definition of causal theories to incorporate axioms (4) and (5).

Definition 7.3 *A* first-order causal theory with persistence *for the tuple* $\langle \mathbf{E}, \mathbf{F} \rangle$ *is a tuple* $\langle \text{EQ}, \text{CONS} \rangle$ *where* EQ *and* CONS *are sets of sentences for the language* $\mathcal{L}(\mathbf{E}, \mathbf{F})$*,* EQ *consists of intervention axioms (1),(2) and the persistence axiom (4) while* CONS *contains at least UNA- and DC-axioms for* $\mathbf{B}, \mathbf{E}$ *and* **F** *and axiom (5).*

Example 10 [Yale Shooting Problem] Armed with this definition, we are able to handle most standard reasoning domains involving more than two time points. As an example, consider the original Yale Shooting domain [16]. In this domain there is a turkey that can be alive or not; there is a gun that can be loaded or not; there is an event 'load' which loads the gun, an event 'wait' which has no effects at all and an event 'shoot', which, if performed when the gun is loaded, causes the turkey not to be alive anymore. We further suppose that the turkey is alive and that the gun is unloaded at time $t = 0$, and that, starting at time $t = 0$, someone first loads the gun, then waits and then shoots. To model this scenario, let T_{YSP} be a causal theory with persistence for sets $\mathbf{E} = \{Load, Wait, Shoot\}$, $\mathbf{F} = \{Alive, Loaded\}$. Apart from the axioms mentioned above, the set CONS further contains the following axioms:

$$Ho(Load, t) \supset Do(Loaded, \mathrm{TRUE}, t+1) \tag{6}$$

$$Ho(Loaded, t) \wedge Ho(Shoot, t) \supset Do(Alive, \mathrm{FALSE}, t+1) \tag{7}$$

$$Ho(Alive, 0) \wedge \neg Ho(Loaded, 0) \tag{8}$$

$$Ho(Load, 0) \wedge Ho(Wait, 1) \wedge Ho(Shoot, 2) \tag{9}$$

The original problem with the Yale Shooting Domain was that naive approaches to modeling persistence end up with two classes of models: an intended one, in which the turkey is not alive anymore at time $t = 3$, and an unintended one, in which the gun becomes unloaded again during the 'waiting' and hence the turkey remains alive when shot at. We will show that we only get the intended class. To see this, notice first that by axiom (5) we have that the only events taking place in *any* model at times $0, 1$ and 2 are those introduced in axiom (9). The circumscription of Do in the definition of causal models then makes sure that in all models of T_{YSP} we have $\neg Do(Loaded, b, t)$ for all b and $t \in \{0, 2\}$. On the other hand, by axioms (9) and (6) we have $Do(Loaded, \mathrm{TRUE}, 1)$ in all models, too. It follows by axiom (1) that we have $Ho(Loaded, 1)$ in all models and by the persistence axiom (4) that we have $Ho(Loaded, 2)$ in all models. Since by axiom (9) we have $Ho(Shoot, 2)$ in all models, the antecedent of axiom (7) holds in all models for t instantiated to 2 and we have $Do(Alive, \mathrm{FALSE}, 3)$ in all models. By axiom (4) we then have $\neg Ho(Alive, 3)$ in all models for T_{YSP}. It is easy to show that such models indeed exist.

We state without proof that we also handle the related 'Stanford Murder Mystery' [2]. By slightly extending our language, we can now also deal with yet another kind of ramifications - those involving so called *dependent fluents*. They will be introduced in the next section:

7.3 Ramifications Again: Dependent Fluents

Giunchiglia and Lifschitz [11] argue that we sometimes want to express dependencies between fluents where these dependencies may be partially unknown. Let us consider Giunchiglia and Lifschitz' [11] motivating example: suppose we are in a room with a baby, an object and a table. The object may be dangerous to the baby (for example, it might be a hammer) but we are not sure about that. We do know however that if the object is placed on the table, it is out of reach of the baby, and hence it is safe. In other words, the safeness of an object depends on whether or not it is on the table, but we do not know exactly in what way: we know it is safe if it is on the table; if it is not on the table, we just know that this fact *determines* whether it is safe or not (see [11] for details). This implies that if we put on object on the table and then remove it again, we want there to be only two possibilities: either the object was safe both before and after it lay on the table or it was unsafe both before and after it lay on the table. Giunchiglia and Lifschitz introduce the high-level action language $\mathcal{ARD}$ in order to deal with this kind of dependencies. It turns out that causal theories can be instantiated for 'dependent fluents' in a straightforward manner, the reason being that we are allowed the use of *assumption symbols* (unobserved variables) in causal theories.

We first extend our formalism by introducing the new subsort of *dependent fluents*; we will assume there are a finite number of them, listed in the set $\mathbf{D}$. Variables of the subsort will be indicated

by d. Like regular fluents, dependent fluents may or may not persist. But unlike the situation for regular fluents, the behaviour of dependent fluents is fully defined in terms of other (regular) fluents and assumption symbols. We first extend our definitions of causal theories to deal with the subsort of persistence. The language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F})$ is as the language $\mathcal{L}(\mathbf{E}, \mathbf{F})$ but now with dependent fluents as a new subsort of observables and with $\mathbf{D}$ indicating the constants of this new subsort.

Definition 7.4 *A* first-order causal theory with persistence and dependent fluents *for the tuple* $\langle \mathbf{D}, \mathbf{E}, \mathbf{F} \rangle$ *is a tuple* $\langle \text{EQ}, \text{CONS} \rangle$ *where* EQ *and* CONS *are sets of sentences for the language* $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F})$, EQ *consists of intervention axioms (1),(2) and the persistence axiom (4) while* CONS *contains at least UNA- and DC-axioms for* $\mathbf{B}, \mathbf{D}, \mathbf{E}$ *and* $\mathbf{F}$ *and axiom (5).*

Note that the only difference between causal theories as defined here and those defined as in the previous section (definition 7.3) is the addition of UNA- and DC-axioms for dependent fluents. The new definition allows us to formalize the example described above:

Example 11 Let T_{DF} be a first-order causal theory with persistence for the language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F})$ containing the additional 0-ary predicate U_{Safe}. Here $\mathbf{D} = \{Safe\}$, $\mathbf{E} = \{Put\_On\_Table, Remove\_From\_Table\}$ and $\mathbf{F} = \{On\_Table\}$ and CONS contains the following additional axioms:

$$Ho(Put\_On\_Table, t) \supset Do(On\_Table, \text{TRUE}, t+1) \tag{10}$$

$$Ho(Remove\_From\_Table, t) \supset Do(On\_Table, \text{FALSE}, t+1) \tag{11}$$

$$Ho(Safe, t) \equiv [\, Ho(On\_Table, t) \vee (\neg Ho(On\_Table, t) \wedge U_{Safe}) \,] \tag{12}$$

U_{Safe} is called an *assumption symbol*. It determines the actual (but unknown) relationship between $\neg On\_Table$ and $Safe$. Suppose further that CONS contains the following observations:

$$\neg Ho(On\_Table, 0) \wedge Ho(Put\_On\_Table, 0) \wedge Ho(Remove\_From\_Table, 1) \tag{13}$$

Notice that (13) together with (10) and (11) make sure that we have $Do(On\_Table, \text{TRUE}, 1)$ and $Do(On\_Table, \text{FALSE}, 2)$ in all models. The circumscription of Do in the definition of models for causal theories makes sure that we have $\neg Do(x, b, t)$ for all other x, b and $t \in \{0, 1\}$. It follows that all models for T_{DF} have

$$\neg Ho(On\_Table, 0) \wedge Ho(On\_Table, 1) \wedge \neg Ho(On\_Table, 2) \tag{14}$$

Now in any model either U_{Safe} holds or it does not. For the class of models $\mathbf{M}$ with $\neg U_{Safe}$, we have by (12) and (14) that $\mathbf{M} \models \neg Ho(Safe, 0) \wedge \neg Ho(Safe, 2)$. For the class $\mathbf{M}'$ with U_{Safe}, we have $\mathbf{M}' \models Ho(Safe, 0) \wedge Ho(Safe, 2)$. It immediately follows that there can be no models with $Ho(Safe, 0) \equiv \neg Ho(Safe, 2)$. It remains to be shown that there do exist models for T_{DF} with U_{Safe} and models with $\neg U_{Safe}$; such models can indeed easily be constructed; we omit the details. Summarizing:

Proposition 7.2 *There is a model* $\mathcal{M}$ *for* T_{DF} *with* $\mathcal{M} \models Ho(Safe, 0) \wedge Ho(Safe, 1) \wedge Ho(Safe, 2)$. *There is a model* $\mathcal{M}'$ *for* T_{DF} *with* $\mathcal{M} \models \neg Ho(Safe, 0) \wedge Ho(Safe, 1) \wedge \neg Ho(Safe, 2)$. *There are no models for* T_{DF} *with any other interpretation of* $Ho(Safe, t)$ *for* $t \in \{0, 1, 2\}$.

We remark that this allows us to specify what Giunchiglia and Lifschitz call 'non-Markovian' theories: the value of a fluent at time t may depend on its 'long-term' history, and not only on the situation in the world at time $t-1$.

7.4 Minimizing Occurrence of Events

We have seen how to deal with domains in which we allow one event at a time to happen. Following Morgenstern & Stein [28, 36], Baral, Gelfond and Provetti [3, 5] and several other authors, we would

like to extend this to the case where more than one event may happen at a time. We want our domains to be subject to the assumption that normally, events don't happen unless there is a specific reason for them to happen.

It is only now that the full strength of Pearlian theories comes to light: it turns out that we can properly model events, just like fluents, using structural equations! The advantage of using structural equations in stead of axioms in CONS will become clear in section 7.5. The new structural equations will say that 'if no intervention takes place that makes an event happen, and if nothing abnormal is the case, then the event will not happen'. The corresponding axiom in EQ will look as follows:

$$\forall e,t\ .\ [\ \neg Do(e,\text{TRUE},t)\ \ \wedge\ \ \neg Do(e,\text{FALSE},t)\ ]\ \supset\ [\ Ho(e,t)\equiv\ \ Ab_1(e,t)\ ] \tag{15}$$

This uses an 'abnormality predicate' $Ab_1(e,t)$ defined for all event-time pairs. An instantiated abnormality predicate plays a role similar to the 'assumption symbols' we have seen in before. But unlike these, which represented things we were completely ignorant about, the abnormalities stand for things we consider abnormal, or, in other words, *unlikely*. For this, we extend the notion of causal model to 'preferred causal model'. We always prefer those models of our theory that are the least 'abnormal'. Since in the next subsection we will encounter domains involving both *rather* and *highly* abnormal eventualities, we need to introduce *two* abnormality predicates in the definition below.

Definition 7.5 *A model $\mathcal{M}$ is a preferred causal model for causal theory T if*

1. $\mathcal{M}\models_{\mathbf{c}} T$ *and*
2. *There is no other* $\mathcal{M}'\models_{\mathbf{c}} T$ *with* $\mathcal{M}'[\![Ab_2]\!]\subset\mathcal{M}[\![Ab_2]\!]$ *and*
3. *There is no other* $\mathcal{M}''\models_{\mathbf{c}} T$ *with* $\mathcal{M}''[\![Ab_2]\!]=\mathcal{M}[\![Ab_2]\!]$ *and* $\mathcal{M}''[\![Ab_1]\!]\subset\mathcal{M}[\![Ab_1]\!]$

We show in appendix 1.3 that the notion of preferred models is already implicitly present in Pearl's original theories. Having defined preferred models, we are in a position to give the definition of causal theories with concurrent events. The only difference to the previous definition (definition 7.4) is that EQ must now also contain the 'no events'-axiom (15) while CONS does not contain the 'one-event-at-a-time' axiom (5) any more.

Definition 7.6 *A* first-order causal theory with persistence, dependent fluents and concurrent events *for the tuple* $\langle\mathbf{D},\mathbf{E},\mathbf{F}\rangle$ *is a tuple* $\langle$EQ, CONS$\rangle$ *where* EQ *and* CONS *are sets of sentences for the language* $\mathcal{L}(\mathbf{D},\mathbf{E},\mathbf{F})$, EQ *consists of intervention axioms (1),(2), persistence axiom (4) and no-events axiom (15) while* CONS *contains at least UNA- and DC-axioms for* $\mathbf{B},\mathbf{D},\mathbf{E}$ *and* $\mathbf{F}$.

Note first that with this definition, we *still* handle standard reasoning domains like the Yale Shooting Problem. To see this, let us consider the theory $T_{\text{YSP},2}$ which is as T_{YSP} but adapted to definition 7.6: EQ now contains axioms (1),(2), (4) and (15), while CONS contains UNA- and DC-axioms and additionally axioms (6)–(9). One sees that the circumscription of CONS rules out all models with $Do(e,b,t)$ for any e,b and t. It follows by axioms (15) and (9) that all models for $T_{\text{YSP},2}$ have the abnormalities

$$Ab_1(Load,0),\ Ab_1(Wait,1),\ Ab_1(Shoot,2) \tag{16}$$

Clearly, there will be no models with even more abnormalities: no more events will happen than those we specified to happen. It is easy to check that his means that in all preferred models for $T_{\text{YSP},2}$, the turkey is not alive anymore at $t=3$.

It may be argued that there is nothing 'abnormal' about the three events in the domain, so that conceptually speaking, using abnormalities may not be a good way to model them. In other words, a priori - if we have no observations about what happens in our domain - the events $Ho(Load,0)$, $Ho(Wait,1)$, $Ho(Shoot,2)$ are considered abnormal. But in situations in which it is *given* that they do happen (i.e. we have the additional axiom (9)), there is nothing abnormal about them any more.

If we take a probabilistic stance towards abnormalities, then there is no problem. In this case, 'abnormal' is identified with 'having small probability' and we always prefer the 'most probable models' [12]; see also appendix 1.3. A priori, the probability of each 'abnormal' event is very small, so the probability of models in which (9) and therefore, also (16) holds is very small too. But if it is given that (9) holds, then the probability of all models becomes *conditioned* on (9), which implies that the events $Ho(Load, 0)$, $Ho(Wait, 1)$, $Ho(Shoot, 2)$ receive probability one: there is nothing abnormal (unlikely) about them any more.

In our simple nonmonotonic formalism, we have no equivalent of the conditioning operator in probability theory. If, using probability theory, we condition on an event A, then we reduce the set of possible models to those in which A holds, but the 'preference ordering' between the models in which A holds is not changed by the conditioning: one can show using the basic rules of probability theory that for any $\mathcal{M}_1$ and $\mathcal{M}_2$, we have that if $\mathcal{M}_1 \models A$ and $\mathcal{M}_2 \models A$ and $P\{\mathcal{M}_1\} > P\{\mathcal{M}_2\}$ then $P\{\mathcal{M}_1|A\} > P\{\mathcal{M}_2|A\}$. Hence the model $\mathcal{M}$ for which $P\{\mathcal{M}|A\}$ is maximal is also the model $\mathcal{M}$ which maximizes $P\{\mathcal{M}\}$ within the set of models in which A holds.

This means that if we take a probabilistic view on nonmonotonic reasoning and interpret all abnormalities as 'things that happen with small probability', then there is nothing wrong with our minimization procedure: the models which have the least abnormalities always coincide with the models in CONS which are the most probable *given* all the axioms in CONS.

7.5 Causal Chains of Events

To see why it makes sense to put the axiom expressing the non-occurrence of events in EQ rather than CONS, we now turn to 'causal chains of events'. The idea here is very simple: suppose an event A 'causes' another event B, i.e. the event A is always accompanied by an intervention that triggers event B. Then, if the event A happens, we do not consider the event B 'abnormal' any more. The structural equation which says that 'event B occurring at time t is abnormal' is replaced by another equation that says 'event B does occur at time t' - just like structural equations concerning regular fluents get replaced if an intervention takes place.

Example 12 Imagine a domain where, if you push somebody, he or she falls down a moment later. We can formalize this using a causal theory according to definition 7.6 such that CONS contains the axiom

$$Ho(Push, t) \supset Do(Fall, \text{TRUE}, t+1)$$

Now suppose CONS further contains axiom $Ho(Push, 0)$. From inspection of axiom (15) and the intervention axioms (1) and (2), we see that all preferred models for this theory will have $Ho(Push, 0) \wedge Ho(Fall, 1)$. But if CONS had not contained $Ho(Push, 0)$, then the preferred models would be those in which no events at all take place.

7.6 Surprise, Surprise

What if a fluent changes value while we have no event in our domain which can account for that? This is the kind of surprise that Lifschitz and Rabinov [20] called a 'miracle'. In order to deal with it, we have to weaken our structural equations concerning regular fluents: persistence may now be broken not only by some intervention, but also by some 'abnormal' (unlikely) external influence. Thus axiom (4) becomes the following *weakened persistence axiom*:

$$\begin{aligned} &\forall f, t \,.\, (t > 0) \supset [\ (\neg Do(f, \text{TRUE}, t) \ \wedge \ \neg Do(f, \text{FALSE}, t)) \ \supset \\ &\quad [\ Ho(f, t) \equiv \ [\ (Ho(f, t-1) \wedge \neg Ab_2(f, t)) \vee (\neg Ho(f, t-1) \wedge Ab_2(f, t))\ ]\ ]\ ] \end{aligned} \tag{17}$$

The notation (17) has been chosen so as to conform to the syntactic form (3). An equivalent, perhaps more intuitive way of stating it is by the following two axioms, whose conjunction is equivalent to (17):

$$[\ (t > 0) \wedge \neg Do(f, \text{TRUE}, t) \wedge \neg Do(f, \text{FALSE}, t) \wedge \neg Ab_2(f, t)\ ] \supset [\ Ho(f, t) \equiv Ho(f, t-1)\ ]$$

$$[\ (t > 0) \wedge \neg Do(f, \text{TRUE}, t) \wedge \neg Do(f, \text{FALSE}, t) \wedge Ab_2(f, t)\ ] \supset \neg[\ Ho(f, t) \equiv Ho(f, t-1)\ ]$$

In other words, if no interventions and no abnormalities take place, then the value of a fluent persists. If no interventions take place and the value of the fluent does not persist, then something abnormal is the case.

Here is our updated definition:

Definition 7.7 *A* first-order causal theory with persistence, dependent fluents, concurrent events and surprises *for the tuple* $\langle \mathbf{D}, \mathbf{E}, \mathbf{F} \rangle$ *is a tuple* $\langle \text{EQ}, \text{CONS} \rangle$ *where* EQ *and* CONS *are sets of sentences for the language* $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F})$*,* EQ *consists of intervention axioms (1),(2),* weakened *persistence axiom (17) and no-events axiom (15) while* CONS *contains at least UNA- and DC-axioms for* $\mathbf{B}, \mathbf{D}, \mathbf{E}$ *and* $\mathbf{F}$*.*

We see that the only difference to the previous definition 7.6 is that persistence axiom (4) has been replaced by its weakened version (17). We illustrate the use of 'surprises' by the following example:

Example 13 [**stolen car problems**] Consider two variations of the 'Stolen Car Problem' [17], modeled by theories $T_{\text{SC},1} = \langle \text{EQ}, \text{CONS}_1 \rangle$ and $T_{\text{SC},2} = \langle \text{EQ}, \text{CONS}_2 \rangle$. Both are causal theories according to def. 7.7 for the tuple $\langle \mathbf{D}, \mathbf{E}, \mathbf{F} \rangle$ with $\mathbf{D} = \emptyset$, $\mathbf{E} = \{Steal\_Car\}$, $\mathbf{F} = \{Car\_In\_Lot\}$. On top of the UNA- and DC-axioms, CONS_1 and CONS_2 both contain observations:

$$Ho(Car\_In\_Lot, 0) \wedge \neg Ho(Car\_In\_Lot, 10)$$

CONS_2 contains no further axioms, while CONS_1 additionally contains:

$$Ho(Steal\_Car, t) \supset Do(Car\_In\_Lot, \text{FALSE}, t+1)$$

Now let t^* be some element of $\{0, \ldots, 9\}$ and consider a model $\mathcal{M}$ with

1. $\mathcal{M} \models Ho(Steal\_Car, t^*)$.
2. $\mathcal{M} \models Do(Car\_In\_Lot, \text{FALSE}, t^*+1)$ and $\mathcal{M} \not\models Do(x, b, t)$ for any other x, b, t.
3. $\mathcal{M} \models \forall f, t \,.\, \neg Ab_2(f, t)$
4. $\mathcal{M} \models Ho(Car\_In\_Lot, t+1) \equiv Ho(Car\_In\_Lot, t)$ for all t except $t = t^*$.

Clearly, $\mathcal{M} \models Circum(\text{CONS}_1, Do)$. It can be easily checked that $\mathcal{M} \models \text{EQ}$ too, so $\mathcal{M} \models_{\mathbf{c}} T_{\text{SC},1}$. Since $\mathcal{M}$ has no abnormalities of the Ab_2-kind, it follows that $\mathcal{M}$ will be preferred over any model which does have these. On the other hand, we have $\mathcal{M} \models Ab_1(Steal\_Car, t^*)$ but since there is no model which has neither Ab_1- nor Ab_2-abnormalities it clearly follows that $\mathcal{M}$ is a *preferred* model for $T_{\text{SC},1}$.

In $T_{\text{SC},2}$ things look different: there is no event which can account for the disappearance of the car. By arguments similar to those above, we find that all preferred models for $T_{\text{SC},2}$ do have $Ab_2(Car\_In\_Lot, t)$ for some $t \in \{1, \ldots, 10\}$.

So we see that in domains where a change of fluents happens for which there *is* an action in the domain that accounts for it, it will preferably be assumed that this action takes place than that some external 'miracle' or 'surprise' happens.

8. Comparison to Other Approaches

We now give a detailed comparison of our causal theories to three existing approaches: those of McCain & Turner [23, 25, 24], Lin [21, 22] and Baral, Gelfond and Provetti [3, 4, 5]. Each of the three approaches will be compared to the instantiation of our theory that is conceptually closest to it. Hence McCain & Turner's approach, which has been specifically designed to handle the ramification problem, will be compared to our 2-point causal theories which are complex enough to deal with ramifications but which contain no additional features that would complicate the comparison. In the same vein, Lin's approach will be compared to first-order theories for persistence in which only one event may happen at a time, and Baral, Gelfond & Provetti's method will be compared to the version of our theory that minimizes occurrences of events but that does not handle surprises.

In the case of McCain & Turner and Baral, Gelfond and Provetti we give and prove 'equivalence theorems' stating that for large classes of reasoning domains these approaches give the same results as our causal theories do. We will also give examples of reasoning domains that fall outside these classes for which, we claim, our approach works better than the one we are comparing it to.

In order to state our equivalence theorems we have to define two 'correspondence relations' $\sim$, such that $T_A \sim T_B$ iff domain descriptions T_A of approach A and T_B of approach B intuitively encode the same domain knowledge. $T_A \sim T_B$ should be pronounced as 'theory T_A *syntactically corresponds* to theory T_B'. In the first correspondence relation, A will be our two-point causal theories while B will be McCain & Turner's approach. In the second relation, A will be our first-order causal theories while B will be Baral, Gelfond & Provetti's approach. All approaches we consider have as their basic objects fluents and events; proving that A and B are equivalent then amounts to showing that for all corresponding theories T_A and T_B (i.e. $T_A \sim T_B$) we have that approach A selects a model $\mathcal{M}_A$ with a particular history of what fluents and events hold at what time iff B selects a model M_B with the same interpretation of event/fluent-time pairs. We will say that such models *semantically correspond*, written as $\mathcal{M}_A \cong \mathcal{M}_B$. Since for many well-formed theories of both approaches, $\sim$ will not be defined, $\sim$ implicitly imposes constraints on the class of reasoning domains for which the equivalence holds.

The importance of the equivalence theorems is that they show that some approaches which at first glance look rather different from ours are actually quite similar; therefore they are also connected to Pearl's sufficient cause principle, albeit implicitly. In our comparison to Lin's [21, 22] approach, we have not bothered to try and prove such a theorem, since it is easy to see, just by looking at the definitions, that his approach is almost equivalent to ours.

8.1 McCain & Turner's Theory of Ramifications and Qualifications

Recently, McCain & Turner (MT) have introduced a 'causal theory' that focuses on handling ramification constraints [23, 25, 24]. We will now compare their approach to our two-point causal theories as defined by definition 4.2 and 4.3.

MT consider theories that are triplets (S, E, C), defined for a propositional language where each atom stands for a fluent. The 'state of the world' S is an interpretation for all fluents. An interpretation is denoted by the set of literals true in it. E is a set of 'explicit effects', i.e. propositional combinations of fluents. Intuitively, they are the formulas that are explicitly caused to hold by some (unspecified) action. C is a set of 'causal laws' that determine the ramifications of effects. MT define the function[4] $\Pi_1(S, E, C)$ such that it gives the set of possible states of the world after an action with effects E has taken place in state S. The exact definition of $\Pi_1(S, E, C)$ can be found in appendix 3.1; here we just give an example of its use:

$$\begin{aligned} S &= \{Alive, Walking\} \\ E &= \{\neg Alive\} \\ C &= \{\neg Alive \Rightarrow \neg Walking\} \end{aligned} \tag{1}$$

In this case, $\Pi_1(S, E, C) = \{\{\neg Alive, \neg Walking\}\}$ i.e. the change of *Alive* brought about a change of *Walking*. However, if we had had $S' = \{\neg Alive, \neg Walking\}$, $E' = \{Walking\}$, $C' = C$, then $\Pi_1(S', E', C')$ would have been empty: '$\Rightarrow$' has a function similar to our '*Do*', enabling changes of right-hand side fluent values given changes of left-hand side values, but *not* the other way around (compare this to example 3). There is one sort of domain constraint that can be expressed in MT's approach but not in ours: MT allow effects to be *any* propositional combination of fluents, and thus an effect may be a disjunction of two fluents. This cannot be expressed by our 2-point causal theories (we have no construct of the form $Do(X \vee Y, b)$). But it is exactly here that MT can give counterintuitive results; to see this, consider the general case where there is an effect X that further causes $Y \vee Z$, and

[4] $\Pi_1(S, E, C)$ is the function McCain & Turner use in their recent paper [25]; it stands at the basis of their 'Causal Theory of Action and Change' [24]. It is a slight modification of their earlier next-state function $Res^4_C(E, S)$ [23]. For the precise relation between $\Pi_1(S, E, C)$ and $Res^4_C(E, S)$, see [25].

an initial state with $\neg X, \neg Y$ and $\neg Z$:

$$\begin{aligned} S &= \{\neg X, \neg Y, \neg Z\} \\ E &= \{X\} \\ C &= \{X \Rightarrow (Y \vee Z)\} \end{aligned}$$

Proposition 8.1 $\Pi_1(S, E, C) = \{\{X, Y, \neg Z\}, \{X, \neg Y, Z\}\}$

The proof of this proposition can be found in appendix 3.2. We have seen in example 7 of section 6 that this seems too strong in general.

Another difference between MT's approach and ours has already been pointed out in section 5.1: there are domain constraints that would be modeled as a single axiom $A \Rightarrow B$ in MT's approach, while we prefer to model them using two separate axioms:

$$\begin{aligned} Do(A(t), \text{TRUE}) &\supset Do(B(t), \text{TRUE}) \\ A(t) &\supset B(t) \end{aligned} \tag{2}$$

While McCain & Turner's semantics treats '$A \Rightarrow B$' as we would treat the single axiom:

$$A(t) \supset Do(B(t), \text{TRUE}) \tag{3}$$

We have seen in section 5.2 that this is in general *not* the same as (2).

But if we translate constraints like $A \Rightarrow B$ indeed as (3), and we restrict ourselves to domains without disjunctive effects, then it turns out that MT and our approach agree on all problem domains that can be represented in the languages of both approaches. In definitions 8.1 and 8.2 (page 23) syntactic ($\sim$) and semantical ($\cong$) correspondence for 2-point causal theories and MT's theories are defined. $\sim$ has been defined such that disjunctive effects cannot occur in corresponding theories and such that rules of the form '$A \Rightarrow B$' are translated into sentences of the form (3). $\cong$ is defined such that $\mathcal{M} \cong (S, S')$ iff the fluents that hold in S hold in $\mathcal{M}$ at time 0 and the fluents that hold in S' hold in $\mathcal{M}$ at time 1. As an example of how definition 8.1 works, we consider the causal theory (S, E, C) defined by (1) at the beginning of this section. The theory described there corresponds to a 2-point causal theory T_{C} (the subscript C stands for 'Causal') such that CONS contains the axioms:

$$\begin{aligned} \mathit{Alive}(0) \wedge \mathit{Walking}(0) & \\ Do(\mathit{Alive}(1), \text{FALSE}) & \\ \neg \mathit{Alive}(1) \supset Do(\mathit{Walking}(1), \text{FALSE}) & \end{aligned} \tag{4}$$

which can be seen to stem from items 3,4 and 5 of definition 8.5, respectively. EQ contains

$$\mathit{Alive}(1) \equiv \mathit{Alive}(0) \ ; \ \mathit{Walking}(1) \equiv \mathit{Walking}(0)$$

which can be seen from item 1 of the definition.

We are now ready to state our equivalence theorem, the proof of which can be found in appendix 3.3.

Theorem 8.1 *For any 2-point causal theory T_{C} and any domain description $T_{\text{MT}} = (S, E, C)$ such that $T_{\text{C}} \sim T_{\text{MT}}$, we have:*

$$S' \in \Pi_1(S, E, C) \Rightarrow \textit{ there exists an } \mathcal{M} \models_{\mathbf{c}} T_{\text{C}} \textit{ with } \mathcal{M} \cong (S, S')$$

$$\mathcal{M} \models_{\mathbf{c}} T_{\text{C}} \Rightarrow \Pi_1(S, E, C) \textit{ contains an } S' \textit{ with } \mathcal{M} \cong (S, S')$$

For any $F_i \in \mathbf{F}$ we will sometimes write $F_i^{\rm TRUE}$ to denote F_i and $F_i^{\rm FALSE}$ to denote $\neg F_i$.

Definition 8.1 *For any 2-point causal theory $T_{\rm C} = \langle \mathbf{V}, \mathbf{U}, {\rm EQ}, {\rm CONS}\rangle$ for sets $\mathbf{E}$ and $\mathbf{F}$ and any $T_{\rm MT} = (S, E, C)$ of MT's approach we define $T_{\rm C} \sim T_{\rm MT}$ ('$T_{\rm C}$ corresponds to $T_{\rm MT}$') to be true iff all of the following hold:*

1. *$\mathbf{E} = \emptyset$; $\mathbf{F} = \{F_1, \ldots, F_n\}$; $F_i \in \mathbf{F}$ iff F_i is an atom in the propositional language for which $T_{\rm MT}$ is defined; $\mathbf{U} = \emptyset$; EQ is defined as required by def. 4.2.*
2. *Each sentence in CONS corresponds either to S or to a sentence in E or to a sentence in C. The set S, the sentences in E and C all correspond to a sentence in CONS. Here 'correspondence' is defined as follows:*
3. *the sentence $\Gamma_{\rm C}$ in CONS corresponds to S (and vice versa) iff $\Gamma_{\rm C}$ is of the form $F_1^{b_1}(0) \wedge \ldots F_n^{b_n}(0)$ and $S = (F_1^{b_1}, \ldots, F_n^{b_n})$.*
   Here $b_i \in \mathbf{B}$ for all $1 \leq i \leq n$. .
4. *A sentence $\Gamma_{\rm C}$ in CONS corresponds to a sentence $\Gamma_{\rm MT}$ in E iff $\Gamma_{\rm C}$ is of the form $Do(F_i(1), b)$ and $\Gamma_{\rm MT}$ is of the form F_i^b. Here $b \in \mathbf{B}$.*
5. *A sentence $\Gamma_{\rm C}$ in CONS corresponds to a sentence $\Gamma_{\rm MT}$ in C iff $\Gamma_{\rm C}$ is of the form ($m \geq 1$)*

$$\Phi \supset Do(F_{i_1}(1), b_1) \wedge \ldots \wedge Do(F_{i_m}(1), b_m) \tag{5}$$

and $\Gamma_{\rm MT}$ is of the form

$$\Phi' \Rightarrow F_{i_1}^{b_1} \wedge \ldots \wedge F_{i_m}^{b_m} \tag{6}$$

Here Φ is a propositional combination of atoms of the form $F(1)$ where F is any element of $\mathbf{F}$. Φ' is the result of replacing all occurrences of $F(1)$ by F. For all j, $1 \leq i_j \leq n$ and $b_j \in \mathbf{B}$.

Definition 8.2 *Given a model $\mathcal{M}_C$ for a causal theory $T_{\rm C}$ and a pair of states (S, S') for a $T_{\rm MT}$ with $T_{\rm C} \sim T_{\rm MT}$ we say that $\mathcal{M}_C$ corresponds to (S, S') iff for all $F_i \in \mathbf{F}$ we have*

$$\mathcal{M}_C \models F_i(0) \Leftrightarrow F_i \in S \quad \text{and} \quad \mathcal{M}_C \models F_i(1) \Leftrightarrow F_i \in S'$$

Lin's Procedure Start with a theory $T_{\rm LIN}$ that consists of 1) unique names axioms for all actions and fluents 2) unique names and domain closure-axioms for truth-values and 3) domain-specific axioms.

1. Add the following basic axioms for the predicate *Caused* to $T_{\rm LIN}$:

$$Caused(f, \text{TRUE}, s) \supset Ho(f, s) \tag{7}$$
$$Caused(f, \text{FALSE}, s) \supset \neg Ho(f, s) \tag{8}$$

2. Circumscribe *Caused* in $T_{\rm LIN}$ with all other predicates fixed. Let $T'_{\rm LIN}$ be the resulting theory.
3. Add to $T'_{\rm LIN}$ the following 'frame axiom' and let $T''_{\rm LIN}$ be the resulting theory.

$$Poss(a, s) \supset \{(\neg Caused(f, \text{TRUE}, Result(a, s)) \wedge \neg Caused(f, \text{FALSE}, Result(a, s))) \supset [Ho(f, Result(a, s)) \equiv Ho(f, s)]\} \tag{9}$$

4. Maximize *Poss* in $T''_{\rm LIN}$ to obtain the final theory $T'''_{\rm LIN}$.

8.2 Lin's Embrace of Causality

Lin [21, 22] has recently introduced a new method for reasoning about action that is based on a version of the situation calculus [27]. In the situation calculus, time is modeled by *situations* s rather than time points. The sort *actions* corresponds to our *events*; however, in Lin's approach symbols of the sort are not necessarily constants (i.e. they can be n-ary function symbols rather than only 0-ary). Similarly, fluents in situation calculus correspond to our fluents, and again, Lin allows fluent symbols to be n-ary. For any action e and situation s, the function[5] $Result(e, s)$ stands for the situation that results when performing e in s. Ordinary situation calculus contains only the functions described above and the predicate *Ho*, defined on fluent-situation pairs. In addition to this, Lin uses two additional predicates *Caused* and *Poss*.

It turns out that Lin's method is *very* similar to ours. We will compare it to the instantiation of causal theories in which only one event is allowed to happen at a time (i.e. definition 7.3), since this instantiation is conceptually closest to the situation calculus. Like us, Lin uses an additional sort *truth values*; a variable of the sort can be either TRUE or FALSE. Lin's *Caused*-predicate is ternary: $Caused(f, v, s)$ is true if the fluent f is caused to have the truth value v in situation s. On page 24 we give a (somewhat extended) quote of [21] which describes the method. The last step of Lin's method is meant to deal with the qualification problem which we do not address in this paper, so it will be of no concern to us. We first need the following:

Proposition 8.2 *If we exchange step 1 and step 2 in Lin's procedure we will arrive at an equivalent final theory* $T'''_{\rm LIN}$.

Proof: If we rewrite (7) and (8) as disjunctions, then *Caused* appears only in negated form in them. It then easily follows from the model-theoretic characterization of circumscription (see appendix 2) that the theory obtained by adding (7) and (8) before the circumscription has the same models as the theory obtained by adding them after the circumscription. □

[5] The actual name for the function used by Lin is '*Do*' in stead of *Result*; we use the name *Result* in order to avoid confusion with our own *Do-predicate*.

Let us write $\text{EQ}_{\text{LIN}} = \{(7) \cup (8) \cup (9)\}$. With this notation, step 1 and step 2 interchanged and step 4 left out, Lin's approach can be rephrased as follows:

$$T''_{\text{LIN}} = \text{EQ}_{\text{LIN}} \cup \mathit{Circum}(T_{\text{LIN}} ; \mathit{Caused}) \tag{10}$$

Now if we compare EQ_{LIN} to the set of axioms EQ in our definition of causal theories (def. 7.3), we see that they are strikingly similar: Once we rename *Caused* to *Do*, (7) and (8) actually become equivalent to our intervention axioms (1) and (2)! If we furthermore realize that $\mathit{Result}(a, s)$ refers to the first time point considered after the time point corresponding to situation s, we see that also (9) is almost the same as our persistence axiom (4). And most importantly, if we compare the characterization of Lin's approach (10) to the definition of models for causal theories (def. 7.2), we see that, if EQ is as in definition 7.3, then (10) above is nearly equivalent to definition 7.2, item 1. The differences are that our approach uses integer time while Lin's uses situations and that our approach does not handle the qualification problem and infinite numbers of fluents and events (CONS contains a domain closure axiom while T_{LIN} does not). On the other hand, Lin's approach does not handle the extensions to our approach introduced after section 7.2. Otherwise, as is hopefully clear from just looking at the respective definitions, Lin's approach is almost identical to ours. Though we have not proven it formally, we conjecture the two approaches to be equivalent on the set of reasoning domains for which both are defined.

Interestingly, in Lin's papers neither the choice to keep *Ho* fixed during circumscription of *Caused*, nor the choice to add the persistence axiom only after this first circumscription is motivated in terms much other than 'if you do not do it, you get counterintuitive results'. Our work thus provides an external motivation for these choices, cf. the remarks in section 4: the axioms (7)-(9) can be interpreted as mimicking the replacement of structural equations. This replacement can only be done properly if the right interpretations of *Do* have been obtained already for each particular interpretation of *Ho*. Note also that, for the 'replacement' to happen correctly, it is *not* crucial that the intervention axioms are added after the circumscription of CONS (which is why Lin can add his version of them to T_{LIN} before circumscribing). It *is* however crucial that the persistence axiom is only added after the circumscription. Otherwise, as can be easily seen from the persistence axiom (4), we would infer that in each model with $\neg[\mathit{Ho}(F, t-1) \equiv \mathit{Ho}(F, t)]$ for some fluent F, we would have $\mathit{Do}(F, b, t)$ for some b. Thus any change would automatically be accompanied by an intervention, and we would select models with spurious changes.

8.3 Baral and Gelfond

We now undertake the most difficult and extensive comparison: we compare our approach to Baral and Gelfond's (BG) approach based on the *action description language* $\mathcal{L}_3$ [3]. $\mathcal{L}_3$ is an extension of Baral, Gelfond and Provetti's language $\mathcal{L}_1$ [5] which in turn is an extension of Lifschitz' language $\mathcal{A}$ [10, 34]. $\mathcal{L}_3$ extends $\mathcal{A}$ to deal with concurrent actions, actions with non-deterministic effects and observations of the actions that take place and the fluents that hold in arbitrary situations. On the other hand, it cannot at all deal with ramifications. The way it treats concurrent actions is similar to the way we treat events: BG always prefer the models of a domain in which as few actions as possible take place. Otherwise, at least on the surface, BG's approach looks quite different from ours. It is definitely not based on Pearl's ideas. Still, it turns out to be equivalent to our approach on most reasoning domains for which both approaches are defined.

BG's approach consists of *domain descriptions* D written in the language $\mathcal{L}_3$. We will compare BG's approach to our instantiation of causal theories in which concurrent events are allowed to happen but surprises are not, i.e. we use definition 7.6. The comparison will follow the same pattern as the comparison to MT's approach did: we first give an example of a reasoning domain where our approach gives better results than BG's does. We will then define syntactical and semantical correspondence relations and provide a theorem stating that, for the subset of possible reasoning domains for which the syntactical correspondence relation is defined, corresponding theories have corresponding models.

We first have to explain the basics of BG's approach though. For more details we refer to [3] and especially to the much more extended [5]. Strictly speaking, the latter reference is about $\mathcal{L}_1$ and not $\mathcal{L}_3$, but $\mathcal{L}_3$ is only a minor extension of $\mathcal{L}_1$.

Review of $\mathcal{L}_3$ BG's approach consists of *domain descriptions* D written in a language $\mathcal{L}_3$. If a domain description D is *consistent*, then it has *models* M which determine what fluents hold at what time and what actions happen when. $\mathcal{L}_3$ consists of the sets of symbols $\mathcal{F}$ (corresponding to our fluents), $\mathcal{A}$ ('unit actions', corresponding to our events) and $\mathcal{S}$ ('situations', corresponding to states of the world at specific points in time). $\mathcal{S}$ contains two special situations s_0 and s_N: the *initial* and *current* situation. A *fluent literal* is a fluent possibly preceded by $\neg$. By a *generalized action* a, BG mean a disjunction of arbitrary sets of unit actions: $a = a_1|\ldots|a_m$ $(m \geq 1)$; $a_i = \{a_{i1}, \ldots, a_{in}\}$ for $1 \leq i \leq n$. Each a_i is called a *compound* action and interpreted as a set of actions which are performed concurrently and which start and stop contemporaneously. If a generalized action is performed, this means that one of the constituent compound actions is performed and it is not known which.

Domain descriptions D in $\mathcal{L}_3$ may contain several kinds of rules. First, there are *effect laws* of the form

$$a \textbf{ causes } f \textbf{ if } p_1, \ldots, p_n$$

where a is a compound action and $f, p_1, \ldots, p_n$ $(n \geq 0)$ are fluent literals. This should be read as 'f is guaranteed to be true after the execution of an action a in any state of the world in which $p_1, \ldots, p_n$ are true'.

Second, there are *fluent facts* of the form f **at** s where f is a fluent literal and s is a situation. This should be read as 'f is observed to be true in situation s'.

Third, there are *occurrence facts* which are expressions of the form α **occurs_at** s where α is a sequence of generalized actions and s is a situation. This says that 'the sequence α of actions was observed to have occurred in situation s'.

Fourth, there are *precedence facts* of the form s_1 **precedes** s_2. This states that situation s_1 occurred before s_2.

The three kinds of facts introduced above are called *atomic*. A *fact* is a propositional combination of atomic facts[6].

An *interpretation* $M = (\Psi, \Sigma)$ for a domain description D contains a 'situation assignment' Σ and a 'causal interpretation' Ψ.

A *situation assignment* is a mapping from $\mathcal{S}$ to sequences of actions, such that 1) $\Sigma(s_0) = []$ ($[]$ denotes the empty sequence) and 2) for every $s_i \in \mathcal{S}$, $\Sigma(s_i)$ is a prefix of $\Sigma(s_n)$. Intuitively, Σ defines an 'action schedule': it says which actions happen in between which situations. A *causal interpretation* Ψ maps sequences of actions to 'states'. A *state* σ is an interpretation of all fluents, denoted by the set of fluents that are interpreted to be true. Hence $\Sigma(s)$ denotes the sequence of actions that have led to situation s, and $\Psi(\Sigma(s))$ denotes the set of fluents that hold in situation s.

If all the rules of a domain description D are true in an interpretation M, we call M a *model* of D. The definition a a rule 'being true in interpretation M' is relatively straightforward; the only complication arises in the case of several actions with contradictory effects that take place at the same time – and it is precisely here that BG's approach can give unintuitive results, see below. The precise definition of 'truth in a model' and of modelhood can be found in appendix 4.1.

A Distinguishing Example We will see that for most of the domains expressible in both formalisms, BG and our causal theories give the same results. However, in domains involving *specificity*, BG and our approach give different results, and we claim that for these domains, BG gives the less intuitive ones. The default assumption of specificity says that 'more specific information about actions overrides less specific information'. We illustrate this using a standard example [37]: suppose that if you lift a

[6] There is one extra kind of rule in $\mathcal{L}_3$, the *hypothesis*. Hypotheses however cannot occur in domain descriptions D and will therefore be of no concern to us.

bowl of soup with either your left or your right hand, but not both, then you will spill the soup and the table will get wet. If you lift the bowl with both hands however, then you will not spill the soup. BG formalize this using the following domain description D, containing only two situations s_0 and s_N:

$$\{\mathit{Lift\_left}\} \quad \textbf{causes} \quad \mathit{Wet} \tag{11}$$

$$\{\mathit{Lift\_right}\} \quad \textbf{causes} \quad \mathit{Wet} \tag{12}$$

$$\{\mathit{Lift\_left}, \mathit{Lift\_right}\} \quad \textbf{causes} \quad \neg \mathit{Wet} \ \textbf{if} \ \neg \mathit{Wet} \tag{13}$$

BG now formalize the assumption of specificity as follows: if the preconditions of (13) hold, then rules (11) and (12) are ignored when determining the effects of *Lift_left* and *Lift_right* – see the paragraph on 'causal models' in appendix 4.1 for details on how this is achieved.

Now in this simple example BG's approach works well. However, suppose your table is placed in your garden, where there is also a sprinkler very near to your table, obeying the following additional rule:

$$\mathit{Turn\_on\_sprinkler} \ \textbf{causes} \quad \mathit{Wet} \tag{14}$$

Now consider a situation in which your table is dry. If in this situation you lift the soup bowl with both hands while somebody else turns on the sprinkler, then the result according to BG's approach turns out to be undefined! More formally, suppose we have the additional facts

$$s_0 \ \textbf{precedes} \ s_N \ \wedge \neg \mathit{Wet} \ \textbf{at} \ s_0 \ \wedge$$
$$[\{\mathit{Turn\_on\_sprinkler}, \mathit{Lift\_left}, \mathit{Lift\_right}\}] \ \textbf{occurs\_at} \ s_0 \tag{15}$$

Proposition 8.3 *In any model M for $D = \{(11) - (15)\}$ it remains undefined what fluents hold and what not in situation s_N.*

The proof of this proposition can be found in appendix 4.2.

One should stress that there is nothing wrong with 'specificity' in itself! From our point of view, specificity just says that, just as you assume that 'no events happen in general' when determining your set of models, you may already assume it in your specification of effect axioms; we just did not build in this feature in our causal theories. The real problem lies in the inappropriate use of **causes** in (13): there is definitely no *sufficient cause* for not getting wet if you lift the soup bowl with two hands: no intervention that sets the value of *Wet* is performed. However, **causes** does receive a semantics in BG's approach as if it would always represent an intervention. From a Pearlian point of view, it comes as no surprise that this may lead to counterintuitive results.

Because we do not feature specificity, we would have to formalize (11) as

$$\mathit{Ho}(\mathit{Lift\_left}, t) \wedge \neg \mathit{Ho}(\mathit{Lift\_right}, t) \supset \mathit{Do}(\mathit{Wet}, \textsc{true}, t+1) \tag{16}$$

and (12) accordingly. (13) would then simply disappear. If the sprinkler were turned on, then according to the definition of models for causal theories (def. 7.2), you would definitely get *Wet*.

Correspondence and Equivalence We will see that BG's approach and ours are equivalent on all domains for which both are defined except those involving specificity. In this section we first define syntactic ($\sim$) and semantic ($\cong$) correspondence, and we then provide a theorem stating that syntactically corresponding theorems yield semantically corresponding models. However, before we can do all this, we have to take care of four technical problems that arise. We consider each of these in turn:

- First, we have to rule out domain descriptions D such as the above which may involve specificity and which thus may be handled differently by the two approaches. These are the *ambiguous* D:

Definition 8.3 *If a domain description D in the language $\mathcal{L}_3$ contains two effect laws*

$$a \textbf{ causes } f \textbf{ if } p_1, \ldots, p_n \quad \text{and} \quad a' \textbf{ causes } \neg f \textbf{ if } p'_1, \ldots, p'_m \tag{17}$$

where 1) $a \cap a' \neq \emptyset$ or $a = \emptyset$ or $a' = \emptyset$ and 2) $\{p_1, \ldots, p_n\} \cap \{\neg p'_1, \ldots, \neg p'_m\} = \emptyset$ then D is said to be **ambiguous**.

The exclusion of ambiguous domain descriptions will be echoed in condition (b) of theorem 8.2 below.

- Another problem is that in BG's approach, there exist domain descriptions D for which a model M *does* exist, but the model is such that, for all states of the world after some particular point in time, it is *undefined* what fluents hold in them (proposition 8.3 illustrates this; there *is* a model for D, but it is undefined what happens in there). From our point of view, such a model containing undefined states is simply not a model at all, and we have to define semantical correspondence '$\cong$' such that no model of our approach corresponds to it. This will be echoed in condition (a) of the definition of 'semantical correspondence' (def. 8.6).

- The third problem is that the definition of modelhood comes in two versions in BG's papers [5]. In one of the two versions, logically consistent initial states are *never* ruled out, even if this would lead to models in which more actions occur than is strictly necessary. In the other version only the models with a minimal number of actions are selected. Since in our approach we always prefer the models with the least number of events (i.e. the smallest, in the subset sense, interpretations of Ab_1), our approach should clearly be compared to the second version. We therefore assume, in the following, that whenever we speak of a model of BG's approach, we mean a model according to the second version of the definition of modelhood – this second version is the definition given in appendix 4.1.

- The fourth problem is that BG use *names* for situations. In order to create corresponding theories, we need names for our time points, too. The definition of our language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F})$ does not allow for this. We therefore have to extend definition 7.6 in the following straightforward way:

  Definition 8.4 *A first-order causal theory $T_{\mathrm{C}} = \langle \mathrm{EQ}, \mathrm{CONS} \rangle$ for a language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F}, \mathbf{T})$ (where $\mathbf{T}$ is a finite set of constants) is a first-order causal theory with persistence, dependent fluents and concurrent events (i.e. a theory according to definition 7.6 for the tuple $\langle \mathbf{D}, \mathbf{E}, \mathbf{F} \rangle$ extended with the constants in $\mathbf{T}$). These constants are all of sort time points.*

  Notice that the definition of models for causal theories (definition 7.2) makes sure that the constants in $\mathbf{T}$, which we will call *time names*, will always be interpreted as nonnegative integers. Whenever in the following we speak of T_{C} , we mean a theory T_{C} defined according to definition 8.4 above.

Having taken care of these problems, we are ready to define syntactic correspondence. This is done in definition 8.5 on page 29. As an example of how that definition works, we will give a simple domain description and a causal theory that corresponds to it. For this, let D be the domain description consisting of (11), (12),(14) and (15). This is just the soup-bowl domain we considered before but without the malfunctioning specificity axiom (13). According to definition 8.5, this domain is equivalent to a causal theory T_{C} with an EQ containing the standard two intervention axioms and our persistence and no-events axiom (15), and a CONS containing UNA- and DC-axioms and on top of

For fluents $F_i \in \mathbf{F}$ we use the following notation: when occurring in a formula of $\mathcal{L}_3$, F_i^{TRUE} should be read as F_i; F_i^{FALSE} should be read as $\neg F_i$. When occurring in a sentence of CONS, $Ho(F_i^{\mathrm{TRUE}}, t)$ should be read as $Ho(F_i, t)$; $Ho(F_i^{\mathrm{FALSE}})$ should be read as $\neg Ho(F_i, t)$.

Definition 8.5 *For any theory T_{C} for a language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F}, \mathbf{T})$ and domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$, we define $T_{\mathrm{C}} \sim D$ ('T_{C} corresponds to D') to hold iff all of the following hold:*

1. **constants** *1)* $\mathbf{D} = \emptyset$, *2)* $\mathbf{E} = \mathcal{A}$; *3)* $\mathbf{F} = \mathcal{F} = \{F_1, \ldots, F_{n_F}\}$; *4)* $\mathbf{T} = \mathcal{S}$; *and 4)* $s_0, s_N \in \mathcal{S}$.

2. **general axioms** EQ *is as required by definition 8.4.* CONS *contains UNA- and DC- axioms for* $\mathbf{E}, \mathbf{F}$ *and* $\mathbf{B}$ *and the additional axiom* $s_0 = 0$.

3. **sentences** *Each sentence in* CONS *that is not equal to one of the sentences mentioned under item 2. above, corresponds to either an effect law or a fact in D. Each effect law and each fact in D corresponds to a sentence in* CONS. *Here 'correspondence' is defined as follows:*

4. **effect laws** *A sentence Γ_{C} corresponds to an effect law L in D iff Γ_{C} is of the form*

$$\forall t.[\, Ho(F_{j_1}^{b_1}, t) \wedge \ldots \wedge Ho(F_{j_m}^{b_m}, t) \;\wedge Ho(a^1, t) \wedge \ldots \wedge Ho(a^n, t)\,] \supset Do(F_i, b, t+1) \quad (18)$$

while L is of the form:

$$\{a^1, \ldots, a^n\} \textbf{ causes } F_i^b \textbf{ if } F_{j_1}^{b_1}, \ldots, F_{j_m}^{b_m} \quad (19)$$

5. **fluent facts** *A sentence Γ_{C} corresponds to a fluent fact F in D iff Γ_{C} is of the form '$Ho(f, s)$' while F is of the form 'f **at** s'.*

6. **occurrence facts** *A sentence Γ_{C} corresponds to an occurrence fact O in D iff O is of the form '$[a_1, \ldots, a_n]$ **occurs_at** s', where $n > 0$, $a_i = a_{i1} \mid \ldots \mid a_{im_i}$, $a_{ij} = \{a_{ij}^1, \ldots, a_{ij}^{k(i,j)}\}$, $a_{ij}^{k(i,j)} \in \mathcal{A}$, while Γ_{C} is of the form:*

$$\begin{array}{l} [\, \mathbf{H}(a_{11}, s) \vee \quad \ldots \quad \vee \mathbf{H}(a_{1m_1}, s)\,] \;\wedge \\ [\, \mathbf{H}(a_{21}, s) \vee \quad \ldots \quad \vee \mathbf{H}(a_{2m_2}, s)\,] \;\wedge \\ \qquad\qquad \vdots \\ \wedge\, [\, \mathbf{H}(a_{n1}, s+n-1) \vee \quad \ldots \quad \vee \mathbf{H}(a_{nm_n}, s+n-1)\,] \end{array} \quad (20)$$

where $\mathbf{H}(a_{ij}, s)$ is short for $Ho(a_{ij}^1, s) \wedge \ldots \wedge Ho(a_{ij}^{k(i,j)}, s)$.

7. **precedence facts** *A sentence Γ_{C} corresponds to a precedence fact P in D iff Γ_{C} is of the form '$s_1 < s_2$' while P is of the form 's_1 **precedes** s_2'.*

8. **non-atomic facts** *A sentence Γ_{C} corresponds to a fact F iff Γ_{C} is a propositional combination of constituents that each correspond to an atomic fact and F is the same propositional combination of the corresponding atomic facts.*

Consider an interpretation $M = (\Psi, \Sigma)$ of a domain description D in $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$. $\Sigma(s_N)$ can be written as a (possibly empty) sequence of compound actions (here $last \in \mathtt{N}_0$):

$$\Sigma(s_N) = [a_0, a_1, \ldots, a_{last}]$$

Similarly, we can write for all $s \in \mathcal{S}$: $\Sigma(s) = [a_0, \ldots, a_t]$ with $t \leq last$. For $t > last$ we define a_t to be the empty set. We define $[a_{-1}]$ to be equal to the empty sequence []. Using this convention, for $0 \leq t \leq last + 1$, we define σ_t as an abbreviation for $\Psi([a_0, \ldots, a_{t-1}])$. For $t > last + 1$, σ_t is defined to be equal to σ_{last+1}.

Definition 8.6 *For any T_C for a language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F}, \mathbf{T})$ and any non-ambiguous domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$ such that $T_C \sim D$, let $M = (\Psi, \Sigma)$ be any interpretation of $(\mathcal{F}, \mathcal{A}, \mathcal{S})$ and $\mathcal{M}$ be any interpretation of our language. We say that M corresponds to $\mathcal{M}$ (written as '$M \cong \mathcal{M}$') iff a) $\Psi(\Sigma(s_N))$ is defined; b) $\mathcal{M}$ interprets all time points as integers and '+' and '<' accordingly; and c) for all $f \in \mathcal{F}, a \in \mathcal{A}$ and $s \in \mathcal{S}$ and all $t \in \mathtt{N}_0$:*

1. $a \in a_t \Leftrightarrow \mathcal{M} \models Ho(a, t)$
2. $f \in \sigma_t \Leftrightarrow \mathcal{M} \models Ho(f, t)$
3. $\Sigma(s)$ *contains exactly* t *elements* $\Leftrightarrow \mathcal{M} \models s = t$

that the axioms:

$$\begin{gathered} s_0 = 0 \\ Ho(Lift\_left, t) \supset Do(Wet, \text{TRUE}, t+1) \\ Ho(Lift\_right, t) \supset Do(Wet, \text{TRUE}, t+1) \\ Ho(Turn\_on\_sprinkler, t) \supset Do(Wet, \text{TRUE}, t+1) \\ (s_0 < s_N) \wedge \neg Ho(Wet, s_0) \wedge \\ [\ Ho(Turn\_on\_sprinkler, s_0) \wedge Ho(Lift\_left, s_0) \wedge Ho(Lift\_right, s_0)\ ] \end{gathered} \tag{21}$$

Here the first axiom is introduced by item 2 of the definition; the second, third and fourth axioms are introduced by items 3 and 4 of the definition and the fifth axiom is introduced by item 8 of the definition. Item 6 of definition 8.5 deserves special attention: it translates any rule in D of the form '$[a_1, \ldots, a_n]$ **occurs_at** s' into an axiom in CONS which expresses that a_1 should hold at time point s, a_2 at time point $s + 1$ etc. In other words, if, in BG's formalism, several actions $[a_1, \ldots, a_n]$ take place sequentially in a situation s, we interpret this as saying, in our formalism, that the first of them happens at the point in time t corresponding to s, the second to the point in time directly thereafter etc. Any situation s' such that s precedes s' in BG's formalism will therefore be mapped to a point in time t' that is sufficiently larger than t so as to allow for all the actions $[a_1, \ldots, a_n]$ to happen in between t and t'.

The semantical correspondence relation '$\cong$' is introduced in definition 8.6 on page 30. Just above that definition, we introduce the notation a_t to stand for the $(t+1)$-th compound action taking place in an interpretation M. Similarly, σ_t stands for the state of the world (i.e. the set of all fluents that hold) just before the $(t + 1)$-th compound action takes place. In the definition itself, it is checked whether for any set of actions denoted by a_t, the same actions take place in $\mathcal{M}$ *at time* t; σ_t is treated similarly.

Having discussed both syntactic and semantical correspondence, we are now ready to state our theorem. The proof can be found in appendix 4.3.

Theorem 8.2 *For any theory T_C for a language $\mathcal{L}(\mathbf{D},\mathbf{E},\mathbf{F},\mathbf{T})$ and any domain description D for a language $\mathcal{L}_3$ of Baral and Gelfond's such that a) $T_C \sim D$ and b) D is not* ambiguous*, we have:*

$$M = (\Psi,\Sigma) \text{ is a model for } D \text{ such that } \Psi(\Sigma(s_N)) \text{ is defined} \Rightarrow$$
$$\text{there exists an } \mathcal{M} \text{ with } M \cong \mathcal{M} \text{ such that } \mathcal{M} \text{ is a preferred model of } T_C$$

and we further have

$$\mathcal{M} \text{ is a preferred model of } T_C \Rightarrow \text{there exists a model } M \text{ for } D \text{ such that } M \cong \mathcal{M}$$

9. A Brief Note on Other Approaches based on Pearl's ideas

There have been several earlier papers on common-sense reasoning about action that were (partially) based on Pearl's ideas; we mention the papers of Darwiche and Pearl [7, 8], Boutilier and Goldszmidt [6] and Geffner [9]. However, in the works of Darwiche, Pearl, Boutilier and Goldszmidt actions are not directly treated as interventions, but rather compiled into nodes in a causal graph (in Darwiche and Pearl's case) or a Bayesian Network (in Boutilier and Goldszmidt's case). The resulting theories then are quite different from ours in that they lack the *Do*-operator, which is fundamental to our solution of the ramification problem. Only the approach of Geffner is able to handle comparable instances of the ramification problem. We plan a more in-depth comparison to Geffner's approach in future work.

Earlier, the present author has introduced the two model selection criteria $\mathbf{S}_0$ and $\mathbf{I}_0$ [14, 13] which were claimed to be based on Pearl's causal graphs. However, the connection was not investigated in detail. The present report can be seen as an extension of this original work, but now with the connection worked out in full detail (see appendix 1). Putting the theory in a form that makes as explicit as possible the connection to Pearl's work has caused the causal theories discussed in this paper to look – superficially – quite different from $\mathbf{S}_0$ and $\mathbf{I}_0$.

10. Conclusion

We have shown that by extending Pearl's causal theories we arrive at a powerful approach to common sense reasoning about action and change. We had to extend Pearl's theory at several places; 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 (in the propositional case, these are elements of $\mathbf{V}$; in the first-order case, instances of *Ho*) and interventions (instances of *Do*). 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 [5, 11, 23, 38]). It is this freedom that makes the representational power of our formalism go beyond most existing ones.

In future work, we would like to make a more extensive comparison between our approach and that of Thielscher [38]. Apart from Pearl, Thielscher is one of the few authors in the field who tries to make completely clear what he means by a causal law, i.e. he attempts to give a semantics to statements of the form 'A causes B' in non-causal terms. Interestingly, Thielscher's is also the only approach we are aware of that can correctly handle reasoning domains for which our approach fails ([38], example 18). On the other hand, our use of *Do* is much less restrictive than Thielscher's use of `causes`. We plan to study the exact differences between Thielscher's and our approach in the near future.

Another approach that deserves further attention is the one based on Sandewall's and Doherty's *occlusion* concept [34, 33, 15]. Especially the way it is used by Gustafsson and Doherty [15] is strongly reminiscent of Lin's use of the *Caused*-predicate and thus also related to our use of *Do*. Even more interestingly, the underlying idea behind occlusion is similar to Pearl's ideas about interventions: in Sandewall, Gustafsson and Doherty's work, if a fluent is *affected* by some action, then it becomes *occluded* for the duration of the action, which means that its value becomes independent of the value

it had directly before the action started. Compare this to the sufficient cause principle: if an action takes place that *sets* the value of some fluent, then the fluent becomes independent of the values of any variables in the domain which normally influence it. We see that the two concepts are closer than their names suggest and we think that they should be compared in detail.

This brings us once again to what may be the most interesting aspect of our approach: 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.

Acknowledgements

The author would like to thank Paul Vitányi for providing some extra time to work on the subject of this paper; Bruno Gaume, for sparking off his interest in the subject; and John-Jules Meijer, for providing lots of encouragement.

1. Appendix: From Pearl's Causal Theories to Ours

In this section we show how exactly to extend Pearl's causal theories in order to arrive at ours. We use the version of Pearl's theories introduced in [31, 32]. We first give Pearl's original definitions. We then show how a set of global constraints CONS is introduced; this gives us a new kind of causal theories that, though already more general than Pearl's, are nevertheless still essentially propositional. In subsection 1.1 we show that they are equivalent to the propositional causal theories used in the main text. In subsection 1.2 we show how to extend them further to first-order causal theories. In subsection 1.3 we show how the notion of 'preferred model' is already implicit in causal theories.

1.1 Propositional Causal Theories

Here is the definition of causal theories as given by Pearl [31, 32]:

Definition 1.1 *A causal theory T is a 4-tuple $T = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \{f_i\} \rangle$ where*

1. $\mathbf{V} = \{X_1, \ldots, X_n\}$ *is a set of observed variables*
2. $\mathbf{U} = \{U_1, \ldots, U_m\}$ *is a set of unobserved variables which represent disturbances, abnormalities or assumptions.*
3. $P(\mathbf{u})$ *is a distribution over $U_1, \ldots, U_m$, and*
4. $\{f_i\}$ *is a set of n deterministic functions, each of the form*

$$X_i = f_i(X_1, \ldots, X_n, U_1, \ldots, U_m) \tag{1}$$

Usually, causal theories come together with one or more *actions* of the form $Do(X_i, x)$. In the following, we assume that $\{X_{i_1}, \ldots, X_{i_l}\}$ is an arbitrary subset of the variables $\mathbf{V}$ and that for all $1 \leq k \leq l$, x_k is a value in the domain of X_{i_k}. Here is how Pearl defines the effect of actions [31, 32]:

Definition 1.2 *(Effect of actions) The effect of the set of actions*

$$\mathbf{A} = \{Do(X_{i_1}, x_1), Do(X_{i_2}, x_2), \ldots, Do(X_{i_l}, x_l)\}$$

on a causal theory T is given by a subtheory $T(\mathbf{A})$ of T, where $T(\mathbf{A})$ obtains by deleting from T all equations corresponding to the X_{i_k} occurring in $\mathbf{A}$ and substituting the equations $X_{i_k} = x_k$ instead.

The definition of *models* for causal theories is straightforward:

Definition 1.3 *A valuation $\mathcal{M}$ for the variables in $\mathbf{V} \cup \mathbf{U}$ belonging to a causal theory T is called a* model *of T iff all of the equations (1) associated with T hold in $\mathcal{M}$. A valuation $\mathcal{M}$ is called* a model for the causal theory T and the set of actions $\mathbf{A}$ *iff $\mathcal{M}$ is a model for $T(\mathbf{A})$, where $T(\mathbf{A})$ is defined as in def. 1.2.*

It is often assumed [32] that the set of equations (1) has a unique solution for $X_i, \ldots, X_n$, given any value of the disturbances $U_1, \ldots, U_m$; in other words, each set of values for the disturbances determines a unique model for T. Therefore the distribution $P(\mathbf{u})$ induces a unique distribution on the set of variables $\mathbf{U} \cup \mathbf{V}$, or, equivalently, on the set of *models*: a model $\mathcal{M}$ such that $U_1 = u_1, \ldots, U_m = u_m$ will receive probability $P\{\mathcal{M}\} = P\{U_1 = u_1, \ldots, U_m = u_m\}$. We can define the notion of *preferred* model in terms of its probability:

Definition 1.4 *For any causal theory T, any valuation $\mathcal{M}$ of the variables in $\mathbf{V} \cup \mathbf{U}$ with maximum probability $P\{\mathcal{M}\}$ will be called a* preferred model *for the causal theory.*

We will use 'preferred models' only in section 1.3. For now we just look at plain models for causal theories, i.e. we do not care about the distribution $P(\mathbf{u})$. In that case, the assumption that the set of

equations (1) has a unique solution for $X_i, \ldots, X_n$, given any value of the disturbances $U_1, \ldots, U_m$ is not needed.

We refer to causal theories as defined above as *Pearlian* or *basic* causal theories. Causal theories which are such that all the variables in $\mathbf{V}$ and $\mathbf{U}$ are propositional will be called *propositional*. For simplicity, we will focus on these propositional theories when defining our extension. Here it is:

Definition 1.5 *An extended propositional causal theory T_{EP} is a 5-tuple*

$$T_{EP} = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \{f_i\}, \text{CONS} \rangle$$

such that

1. $\langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \{f_i\} \rangle$ *is a propositional basic causal theory as defined in def. 1.1.*
2. CONS*, the set of* constraints*, is a finite set of propositional formulas over variables $\mathbf{V} \cup \mathbf{U} \cup \mathcal{A}(\mathbf{V})$. Here $\mathcal{A}(\mathbf{V})$ is defined as the set of propositional variables*

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

For any extended propositional causal theory $T_{EP} = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \{f_i\}, \text{CONS} \rangle$, the associated basic causal theory will be called T^*_{EP}. If all the constraints in CONS hold for a valuation $\mathcal{M}$ of the variables in $\mathbf{V} \cup \mathbf{U} \cup \mathcal{A}(\mathbf{V})$ we will write $\mathcal{M} \models \text{CONS}$. Notice that a T_{EP} is just a causal theory together with a set of formulas that may directly involve interventions. For such theories, we also need to extend the definition of models. The idea here is that we want to make sure that the axioms in CONS hold for all models of causal theories while still, the sufficient cause principle dictates how interventions should be handled.

Definition 1.6 *A model for an extended causal theory $T_{EP} = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \{f_i\}, \text{CONS} \rangle$ 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 of the basic causal theory $T^*_{EP}(\mathbf{A})$. Here $T^*_{EP}(\mathbf{A})$ is the effect of the set of actions $\mathbf{A}$ on the basic causal theory T^*_{EP}, where*

$$\mathbf{A} = \{Do(X_i, x) \mid \mathcal{M} \models Do(X_i, x), X_i \in \mathbf{V}, x \in \mathbf{B}\}$$

A preferred model for T_{EP} is defined as a valuation $\mathcal{M}$ for variables in $\mathbf{V} \cup \mathbf{U} \cup \mathcal{A}(\mathbf{V})$ such that 1) $\mathcal{M} \models \text{CONS}$ and 2) the restriction of $\mathcal{M}$ to $\mathbf{V} \cup \mathbf{U}$ is a preferred model of $T^*_{EP}(\mathbf{A})$.

Note that the set $\mathbf{A}$ in item 2 of this definition depends on the model $\mathcal{M}$, i.e. it can be different for different $\mathcal{M}$. Now for any extended causal theory $T_{EP} = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \{f_i\}, \text{CONS} \rangle$, let T'_{EP} be the theory that results from deleting from T_{EP} all functions f_i in $\{f_i\}$ of the form $X_i = U$ where U is some element of $\mathbf{U}$. It is easy to see the following:

Proposition 1.1 *$\mathcal{M}$ is a model for T_{EP} iff $\mathcal{M}$ is a model for T'_{EP}.*

Hence as long as we work with extended propositional theories, we do not have to worry about specifying functions f_i for *all* the $X_i \in X$ – if an f_i is left out, it just means that we are indifferent about the corresponding X_i. But now notice that, since we are working with propositional variables, we can replace the equality sign '=' in the structural equations (1) by logical equivalence '$\equiv$' without changing their meaning. From this, together with the proposition (1.1) above and the definition of effects (def. 1.2) it immediately follows that the definitions of extended propositional causal theories T_{EP} and their models (definitions 1.5 and 1.6, resp.) are completely equivalent to the definitions of causal theories and models given in the main text (definitions 3.1 and 3.2, resp.). This shows how the definitions in the main text arise as natural extensions of Pearl's definitions.

1.2 From propositional to first-order causal theories
A potential difficulty with propositional causal theories $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS} \rangle$ is the fact that their models are defined by performing syntactic operations on T - some of the equations in EQ get replaced, depending on the axioms in CONS. It would be convenient if we could express these syntactic operations as axioms *within* EQ itself. A model for a causal theory for which EQ is extended with such axioms would then simply be a valuation $\mathcal{M}$ with $\mathcal{M} \models \text{EQ} \cup \text{CONS}$, where '$\models$' denotes ordinary propositional entailment. It turns out that EQ can indeed be extended with such axioms, and moreover, we find that this is exactly what we need to extend our theories to the first-order case.

A Reformulation of Propositional Causal Theories We first give a reformulation of propositional causal theories that is equivalent with the original definition but in which the replacement of axioms is avoided in the way indicated above:

For any causal theory $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS} \rangle$ defined according to definition 3.1, let T' be the propositional theory over variables $\mathbf{U} \cup \mathbf{V} \cup \mathcal{A}(V)$ such that $T' = \text{CONS} \cup \text{EQ}_{\text{new}}$. Here EQ_{new} results from EQ by replacing any structural equation $X_i \equiv \Phi$ in EQ by the following three axioms:

$$[\ \neg Do(X_i, \text{TRUE}) \wedge \neg Do(X_i, \text{FALSE})\ ] \supset (X_i \equiv \Phi) \tag{2}$$

$$Do(X_i, \text{TRUE}) \supset X_i \tag{3}$$

$$Do(X_i, \text{FALSE}) \supset \neg X_i \tag{4}$$

Here the first axiom represents the original structural equation: if no interventions take place, then X_i should still be equivalent to Φ. The second and third axioms represent the effect of actions.

Theorem 1.1 *For any causal theory T and propositional theory T' obtained from T as described above, we have*

$$\mathcal{M} \text{ is a model for causal theory } T \Leftrightarrow \mathcal{M} \models T'$$

Proof: ($\Rightarrow$) Suppose $\mathcal{M}$ is a model for causal theory T. Then, according to definition 3.2, $\mathcal{M} \models \text{CONS}$. It remains to be shown that $\mathcal{M} \models \text{EQ}_{\text{new}}$. For this, let first ϕ be any axiom in EQ_{new} of form (2). Notice that EQ must contain the structural equation $S = `X_i \equiv \Phi'$. Either $\mathcal{M} \models \neg Do(X_i, \text{TRUE}) \wedge \neg Do(X_i, \text{FALSE})$ or not. In the latter case $\mathcal{M} \models \phi$ trivially. In the former case, EQ$'$ in definition 3.2 must clearly contain structural equation S to. Since $\mathcal{M} \models \text{EQ}'$, we have $\mathcal{M} \models S$ and therefore $\mathcal{M} \models \phi$ and we are done.

Now let ϕ be any axiom in EQ_{new} of form (3). If $\mathcal{M} \not\models Do(X_i, \text{TRUE})$ we are done, so let us suppose $\mathcal{M} \models Do(X_i, \text{TRUE})$. In this case, EQ$'$ in definition 3.2 contains the structural equation $S = `X_i \equiv \text{TRUE}'$. Since $\mathcal{M} \models \text{EQ}'$, we have $\mathcal{M} \models S$ and therefore $\mathcal{M} \models \phi$ and we are done.

In the same way one can show that $\mathcal{M} \models \phi$ for any ϕ in EQ_{new} of form (4).

($\Leftarrow$) Suppose $\mathcal{M} \models T'$. Then $\mathcal{M} \models \text{CONS}$ so condition 1 of definition 3.2 is satisfied. It remains to be shown that $\mathcal{M} \models \text{EQ}'$ where EQ$'$ is obtained form EQ as in condition 2 of definition 3.2.

For this, note there are two kinds of axioms in EQ$'$: those of form $X_i \equiv b$ for some $X_i \in \mathbf{V}$ and $b \in \mathbf{B}$ and those of form $X_i = \Phi$ for some $X_i \in \mathbf{V}$. We will show that for any ϕ in EQ$'$ of either of the two kinds, $\mathcal{M} \models \phi$. If ϕ is an axiom of the first kind, then, by definition 3.2, we must have $\mathcal{M} \models Do(X_i, b)$. But then, since $\mathcal{M} \models \text{EQ}_{\text{new}}$ and EQ_{new} contains (the proper instances of) axioms (3) and (4), we also have $\mathcal{M} \models (X_i \equiv b)$ and hence also $\mathcal{M} \models \phi$.

If ϕ is of the second kind, then, by definition 3.2, we must have $\mathcal{M} \models \neg Do(X_i, \text{TRUE}) \wedge \neg Do(X_i, \text{FALSE})$. But then, since $\mathcal{M} \models \text{EQ}_{\text{new}}$ and EQ_{new} contains (the proper instance of) axiom (2), we have $\mathcal{M} \models \phi$ too. □

Quantifying over Structural Equations Let us suppose for the moment that we want to use causal theories for persistence involving $n+1$ points in time. Theorem 1.1 shows that in this case, all axioms

in EQ can equivalently be represented by a set EQ_{new} which contains:

$$\begin{array}{lcl} \neg Do(F_i(1), \text{TRUE}) \wedge \neg Do(F_i(1), \text{FALSE}) & \supset & (F_i(1) \equiv F_i(0)) \\ \neg Do(F_i(2), \text{TRUE}) \wedge \neg Do(F_i(2), \text{FALSE}) & \supset & (F_i(2) \equiv F_i(1)) \\ & \vdots & \\ \neg Do(F_i(n), \text{TRUE}) \wedge \neg Do(F_i(n), \text{FALSE}) & \supset & (F_i(n) \equiv F_i(n-1)) \end{array} \tag{5}$$

for all $F_i \in \mathbf{F}$, with corresponding axioms

$$Do(F_i(t), b) \supset (F_i(t) \equiv b)$$

for all $F_i \in \mathbf{F}, b \in \mathbf{B}$ and $t \in \{0, 1, \ldots, n\}$.

It is now evident that this scheme of axioms can be extended to a countably infinite number of time points by universally quantifying over both groups of axioms and letting $F_i(t)$ denote a predicate involving object t rather than an atomic propositional variable. We thus end up with two groups of axioms. First,

$$\forall t > 0 \; . \; \neg Do(F_i(t), \text{TRUE}) \wedge \neg Do(F_i(t), \text{FALSE}) \supset (F_i(t) \equiv F_i(t-1)) \tag{6}$$

for all $F_i \in \mathbf{F}$ and second

$$\begin{array}{lcl} \forall t \; . \; Do(F_i(t), \text{TRUE}) & \supset & (F_i(t) \equiv \text{TRUE}) \\ \forall t \; . \; Do(F_i(t), \text{FALSE}) & \supset & (F_i(t) \equiv \text{FALSE}) \end{array} \tag{7}$$

Axioms of form (6) will be called *structural equation axioms*. Notice that they really contain an infinitude of structural equations! Axioms of form (7) will be called *intervention axioms*.

In theories for temporal reasoning domains it will be useful not only to quantify over time points but also over fluents. This can easily be accomplished by treating F_i as an object rather than a predicate constant, introducing a new predicate *Ho* and writing $Ho(F_i, t)$ in stead of $F_i(t)$, a usual practice in common-sense temporal reasoning.

1.3 Preferred Models of Causal Theories

For simplicity, we consider extended *propositional* causal theories again as defined in appendix 1.1, definitions 1.5 and 1.6. For the moment we assume that the set of equations (1) has a unique solution for $X_i, \ldots, X_n$, given any value of the disturbances $U_1, \ldots, U_m$; in other words, given a causal theory T, each set of values for the disturbances determines a unique model for T. Taking a probabilistic stance towards reasoning under uncertainty (as, for example, in [12]) we assume that the distribution $P(\mathbf{u})$ represents which external influences we consider 'abnormal' and which external influences we are completely ignorant about. Having specified such a distribution $P(\mathbf{u})$, we always prefer the models which are the most probable according to $P(\mathbf{u})$, in accordance with definition 1.4. We get a particularly simple case if we assume that $\mathbf{U}$ can be partitioned into two subsets: variables standing for unlikely or abnormal occurrences (having small probability) and variables standing for external influences we are ignorant about (having probability 1/2). We call $P(\mathbf{u})$'s for which this holds *simple*:

Definition 1.7 *For any extended propositional causal theory T_{EP}, we call the associated $P(\mathbf{u})$* simple *if $\mathbf{U}$ and $P(\mathbf{u})$ are such that:*

- *$\mathbf{U}$ can be partitioned as follows: $\mathbf{U} = \mathbf{Ind} \cup \mathbf{Ab}$*
- *$P(\mathbf{u})$ is defined as follows:*
  - *All variables in $\mathbf{U}$ are independent: for all $U_1, U_2 \in \mathbf{U}$, $P(U_1 \wedge U_2) = P(U_1) \cdot P(U_2)$.*

- *for all* $Ind \in \mathbf{Ind}$, $P(Ind = \text{TRUE}) = P(Ind = \text{FALSE}) = 1/2$.
- *for all* $Ab \in \mathbf{Ab}$, $P(Ab = \text{TRUE}) = \epsilon_{Ab}$. *Here* $0 < \epsilon_{Ab} < 1/2$.

Now let the sets $\mathbf{V}$, $\mathbf{U}$, $\{f_i\}$ and CONS be given and let $\mathcal{T}(\mathbf{V}, \mathbf{U}, \{f_i\}, \text{CONS})$ be the class of extended propositional causal theories $\langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \{f_i\}, \text{CONS} \rangle$ for which $P(\mathbf{u})$ is simple. Hence the class $\mathcal{T}(\mathbf{V}, \mathbf{U}, \{f_i\}, \text{CONS})$ consists of causal theories T_{EP} which all share the same *models* but not the same *preferred models*. We say that a model $\mathcal{M}$ is *potentially preferred* iff there is a theory in $\mathcal{T}(\mathbf{V}, \mathbf{U}, \{f_i\}, \text{CONS})$ for which it is a preferred model. In other words, there exists a simple $P(\mathbf{u})$ such that $\mathcal{M}$ is among the most probable models for $T = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \{f_i\}, \text{CONS} \rangle$. If we are given $\mathbf{V}, \mathbf{U}, \{f_i\}$ and CONS but the only thing we know about $P(\mathbf{u})$ is that it is simple, it seems reasonable to consider *any* model that is potentially preferred. The proposition below shows that the 'minimal models' in the usual sense of nonmonotonic logic coincide exactly with the models that are 'potentially preferred'.

Proposition 1.2 *$\mathcal{M}$ is potentially a preferred model for the class of theories $\mathcal{T}^* = \mathcal{T}(\mathbf{V}, \mathbf{U}, \{f_i\}, \text{CONS})$ iff there is no model $\mathcal{M}'$ for the theories in $\mathcal{T}^*$ with*

$$\{Ab \mid Ab \in \mathbf{Ab}, \mathcal{M}' \models Ab\} \subset \{Ab \mid Ab \in \mathbf{Ab}, \mathcal{M} \models Ab\} \tag{8}$$

Proof: (only-if) Suppose $\mathcal{M}$ is a potentially preferred model for $\mathcal{T}^*$. So $\mathcal{M}$ is preferred for some $T \in \mathcal{T}^*$. Let $P_T(\mathbf{u})$ be the distribution associated with this T. Remember that we identified the probability of a single model under distribution $P(\mathbf{u})$ with the joint probability of all $U \in \mathbf{U}$ (cf. the remark previous to definition 1.4 on page 33).

Now let us assume, by way of contradiction, that there is a model $\mathcal{M}'$ for one of the theories T' in $\mathcal{T}^*$ such that $\mathcal{M}'$ has property (8). Since $\mathcal{M}'$ is a model for T', it is also a model for T (since T' and T share the same $\mathbf{V}, \mathbf{U}, \{f_i\}$ and CONS). We will show that $\mathcal{M}'$ has higher probability under $P_T(\mathbf{u})$ than $\mathcal{M}$. This implies that $\mathcal{M}$ is *not* a preferred model of T and hence we arrive at the desired contradiction.

We show that indeed $P\{M'\} > P\{\mathcal{M}\}$. First, we define $\mathcal{M}[U_i]$ to be the $u_i \in \mathbf{B}$ for which $\mathcal{M} \models U_i \equiv u_i$. We also define for any $Ab \in \mathbf{Ab}$: $P\{Ab\} := P\{Ab \equiv \text{TRUE}\}$. Note that since $P(\mathbf{u})$ is simple, we have by definition 1.7:

$$P(\mathcal{M}') = \prod_{U \in \mathbf{U}} P\{U_i = \mathcal{M}'[U_i]\} = P\{\mathcal{M}\} \cdot \prod_{Ab \in \mathbf{D}} \frac{1 - P\{Ab\}}{P\{Ab\}} = P\{\mathcal{M}\} \cdot \prod_{Ab \in \mathbf{D}} \frac{1 - \epsilon_{Ab}}{\epsilon_{Ab}}$$

where $\mathbf{D}$ is the non-empty set $\{Ab \in \mathbf{Ab} \mid \mathcal{M} \models Ab \text{ and } \mathcal{M}' \not\models Ab\}$. Since all $\epsilon_{Ab} < 1/2$, this shows that $P\{\mathcal{M}'\} > P\{\mathcal{M}\}$.

(if) In this case, let T be a theory $\langle \mathbf{V}, \mathbf{U}, \{f_i\}, P(\mathbf{u}), \text{CONS} \rangle$ with $P(\mathbf{u})$ such that for all $Ab \in \mathbf{Ab}$ with $\mathcal{M} \models Ab$, $P(Ab) = 1/2 - \epsilon$ where $\epsilon > 0$ is some small number, while $P(Ab') = \epsilon$ for all $Ab' \in \mathbf{Ab}$ with $\mathcal{M} \not\models Ab'$. Clearly, $P(\mathbf{u})$ is simple so indeed $T \in \mathcal{T}^*$. We are assuming that there is no model $\mathcal{M}'$ for which (8) holds. It is easy to see that this implies that for ϵ small enough, $\mathcal{M}$ must be the model for T that has the highest probability. □

We see that the probabilistic view on uncertainty taken in Pearl's original definition can be reconciled with a minimal subset-minimization policy as is adopted in the main text, as long as we interpret the propositional variables $Ab \in \mathbf{Ab}$ as things that occur with small ($< 1/2$) but, apart from that, unknown probabilities. We can further extend the concept of 'potentially preferred models' to coincide with the first-order 'minimal models' as defined in section 7.4. We will not do this in full detail; in stead, we will only roughly indicate how such an extension may be achieved:

First, we may extend our concept of 'potentially preferred models' to several 'levels' of abnormality by further partitioning $\mathbf{U}$. For example, we may set $\mathbf{U} = \mathbf{Ind} \cup \mathbf{Ab}_1 \cup \mathbf{Ab}_2$ where $\mathbf{Ab}_1$ contains propositions Ab_1 with $P(Ab_1) < 1/2$ while $\mathbf{Ab}_2$ contains propositions Ab_2 with $0 < P(Ab_2) <$

$\min_{Ab_1 \in \mathbf{Ab}_1}(P(Ab_1))^k$ where k is the number of abnormality propositions in $\mathbf{Ab}_1$. In this way, models without abnormalities in $\mathbf{Ab}_2$ will always be preferred over models with abnormalities in $\mathbf{Ab}_2$, no matter the number of abnormalities in $\mathbf{Ab}_1$.

We also need to take care of the following: our notion of preferred models depends on the assumption that the set of structural equations (1) has a unique solution for $X_i, \ldots, X_n$, given any instantiation of the variables in $\mathbf{U}$. This need not be the case for (both propositional and first-order) causal theories as defined in the main text. For example, there will be no structural equations that indicate the value of a fluent at time 0: 2-point causal theories do not contain structural equations of the form $F_i(0) \equiv \Phi$, since, in the absence of axioms in CONS, we want both models with $\mathcal{M} \models F_i(0)$ and models with $\mathcal{M} \not\models F_i(0)$. But, in light of proposition 1.1 on page 34 we can always transform such an 'incompletely specified' set of structural equations into a complete one by introducing new assumption symbols! In our example, we can add $F_i(0) = \mathrm{Ind}_i$ for all $F_i \in \mathbf{F}$ to our set of structural equations. In this case, the value of Ind_i determines a unique solution for $F_i(0)$. If we let $P(\mathrm{Ind}_i = \mathrm{TRUE}) = P(\mathrm{Ind}_i = \mathrm{FALSE}) = 1/2$, then neither of the two values for Ind_i will be preferred.

Finally, we note that the notion of 'preferred models for theories with simple $P(\mathbf{u})$' can be extended to the first-order case considered in the main text, simply by replacing the set of abnormality propositions by a single abnormality predicate Ab_1 and letting each instantiation of Ab_1 stand for a separate abnormality.

2. Appendix: The Model-Theoretic Characterization of Circumscription

The following is all taken from [18].

Let T be a first-order theory for some language $\mathcal{L}$; let P be a tuple of predicate constants for $\mathcal{L}$ and let Z be a tuple of function and/or predicate constants for $\mathcal{L}$ disjoint with P. For any two structures $\mathcal{M}_1$ and $\mathcal{M}_2$ for the language L, we write $\mathcal{M}_1 \leq^{P;Z} \mathcal{M}_2$ if

(i) $|\mathcal{M}_1| = |\mathcal{M}_2|$

(ii) $\mathcal{M}_1[\![K]\!] = \mathcal{M}_2[\![K]\!]$ for every constant K not in P, Z

(iii) $\mathcal{M}_1[\![P_i]\!] \subseteq \mathcal{M}_2[\![P_i]\!]$ for every P_i in P.

If $\mathcal{M}_1 \leq^{P;Z} \mathcal{M}_2$ but not $\mathcal{M}_2 \leq^{P;Z} \mathcal{M}_1$ we write $M_1 <^{P;Z} M_2$. We call a structure $\mathcal{M}$ *minimal* in a class M of structures if $\mathcal{M} \in$ M and there is no structure $\mathcal{M}' \in$ M such that $\mathcal{M}' <^{P;Z} M$.

The 'circumscription of P in T with Z varied' which we write as $Circum(T; P; Z)$ is defined as a formula in second-order logic (we will not repeat this formula here, see [26, 18]). However, we may also characterize circumscription as follows:

Proposition 2.1 (Lifschitz 1985) *A structure $\mathcal{M}$ is a model of $Circum(T; P; Z)$ iff $\mathcal{M}$ is minimal in the class of models of T with respect to $<^{P;Z}$.*

3. Appendix: McCain & Turner vs. 2-point causal theories

3.1 Formal Definition of MT's Next-State Function

The definitions in this section have all been copied from [25, 24]. MT start with a propositional language which includes the 0-ary logical connectives TRUE and FALSE. TRUE and ¬FALSE are tautologies in which no atoms occur. Each interpretation is identified with the set of literals true in it. A *causal law* is defined to be an expression of the from $\phi \Rightarrow \psi$ where ϕ and ψ are formulas of the propositional language. A set of causal laws is called a *causal theory*. Now for every causal theory D and interpretation I, we let

$$D^I = \{\psi : \text{for some } \phi,\ \phi \Rightarrow \psi \in D \text{ and } I \models \phi\}\ .$$

That is, D^I is the set of consequents of all causal laws in D whose antecedents are true in I.

Definition 3.1 *Let D be a causal theory and I be an interpretation. We say that I is* causally explained *according to D if I is the unique model of D^I.*

We now define $\Pi_1(S, E, C)$. It is a function on *initial states* S, *explicit effects* E and *background knowledge* C. A state is just an interpretation; an explicit effect is a set of formulas; the background knowledge is a set of causal laws.

Definition 3.2 *For any interpretation S, set of propositional formulas E and set of causal laws C, $\Pi_1(S, E, C)$ is defined to be the set consisting of all states that are causally explained according to the causal theory*

$$\{L \Rightarrow L : L \in S\} \cup \{\text{TRUE} \Rightarrow \phi : \phi \in E\} \cup C$$

We will repeatedly use the following lemma which gives a precise characterization of $\Pi_1(S, E, C)$. It was first presented by McCain & Turner [25] but without proof. The proof however is not difficult; we provide one below.

Lemma 3.1 (McCain & Turner) *For any interpretation S, set of propositional formulas E and set of causal laws C, a state S' belongs to $\Pi_1(S, E, C)$ if and only if*

- $S' \models E \cup C^{S'}$
- $S' \setminus S \subseteq \{\phi : \phi$ *is a literal and* $(S \cap S') \cup E \cup C^{S'} \models \phi\}$

Proof: To prove the lemma, we first need the following observation (the proof of which is trivial and omitted):

Claim For any two causal theories A and B and any interpretation I and any propositional formula ψ, we have $\psi \in (A \cup B)^I \Leftrightarrow \psi \in A^I \cup B^I$.

From repeatedly applying the claim and inspection of definition 3.2 it follows that it is sufficient to prove the following:

Restated Lemma S' is the unique model of $X = (S \cap S') \cup E \cup C^{S'}$ iff

C1. $S' \models E \cup C^{S'}$ and

C2. $S' \setminus S \subseteq \{\phi : \phi$ is a literal and $(S \cap S') \cup E \cup C^{S'} \models \phi\}$

We first show the only-if direction, so let us assume that S' is the unique model of X. Then **C1** holds trivially. To see that **C2** holds, notice that S' is the *unique* model of X, so for all literals ϕ in S', $X \models \phi$. Hence $S' = \{\phi : \phi$ is a literal and $X \models \phi\}$ from which **C2** follows.

Now for the if-direction. For this, suppose S' is *not* the unique model of X. Then either S' is not a model of X at all or there is more than one model for X. In the first case, note that *always* $S' \models S \cap S'$, so if $S' \models E \cup C^{S'}$ then also $S' \models (S \cap S') \cup E \cup C^{S'}$ and thus $S' \models X$. From this it follows that if $S' \not\models X$, then also $S' \not\models E \cup C^{S'}$ so **C1** does not hold and we are done.

In the second case (i.e. there is more than one model for X), note that there must be a literal ϕ such that neither ϕ nor $\neg\phi$ is contained in X. This implies $\phi \notin S \cap S'$ and $\neg\phi \notin S \cap S'$, it follows that either ($S \models \neg\phi$ and $S' \models \phi$) or ($S \models \phi$ and $S' \models \neg\phi$). In the first case, $\phi \in S' \setminus S$ while by definition of ϕ, $\phi \notin X$. This means that **C2** does not hold which is what we had to prove. In the second case, one can prove that **C2** is violated in an analogous way. □

3.2 Proof of proposition 8.1

Take S, E and C as in the statement of the proposition. Let S' be any member of $\Pi_1(S, E, C)$. By lemma 3.1 in appendix 3.1 it follows that $S' \models E$ and hence $X \in S'$. This means $C^{S'}$ must contain $Y \vee Z$. Applying lemma 3.1 again, we find $S' \models Y \vee Z$. So there are three possibilities left for S': $S' \in \{S_1, S_2, S_3\}$ where $S_1 = \{X, Y, Z\}$, $S_2 = \{X, Y, \neg Z\}$, $S_1 = \{X, \neg Y, Z\}$. Now the second part of lemma 3.1 does not hold for $S' = S_1$ while it does hold for $S' = S_2$ and $S' = S_3$. The first part of the lemma also holds for $S' = S_2$ and $S' = S_3$, so $\Pi_1(S, E, C) = \{S_2, S_3\}$.

3.3 Proof of theorem 8.1
Our proof makes repeated use of lemma 3.1 of appendix 3.1. We first prove the first part of theorem 8.1, which is restated below. We use the notation introduced in def. 8.1, i.e. we will sometimes write F_i^{TRUE} to denote F_i and F_i^{FALSE} to denote $\neg F_i$.

Theorem 3.1 (part I) *For any theory T_{C} and any domain description $T_{\mathrm{MT}} = (S, E, C)$ such that $T_{\mathrm{C}} \sim T_{\mathrm{MT}}$, we have $S' \in \Pi_1(S, E, C) \Rightarrow$ there exists an $\mathcal{M}$ with $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$ and $\mathcal{M} \cong (S, S')$*

Proof: For any given $S, E, C, S' \in \Pi_1(S, E, C)$ and T_{C} we will now first construct a model $\mathcal{M}$ with $\mathcal{M} \cong (S, S')$. We then show $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$.

Construction of $\mathcal{M}$ We construct a causal model $\mathcal{M}$ according to the following definition, in which we identify sets containing conjunctions of literals with the sets containing just the constituting literals:

1. The initial state of $\mathcal{M}$ is S; the final state of $\mathcal{M}$ is S'. In other words: $F_i \in S \Leftrightarrow \mathcal{M} \models F_i(0)$; $F_i \in S' \Leftrightarrow \mathcal{M} \models F_i(1)$.
2. $F_i^b \in E \cup C^{S'} \Leftrightarrow \mathcal{M} \models Do(F_i(1), b)$.

It is clear that a model $\mathcal{M}$ satisfying all the conditions above can indeed be constructed. It follows immediately from the definition of '$\cong$' (def. 8.2) and item 1 in the construction of $\mathcal{M}$ that $\mathcal{M} \cong (S, S')$. We now prove that $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$. We show this in three stages: in stage 1, we show $\mathcal{M} \models$ CONS. We use this to show in stage 2 that $\mathcal{M}$ is a minimal model for CONS of $\mathcal{A}(\mathbf{V})$ within context $\mathcal{M}\rfloor_{\mathbf{V}\cup\mathbf{U}}$. This shows that $\mathcal{M}$ satisfies condition (1) of definition 4.3. In stage 3, we show that $\mathcal{M}$ also satisfies condition (2) of that definition.

Stage 1 We have to show that $\mathcal{M} \models \phi$ for all axioms ϕ contained in CONS. From def. 8.1, item 2, it follows that it is sufficient to show that $\mathcal{M} \models \phi$ for all sentences in CONS that correspond to either S or to the sentences in E and C. By items 3 and 4 of the same definition and the construction of $\mathcal{M}$ above, it follows immediately that indeed $\mathcal{M} \models \phi$ for the sentences ϕ that correspond to S and E.

Concerning C, let Γ be any sentence in CONS of the form (5). We will show $\mathcal{M} \models \Gamma$. Let us write $\Gamma \equiv \Phi \supset \Psi$ with Ψ of the form $Do(F_{i_1}(1), b_1) \wedge \ldots \wedge Do(F_{i_m}(1), b_m)$. If $\mathcal{M} \not\models \Phi$ we are done, so let us suppose $\mathcal{M} \models \Phi$. By construction of $\mathcal{M}$, we have $S' \models \Phi'$ where Φ' is obtained from Φ by replacing all occurrences of $F_i(1)$ in Φ by F_i. By item 5 of def. 8.1, C must contain a rule of form $\Phi' \Rightarrow F_{i_1}^{b_1} \wedge \ldots \wedge F_{i_m}^{b_m}$ and, since $S' \models \Phi'$, we conclude that $F_{i_j}^{b_j} \in C^{S'}$ for all $1 \leq j \leq m$. Hence by item 2 of the construction of $\mathcal{M}$, $\mathcal{M} \models \Psi$ and hence $\mathcal{M} \models \Gamma$.

Stage 2 We have to show that $\mathcal{M}$ is a minimal element for CONS of $\mathcal{A}(\mathbf{V})$ within the set

$$\mathbf{M}\ (\mathcal{M}) = \{\mathcal{M}' \mid \mathcal{M}' \text{ and } \mathcal{M} \text{ have the same interpretation of } \mathbf{V} \cup \mathbf{U}\}$$

For this, let $\mathcal{M}' \in \mathbf{M}\ (\mathcal{M})$ be a model for CONS. Suppose, by way of contradiction, that $\{\phi \in \mathcal{A}(\mathbf{V}) \mid \mathcal{M}' \models \phi\}$ is a proper subset of $\{\phi \in \mathcal{A}(\mathbf{V}) \mid \mathcal{M} \models \phi\}$. Then there is a pair (F_i, b) such that $\mathcal{M}' \not\models Do(F_i, b)$ while $\mathcal{M} \models Do(F_i, b)$. $\mathcal{M}$ is constructed such that $F_i^b \in E \cup C^{S'}$. Now if $F_i^b \in E$, then, by def. 8.1, item 4, $\mathcal{M}' \not\models$ CONS and we arrive at a contradiction. So suppose $F_i^b \in C^{S'}$. It follows from the definition of $C^{S'}$ that $S' \models \Phi'$ for a Φ' such that C contains some sentence of form $\Phi' \Rightarrow \Psi'$ and $F_i^b \in \Psi'$. Hence by def. 8.1, item 5, CONS contains the sentence $\Phi \supset \Psi$ where Φ, Ψ stand to Φ', Ψ' as required in that item of the definition. Since $S' \models \Phi'$, by the construction of $\mathcal{M}$, we have $\mathcal{M} \models \Phi$. Since $\mathcal{M}'$ and $\mathcal{M}$ share the same interpretations of the variables in $\mathbf{V}$, we also have $\mathcal{M}' \models \Phi$. We assumed $\mathcal{M}' \models$ CONS so also $\mathcal{M}' \models \Psi$ and in particular $\mathcal{M}' \models Do(F_i, b)$. But we assumed the contrary so we have arrived at a contradiction.

Stage 3 We have to show that $\mathcal{M} \models \text{EQ}'$ where EQ′ is the updated set of structural equations referred to in condition (2) of def. 4.3. By construction of $\mathcal{M}$, there are two cases for each F_i: either EQ′ contains $F_i(1) \equiv b$ for some b or EQ′ contains $F_i(1) \equiv F_i(0)$.

The first case happens if $\mathcal{M} \models Do(F_i(1), b)$. By construction of $\mathcal{M}$, it follows that in this case $F_i^b \in E \cup C^{S'}$. By lemma 3.1 and the fact that $S' \in \Pi_1(S, E, C)$ we have $S' \models E \cup C^{S'}$ and thus $S' \models F_i^b$. Then by construction of $\mathcal{M}$, indeed $\mathcal{M} \models F_i(1) \equiv b$ and we are done.

The second case happens if $\mathcal{M} \not\models Do(F_i, b)$ for any $b \in \mathbf{B}$. It follows from the construction of $\mathcal{M}$ that in this case there is no b such that $F_i^b \in E \cup C^{S'}$. Suppose, by way of contradiction, that $\mathcal{M} \not\models F_i(1) \equiv F_i(0)$. Then (by construction of $\mathcal{M}$), $F_i^b \in S' \setminus S$ for some b. By lemma 3.1, $(S \cap S') \cup E \cup C^{S'} \models F_i^b$. Clearly, $F_i^b \notin S \cap S'$. We have already shown that F_i^b cannot be in $E \cup C^{S'}$ either. Since both $S \cap S'$ and $E \cup C^{S'}$ only contain conjunctions of literals, it follows that $(S \cap S') \cup E \cup C^{S'} \not\models F_i^b$ and we have reached a contradiction. □

We now restate and prove the second part of theorem 8.1.

Theorem 3.1 (part II) *If we have a theory* T_{C} *and a* $T_{\text{MT}} = (S, E, C)$ *such that* $T_{\text{C}} \sim T_{\text{MT}}$, *then*

$$\mathcal{M} \models_{\mathbf{c}} T_{\text{C}} \;\Rightarrow\; \Pi_1(S, E, C) \text{ contains an } S' \text{ with } \mathcal{M} \cong (S, S')$$

Proof: We will define S' in terms of a given model $\mathcal{M}$ with $\mathcal{M} \models_{\mathbf{c}} T_{\text{C}}$, as follows: for all F_i: $F_i \in S' \Leftrightarrow \mathcal{M} \models F_i(1)$. We will show that this S' is contained in $\Pi_1(S, E, C)$ and that $\mathcal{M} \cong (S, S')$.

Since $\mathcal{M} \models_{\mathbf{c}} T_{\text{C}}$ and thus $\mathcal{M} \models$ CONS we know by items 2 and 3 of def. 8.1 that for all F_i: $F_i \in S \Leftrightarrow \mathcal{M} \models F_i(0)$. From this and the definition of S' above it follows $\mathcal{M} \cong (S, S')$. It remains to be shown that $S' \in \Pi_1(S, E, C)$. By lemma 3.1 it is sufficient to prove claims 1 and 2 below.

Claim 1 $S' \models E \cup C^{S'}$.
To prove this, we show that for all F_i^b with $F_i^b \in E$ and for all F_i^b with $F_i^b \in C^{S'}$ we have $S' \models F_i^b$. In the first case ($F_i^b \in E$), as $\mathcal{M} \models$ CONS and by def. 8.1, item 4, we have $\mathcal{M} \models Do(F_i(1), b)$. Since $\mathcal{M} \models \text{EQ}'$ this means $\mathcal{M} \models F_i^b(1)$. From the definition of S', we now have $S' \models F_i^b$ and we are done. In the second case, i.e. $F_i^b \in C^{S'}$, C must contain a law of form $\Phi' \Rightarrow \Psi'$ such that $S' \models \Phi'$ and one of the conjuncts in Ψ' is F_i^b. By def. of S' and definition 8.1, item 5, we also have $\mathcal{M} \models \Phi$ where $\Phi \supset \Psi$ is the sentence in CONS that corresponds to $\Phi' \supset \Psi'$. As $\mathcal{M} \models$ CONS, we have $\mathcal{M} \models \Psi$ and in particular $\mathcal{M} \models Do(F_i(1), b)$. As $\mathcal{M} \models \text{EQ}'$, it follows $\mathcal{M} \models F_i^b(1)$ and hence, by def. of S', $S' \models F_i^b$.

Claim 2 $S' \setminus S \subseteq \{\phi : (S \cap S') \cup E \cup C^{S'} \models \phi\}$.
To prove this, we show that for any F_i^b, if $F_i^b \in S' \setminus S$, then we must also have $F_i^b \in (S \cap S') \cup E \cup C^{S'}$. So suppose $F_i^b \in S' \setminus S$. By def. of S', we have for such an F_i^b that $\mathcal{M} \models F_i^b(1) \wedge \neg F_i^b(0)$. Since $\mathcal{M} \models \text{EQ}'$, the structural equation $F_i(1) \equiv F_i(0)$ cannot be in EQ′ and thus there must have been an intervention: we must have $\mathcal{M} \models Do(F_i(1), b)$. Now as $\mathcal{M} \models_{\mathbf{c}} T_{\text{C}}$, there can be no other $\mathcal{M}' \models$ CONS with the same interpretations of $\mathbf{V}$ and $\mathbf{U}$ but with a smaller (in the subset sense) interpretation of the '*Do*' variables. This means that CONS must contain an axiom somehow mentioning $Do(F_i(1), b)$. From def. 8.1 we see there are two possibilities: The first possibility is that CONS contains the axiom $Do(F_i(1), b)$, i.e. an axiom of the form given in item 4 of def. 8.1. In this case, it follows $F_i^b \in E$ so $F_i^b \in (S \cap S') \cup E \cup C^{S'}$ and we are done.

The second possibility is that CONS contains an axiom of form (5) while $\mathcal{M} \models \Phi$ where Φ is as in (5) and one of the conjuncts on the right-hand side of the axiom is $Do(F_i(1), b)$. In that case, C contains the corresponding rule $\Phi' \Rightarrow \Psi'$ with $F_i^b \in \Psi'$ and, by def. of S', $\Phi' \in S'$. Therefore, $F_i^b \in C^{S'}$ and hence $F_i^b \in (S \cap S') \cup E \cup C^{S'}$ and we are done. □

4. Appendix: $\mathcal{L}_3$ and first-order causal theories

4.1 Formal definition of modelhood

The definitions in this section have all been copied from [3]. In order to formally define the notion of 'model', we first need to introduce a menagerie of other concepts. For detailed explanation of these concepts and clarifying examples, we refer to [3]. Roughly, the definition of 'model' for BG's domain descriptions D (def. 4.4) makes use of two other, more basic concepts: the first is the notion of *causal model*, which will be introduced below. The second is the notion of a fact being true in an interpretation M, a notion which will be introduced directly after.

In all that follows, $\circ$ stands for *concatenation*: For a given sequence of objects α and an object a, the 'concatenation of α and a', written as $\alpha \circ a$, stands for the sequence of α followed by a.

Causal Models A *state* is a set of fluent names. A *causal interpretation* is a partial function Ψ from sequences of actions to states such that 1) the empty sequence [] belongs to the domain of Ψ and 2) Ψ is prefix-closed, i.e. for any sequence of actions α and any action a, if $\alpha \circ a$ is in the domain of Ψ then so is α.

Given a fluent f and a state σ, we say that f is *true* in σ if $f \in \sigma$; $\neg f$ is *true* in σ if $f \notin \sigma$. The truth of a propositional formula C with respect to σ is defined as usual.

A fluent f is an *immediate effect* of (executing) a in σ if there is an effect law 'a **causes** f **if** $p_1, \ldots, p_n$' in D whose preconditions $p_1, \ldots p_n$ hold in σ.

A fluent f is an *inherited effect* of (executing) a in σ if there is $b \subset a$ such that 1) f is an immediate effect of b in σ; b) there is no action c such that $b \subset c \subseteq a$ and $\neg f$ is an immediate effect of c in σ.

f is an *effect* of (executing) a in σ if either f is an immediate effect of a in σ or f is an inherited effect of a in σ.

Now let $E_a^+(\sigma) = \{f : f \text{ is an effect of } a \text{ in } \sigma\}$,$E_a^-(\sigma) = \{f : \neg f \text{ is an effect of } a \text{ in } \sigma\}$ and $Res(a, \sigma) = \sigma \cup E_a^+(\sigma) \setminus E_a^-(\sigma)$.

Definition 4.1 *A causal interpretation* Ψ satisfies *effect laws of D if for any sequence* $\alpha \circ a$ *from the language of D:*

$$\Psi(\alpha \circ a) = Res(a, \Psi(a)) \text{ if } E_a^+(\Psi(\alpha)) \cap E_a^-(\Psi(\alpha)) = \emptyset$$

and undefined otherwise.
We say that Ψ *is a* causal model *of D if it satisfies all the effect laws of D.*

Truth of Facts We already defined a *situation assignment* on page 26. We will now make our definition of 'interpretation' (also page 26) more precise:

For any domain description D and any causal interpretation Ψ that is a causal model of D, an *interpretation* M of $\mathcal{L}_3$ is a pair (Ψ, Σ) where Ψ is a causal model of D, Σ is a situation assignment of $\mathcal{S}$ and $\Sigma(s_N)$ belongs to the domain of Ψ.

Definition 4.2 *For any interpretation* $M = (\Psi, \Sigma)$,

1. $(f$ **at** $s)$ *is* true *in M (or satisfied by M) if f is true in* $\Psi(\Sigma(s))$.
2. *Atomic fact* $([a_1, \ldots, a_n]$ **occurs_at** $s)$ *is true in M if there is a sequence* $\beta = [b_1, \ldots, b_n]$ *of actions such that 1) The sequence* $\Sigma(s) \circ \beta$ *is a prefix of the actual path of M, and 2) for any* $i, 1 \leq i \leq n$, $a_i^* \subseteq b_i$. *Here* a_i^* *is an instance*[7] *of* a_i.
3. $(s_1$ **precedes** $s_2)$ *is true in M if* $\Sigma(s_1)$ *is a proper prefix of* $\Sigma(s_2)$.
4. *Truth of non-atomic facts in M is defined as usual.*

A set of facts is true in interpretation M if all its members are true in M.

[7] For any compound action a^* and any generalized action a, we say a^* is an instance of a if $a = a_1|\ldots|a_m$ and for some $1 \leq i \leq m$, $a_i = a^*$.

Models of domain descriptions D First, we need to introduce one more concept:

Definition 4.3 *Let $\alpha = [a_1, \ldots, a_n]$ and $\beta = [b_1, \ldots, b_m]$ be sequences of actions. We say that $\alpha \leq \beta$ if there exists a subsequence*[8] *$[b_{i_1}, \ldots, b_{i_n}]$ of β such that for every $a_k \in \alpha, a_k \subseteq b_{i_k}$.*

We can now finally define what it means to be a model:

Definition 4.4 [3] *An interpretation $M = (\Psi, \Sigma)$ is called a* model *of a domain description D in $\mathcal{L}_3$ if the following conditions are satisfied:*

C1. *Ψ is a causal model of D*

C2. *facts of D are* true *in M*

C3. *there is no other interpretation $N = (\Psi', \Sigma')$ such that N satisfies the conditions* **C1.** *and* **C2.** *and $\Sigma'(s_N) \leq \Sigma(s_N)$ and $\Sigma'(s_N) \neq \Sigma(s_N)$ and $\Psi'(\Sigma'(s_N))$ is defined.*

The definition given above is slightly different from the one given in [3, 5]. This is because we opt for the version of BG's approach in which the models with the least number of actions are always preferred (see page 28). At page 10 of [5], it is claimed that in this case, the definition should be with **C1** and **C2** as above, but **C3** replaced by

C3'. there is no other interpretation $N = (\Psi', \Sigma')$ such that N satisfies the conditions **C1.** and **C2.** and $\Sigma'(s_N) \leq \Sigma(s_N)$.

However, we must assume that BG have made a mistake here: it is easy to show that if one used **C3'** in stead of **C3**, then there will be no models for *any* domain description D whatsoever. We therefore opt for the unproblematic version of **C3** given above.

4.2 Proof of proposition 8.3
We abbreviate all fluent and action names in the obvious way.

We have $\mathcal{F} = \{W\}, \mathcal{A} = \{LL, LR, T\}, \mathcal{S} = \{s_0, s_N\}$. Suppose $M = (\Psi, \Sigma)$ is a *model* of D. We then have by definition 4.4, **C2** and (15) that $\neg W$ **at** s_0 must hold in M. By def. 4.2, item 2 it follows that $\Psi([]) = \emptyset$ (i.e. $\neg W$ is true in $\Psi([])$). Also, Ψ must be a causal model of D. It is straightforward to verify that for $a = \{T, LL, LR\}$ we have $E_a^+(\Psi(\alpha)) \cap E_a^-(\Psi(\alpha)) = \{W\}$. Definition. 4.1 now tells us that $\Psi([\{T, LL, LR\}])$ is undefined. Now since by def. 4.2, item 2, we have for any model $M = (\Psi, \Sigma)$ of D that $[\{T, LL, LR\}]$ is a prefix of $\Sigma(s_N)$, it follows that $\Psi(\Sigma(s_N))$ is undefined too, i.e. M does not determine what happens after s_0. □

4.3 Proof of theorem 8.2
The proof of theorem 8.2 is necessarily quite involved, the reason being that both our definition of 'preferred models' and BG's definition of models contain a self-reference: a model $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$ is preferred if there is *no other* model $\mathcal{M}' \models_{\mathbf{c}} T_{\mathrm{C}}$ with a smaller interpretation of Ab_1. Similarly, M is a model of D if there is *no other* model satisfying conditions **C1** and **C2** of def. 4.4 in which fewer actions happen. For this reason, we will first show that for corresponding M and $\mathcal{M}$, we have that the same fluents and events hold at the same time (lemma 4.1). We use this result to show that for corresponding M and $\mathcal{M}'$, we have $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$ iff conditions **C1** and **C2** hold for M (this is essentially what happens in lemmas 4.2,4.3 and 4.4). Only then will we be in a position to prove the theorem itself. We start by proving the four lemmas mentioned.

[8] Given a sequence $X = x_1, \ldots, x_m$, another sequence $Z = z_1, \ldots, z_n$ is a subsequence of X if there exists a strictly increasing sequence $i_1, \ldots, i_n$ of indices of X such that for all $j = 1, 2, \ldots, n$, we have $x_{i_j} = z_j$.

Correspondence of Facts In the following, whenever we mention T_{C} , we mean a causal theory T_{C} defined for a language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F}, \mathbf{T})$ according to definition 8.4.

Lemma 4.1 *For any T_{C} for an instance of our language and any non-ambiguous domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$ such that $T_{\mathrm{C}} \sim D$, let $M = (\Psi, \Sigma)$ be any interpretation of $(\mathcal{F}, \mathcal{A}, \mathcal{S})$ and $\mathcal{M}$ be any interpretation of our language such that $M \cong \mathcal{M}$. Let F be any fact in D and ϕ be any axiom in* CONS *such that F and ϕ are corresponding facts (where correspondence is defined as in def. 8.5, items 5-8). We have:*

$$F \text{ is true in } M \quad \Leftrightarrow \mathcal{M} \models \phi \tag{1}$$

Proof of lemma 4.1: The fact F mentioned in the lemma can either be atomic or non-atomic. We first consider atomic facts. They can be of three kinds: fluent facts, occurrence facts and precedence facts. We only give the proof for fluent facts. The proofs for the other two cases follow the same scheme as the one for fluent facts so we omit them.

So suppose that F is a fluent fact 'f **at** s'. We show that if F is true in M, then the corresponding fluent fact ϕ holds in $\mathcal{M}$. Notice ϕ equals '$Ho(f, s)$'.

So suppose F is true in M. It follows that f is true in $\Psi(\Sigma(s))$, so $f \in \sigma_t$ for the t that is equal to the length of the sequence $\Sigma(s)$. Now since $M \cong \mathcal{M}$, it follows from def. 8.6 that $\mathcal{M} \models Ho(f, t)$ and $\mathcal{M} \models s = t$, so $\mathcal{M} \models Ho(f, s)$, so $\mathcal{M} \models \phi$.

If ϕ holds in $\mathcal{M}$, then it follows that F is true in D by an analogous line of reasoning, but in the other direction. We omit the details.

We now consider non-atomic facts. The truth of non-atomic facts F in an interpretation M is defined in terms of its constituent atomic facts in exactly the same manner as the truth of non-atomic sentences for a model $\mathcal{M}$ is defined in terms of its constituent atoms. Since we have already proven that the lemma holds for all atomic facts, it follows that it still holds for non-atomic facts. □

Causal Consistency and its Lemma's

Definition 4.5 (Causally Consistent) *For any T_{C} for an instance of our language and any non-ambiguous domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$ such that $T_{\mathrm{C}} \sim D$, let $M = (\Psi, \Sigma)$ be any interpretation of $(\mathcal{F}, \mathcal{A}, \mathcal{S})$ and $\mathcal{M}$ be any interpretation of our language. We say that M and $\mathcal{M}$ are* causally consistent *iff*

1. *$M \cong \mathcal{M}$.*
2. *$\mathcal{M} \models \forall e, b, t.\ \neg Do(e, b, t)$*
3. *For all F_i^b: a) $\mathcal{M} \models \neg Do(F_i, b, 0)$. b) for all $t \in \{1, 2, \dots\}$: $\mathcal{M} \models Do(F_i, b, t)$ iff* CONS *contains an axiom ϕ of form (18) such that 1) the F_i and b mentioned in the right-hand side of ϕ are equal to the F_i and b mentioned here and 2) the left-hand side of ϕ holds in $\mathcal{M}$.*
4. *$\mathcal{M} \models \forall e, t\ .\ Ab_1(e, t) \equiv Ho(e, t)$.*
5. *$|\mathcal{M}|_b = \mathbf{B}; |\mathcal{M}|_e = \mathbf{E}; |\mathcal{M}|_f = \mathbf{F}$; $\mathcal{M}$ interprets all constants in $\mathbf{B}$, $\mathbf{E}$ and $\mathbf{F}$ as themselves.*
6. *Ψ is a causal model of D.*

Lemma 4.2 *For any T_{C} for an instance of our language and any non-ambiguous domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$ such that $T_{\mathrm{C}} \sim D$, let $M = (\Psi, \Sigma)$ be any interpretation of $(\mathcal{F}, \mathcal{A}, \mathcal{S})$ and $\mathcal{M}$ be any interpretation of our language such that $M \cong \mathcal{M}$. We have:*

$$\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}} \Rightarrow \textit{All facts of } D \textit{ are true in } M \tag{2}$$

and furthermore

$$\textit{All facts of } D \textit{ are true in } M \ \&\ M \textit{ and } \mathcal{M} \textit{ are causally consistent} \tag{3}$$

$$\Rightarrow$$

$$\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$$

Lemma 4.3 *For any T_C for an instance of our language and any domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$ such that $T_C \sim D$, let $M = (\Psi, \Sigma)$ be any interpretation of D such that 1) M satisfies conditions* **C1** *and* **C2** *of definition 4.4 and 2) $\Psi(\Sigma(s_N))$ is defined. Then there exists an interpretation $\mathcal{M}$ of our language such that M and $\mathcal{M}$ are causally consistent.*

Lemma 4.4 *For any T_C for an instance of our language and any non-ambiguous domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$ such that $T_C \sim D$, let $\mathcal{M} \models_c T_C$. Then there exists an interpretation $M = (\Psi, \Sigma)$ of $\mathcal{L}_3$ such that M and $\mathcal{M}$ are causally consistent.*

Proofs of lemma's 4.2-4.4

Proof of lemma 4.2 (first part) We prove the contrapositive of (2). Suppose not all facts of D hold in M. By lemma 4.1, the fact that $M \cong \mathcal{M}$ and the definition of corresponding theories (def. 8.5) it follows that $\mathcal{M} \not\models$ CONS and therefore $\mathcal{M} \not\models_c T_C$. □

Proof of lemma 4.2 (second part) Suppose we have an M and $\mathcal{M}$ such that $M \cong \mathcal{M}$ and (3) holds. We show in three stages that $\mathcal{M} \models_c T_C$: in stage 1, we show $\mathcal{M} \models$ CONS. We use this to show in stage 2 that $\mathcal{M} \models$ *Circum*(CONS; *Do*). In stage 3, we show that $\mathcal{M} \models$ EQ. Stage 1–3 together show that $\mathcal{M}$ satisfies condition (1) of the definition of models for causal theories (def. 7.2). That $\mathcal{M}$ satisfies condition (2) of that definition follows already from the definition of '$\cong$' (def. 8.6, item (b)).

Stage 1 We show that $\mathcal{M} \models \phi$ for all axioms ϕ contained in CONS. From def. 8.5, item 2, it follows that CONS contains UNA- and DC-axioms for **B**, **E** and **F**. It follows directly from the fact that M and $\mathcal{M}$ are causally consistent that these axioms hold for $\mathcal{M}$ too. We now consider the axioms of form (18) in CONS (def. 8.5 item 4). Let us denote by ϕ any of these axioms. ϕ can be written as $\forall t.\mathbf{H}(t) \supset Do(F_i^b, t+1)$ where **H** is some conjunction of instances of *Ho*. We now show that for all $t \geq 0$, $\mathcal{M} \models \mathbf{H}(t) \supset Do(F_i, b, t+1)$. Since we assume that *time points* are interpreted as the nonnegative integers, this is sufficient to show $\mathcal{M} \models \phi$. So take any $t \geq 0$. For this t, either $\mathcal{M} \not\models \mathbf{H}(t)$ and we are done. If $\mathcal{M} \models \mathbf{H}(t)$ then by the definition of causal consistency, we have $\mathcal{M} \models Do(F_i, b, t+1)$ and we are done again.

It follows from lemma 4.1 and the fact that all facts of D are true in M that $\mathcal{M} \models \phi$ for all axioms ϕ in T_C mentioned in items 5–8 of definition 8.5 (all these axioms correspond to some fact in D). From item 3 of def. 8.5, we conclude that CONS contains no more axioms than those for which we already checked that they hold in $\mathcal{M}$. This completes stage 1.

Stage 2 We show that $\mathcal{M} \models$ *Circum*(CONS; *Do*). Notice first that by proposition 7.1 all models share the same universes for CONS. From this and the characterization of circumscription given in proposition 2.1 it follows that it is sufficient to show that among all the models for CONS with the same interpretation of *Ho*, there is no $\mathcal{M}'$ whose interpretation of *Do* is a proper subset of that of $\mathcal{M}$. Suppose, by means of contradiction, that such an $\mathcal{M}'$ exists, i.e. $\mathcal{M}' \models$ CONS, $\mathcal{M}[\![Ho]\!] = \mathcal{M}'[\![Ho]\!]$ but $\mathcal{M}'[\![Do]\!] \subset \mathcal{M}[\![Do]\!]$. By the definition of causal consistency, item 2, $\mathcal{M}$ and therefore also $\mathcal{M}'$ have $\neg Do(e, b, t)$ for any e, b, t. It follows there is a pair $(F_{i^*}^{b^*}, t^*)$ with $\mathcal{M}' \models \neg Do(F_{i^*}, b^*, t^*+1)$ while $\mathcal{M} \models Do(F_{i^*}, b^*, t^*+1)$. It follows from the definition of causal consistency, item 3 that T_C contains an axiom ϕ of form (18) where 1) the F_i,b and t mentioned in the right-hand side of ϕ are equal to F_{i^*},b^* and t^* and 2) the left-hand side of ϕ holds in $\mathcal{M}$, and therefore also in $\mathcal{M}'$. But also $\mathcal{M}' \models$ CONS so $\mathcal{M}' \models Do(F_{i^*}, b^*, t^*)$ and we arrive at a contradiction.

Stage 3 We show that $\mathcal{M} \models$ EQ by showing that axioms (1) and (2), (4) and (15) hold in $\mathcal{M}$. By items 2 and 4 of the definition of causal consistency and the fact that M and $\mathcal{M}$ are causally consistent it follows that (15) holds in $\mathcal{M}$. Concerning (1) and (2), we must show that for all $x \in \mathbf{E} \cup \mathbf{F}, t \in \mathbb{N}_0$, if $\mathcal{M} \models Do(x, \text{TRUE}, t)$ then we also have $\mathcal{M} \models Ho(x, t)$ and similarly, if $\mathcal{M} \models Do(x, \text{FALSE}, t)$ then we also have $\mathcal{M} \models \neg Ho(x, t)$. Let us thus assume $\mathcal{M} \models Do(x, b, t)$ for some $x \in \mathbf{E} \cup \mathbf{F}$, $t \in \mathbb{N}_0$ and $b \in \mathbf{B}$. Since $\mathcal{M} \models \neg Do(e, b, t)$ for any e, b and t (see the definition of causal consistency), we may assume $x =$

F_i^b for some $F_i \in \mathbf{F}$. From the same definition, it follows $t \geq 1$ and that CONS must contain an axiom ϕ of form (18) such that the right-hand side of ϕ is equal to $Do(F_i, b, t)$ and the left-hand side holds in $\mathcal{M}$. It follows D contains an effect law of form (19), i.e. of form $\{a^1, \ldots, a^n\}$ **causes** F_i^b **if** $F_{j_1}^{b_1}, \ldots, F_{j_m}^{b_m}$. Since $M = (\Psi, \Sigma)$ corresponds to $\mathcal{M}$ we have that $\{a^1, \ldots, a^n\} \subseteq a_{t-1}$ and $F_{j_1}^{b_1}, \ldots, F_{j_m}^{b_m}$ all hold in σ_{t-1}. By the following claim we have that F_i^b is an effect of a_{t-1} in σ_{t-1}.

Sublemma 4.1 *Let σ be a state for a non-ambiguous domain description D. Let f be a fluent literal and a be a compound action. Now suppose there is $b \subseteq a$ such that f is an* immediate *effect of b in σ. Then f is an (ordinary) effect of a in σ.*

Proof: follows almost directly from the definitions of immediate effect, inherited effect (section 4.1 and ambiguity (def. 8.3) . We omit the details. □

By the definition of causal consistency Ψ is a causal model of D and $M \cong \mathcal{M}$. Since $M \cong \mathcal{M}$, $\Psi(\Sigma(s_N))$ is defined. We now have by the claim below that σ_t is defined.

Claim Let $M = (\Psi, \Sigma)$ be an interpretation for some domain description D. If $\Psi(\Sigma(s_N))$ is defined, then for all $t \geq 0$, σ_t is defined.
Proof: is immediate from the definition of σ_t in def. 8.6. □

Since σ_t is defined, it follows by def. 4.1 that $\sigma_t = Res(a_{t-1}, \sigma_{t-1})$. Since F_i^b is an effect of a_{t-1} in σ_{t-1}, this means F_i^b is contained in σ_t. Since $M \cong \mathcal{M}$, we have $\mathcal{M} \models Ho(F_i^b, t)$, which is what we had to prove.
We will now show that $\mathcal{M} \models (4)$. For this, suppose $\mathcal{M} \models \neg Ho(F_i^b, t-1) \wedge Ho(F_i^b, t)$ for some F_i^b and t. One can show $\mathcal{M} \models Do(F_i, b, t)$ in a manner analogous to (but simpler than) the reasoning in the first part of stage 3, above. We omit the details of this part of the proof. Since we can show this for *any* $t > 0$, *any* b and *any* F_i, it follows by proposition 7.1 that the following axiom holds for $\mathcal{M}$:

$$\forall f, t\ .(t > 0) \supset$$
$$\neg[\ Ho(f, t) \equiv\ Ho(f, t-1)\ ] \supset [\ Do(f, \text{TRUE}, t) \quad \vee \quad Do(f, \text{FALSE}, t)\ ]$$

which is clearly equivalent to (4). This ends the proof of lemma 4.2. □

Proof of lemma 4.3 $\Psi(\Sigma(s_N))$ is defined; this means definition 8.6 applies. Items (1) and (2) of the definition determine the interpretation of $Ho(x, t)$ for all $x \in \mathbf{E} \cup \mathbf{F}$ and all $t \geq 0$ in a way that clearly cannot lead to any contradiction. Also, it is clear that item (3) in the definition cannot lead to any contradictory assignment in $\mathcal{M}$ of time name *symbols* to time *points*. Hence an interpretation $\mathcal{M} \cong M$ exists. Now items 2 to 5 of the definition of causal consistency determine the interpretations of all atoms in $\mathcal{M}$ other than those of the form $Ho(x, t)$, again in a way that cannot give rise to any contradiction. □

Proof of lemma 4.4 It follows immediately from the fact that $\mathcal{M} \models_{\mathbf{c}} T_{\text{C}}$ and the form that CONS must have according to def. 8.5, that conditions 2 and 3 of the definition of causal consistency hold for $\mathcal{M}$. Since condition 2 holds and $\mathcal{M} \models_{\mathbf{c}} T_{\text{C}}$ and so, in particular, $\mathcal{M} \models Circum(\text{CONS}; Do)$, condition 4 holds too. By proposition 7.1 condition 5 holds too. Hence it only remains to prove that an interpretation $M = (\Psi, \Sigma)$ exists with $M \cong \mathcal{M}$ such that Ψ is a causal model for D. Using def. 8.6 we can construct such an M as follows:

1. We set $\Sigma(s_N) = [a_0, \ldots, a_{last}]$ where a_i is defined as in def. 8.6, item 1. We define, for any $s \in \mathcal{S}$, $\Sigma(s)$ to be the prefix of $\Sigma(s_N)$ of length t for the t for which $\mathcal{M} \models s = t$. This completes the definition of Σ.
2. We now define Ψ: we set for all $t \geq 0$, $\Psi([a_0, \ldots, a_{t-1}]) = \sigma_t$ (we use the convention that $[a_{-1}] = []$, so $\Psi([]) = \sigma_0$), where the a_i's are defined as above and σ_t is defined as in def. 8.6,

item 2. Let us denote by the Σ-*sequences* those sequences that start with a prefix of $\Sigma(s_N)$ (note this includes the empty sequence). Now for *any* sequence of compound actions that is not a Σ-sequence we define $\Psi(\alpha)$ to be such that it satisfies def. 4.1. Note that this is always possible, since according to that definition $\Psi(\alpha)$ is allowed to be undefined. It can be seen from that definition that, since $\Psi([])$ has already been defined above, this uniquely defines the function Ψ for any sequence that does not start with a prefix of $\Sigma(s_N)$. This completes the definition of Ψ.

For any sequence of compound actions except possibly the Σ-sequences, we know (from the way we defined Ψ) that Ψ satisfies all the effect laws of D. If we can prove that the Σ-sequences, too, satisfy all the effect laws of D, we have by def. 4.1 that Ψ is a causal model of D and we are done.

We now show that indeed for any Σ-sequence α, $\Psi(\alpha)$ satisfies all effect laws of D. We start with the following claim:

Claim For any Σ-sequence $\alpha \circ a_t = [a_0, \ldots a_{t-1}, a_t]$, we have $E^+_{a_t}(\Psi(\alpha)) \cap E^-_{a_t}(\Psi(\alpha)) \neq \emptyset$. **Proof**: Suppose the above is not the case. Then by def. 4.1, $E^+_{a_t}(\sigma_t) \cap E^-_{a_t}(\sigma_t) = \emptyset$ (here σ_t is defined as in def. 8.6, item 2. So there is an $F_i \in \mathcal{F}$ such that both F_i and $\neg F_i$ are effects of a_t in σ_t. Hence D contains two effect laws of form (19) with right-hand sides F_i^{TRUE} and F_i^{FALSE} respectively, such that the preconditions of both laws hold in σ_t and the actions mentioned in both laws are subsets of a_t. We call these laws A and A'. Since $T_C \sim D$, CONS contains two axioms, ϕ and ϕ', of the form (18) that correspond to A and A', respectively. Now $\mathcal{M} \models$ CONS, but further, since we defined M such that def. 8.6, items 1 and 2 hold for $\mathcal{M}$ and M, we have that $\mathcal{M} \models Ho(x,t)$ where for x we can substitute any precondition $F^{b_k}_{j_k}$ or unit action a^l mentioned in either A or A'. This means that the left-hand sides of ϕ and ϕ' hold in $\mathcal{M}$. It follows $\mathcal{M} \models Do(F_i, \mathrm{TRUE}, t) \wedge Do(F_i, \mathrm{FALSE}, t)$. But then $\mathcal{M} \not\models$ EQ and thus $\mathcal{M} \not\models_{\mathbf{c}} T_C$, which is in contradiction to our assumptions. □

From the claim, it follows that we must prove that for any Σ-sequence $\alpha \circ a_t = [a_0, \ldots a_{t-1}, a_t]$, for any $t \geq 0$, we have $\Psi(\alpha \circ a_t) = Res(a_t, \Psi(\alpha))$. From the definition of σ_t and definition 4.1, we know that this is equivalent to showing that for all $F_i \in \mathcal{F}$, for all $t \geq 0$,

$$F_i \in \sigma_{t+1} \Leftrightarrow F_i \in [\sigma_t \cup E^+_{a_t}(\sigma_t)] \setminus E^-_{a_t}(\sigma_t)] \tag{4}$$

We will prove (4) using the following

Claim Let either $b = \mathrm{TRUE}$ and $\overline{b} = \mathrm{FALSE}$ or vice versa. For any $\mathcal{M}$ with $\mathcal{M} \models_{\mathbf{c}} T_C$ and any t and F_i we have

$$\mathcal{M} \models Ho(F_i^b, t+1) \Leftrightarrow \mathcal{M} \models [Ho(F_i^b, t) \vee Do(F_i, b, t+1)] \wedge \neg Do(F_i, \overline{b}, t+1)$$

Proof: trivial & omitted. □

We show the if-direction of (4). Suppose $F_i \in \sigma_{t+1}$. Then $\mathcal{M} \models Ho(F_i, t+1)$ and by the claim above, $\mathcal{M} \models \neg Do(F_i, \mathrm{FALSE}, t+1)$. This means CONS contains no axiom of form (18) with right-hand side $Do(F_i, \mathrm{FALSE}, t+1)$ and left-hand side such that it holds at time t. As $T_C \sim D$ and $M \cong \mathcal{M}$, there is no rule in D of form (19) such that $\{a^1, \ldots, a^n\} \subseteq a_t$ and $F^{b_1}_{j_1}, \ldots, F^{b_m}_{j_m}$ all hold in σ_t. This means $\neg F_i$ is not an effect of a_t in σ_t so $F_i \notin E^-_{a_t}(\sigma_t)$.

From our assumption $F_i \in \sigma_{t+1}$ it also follows that either $\mathcal{M} \models Ho(F_i, t)$ or $\mathcal{M} \models Do(F_i, \mathrm{TRUE}, t+1)$. In the first case, $F_i \in \sigma_t$ and we are done; in the second case, since $\mathcal{M} \models_{\mathbf{c}} T_C$ and thus $\mathcal{M} \models Circum(\mathrm{CONS}; Do)$, using propositions 2.1 and 7.1, there are no models $\mathcal{M}'$ for CONS with $\mathcal{M}'[\![Ho]\!] = \mathcal{M}[\![Ho]\!]$ and $\mathcal{M}' \not\models Do(F_i, \mathrm{TRUE}, t+1)$. This means that for some t, CONS must mention $Do(F_i, \mathrm{TRUE}, t+1)$ in one of its axioms. This can only be an axiom ϕ of form (18) and it follows that the left-hand side of this axiom, instantiated to time t, holds in $\mathcal{M}$ and hence that D contains an effect law of form (19) corresponding to ϕ such that $\{a^1, \ldots, a^m\} \subseteq a_t$ and $F^{b_1}_{j_1}, \ldots, F^{b_m}_{j_m}$ all hold in σ_t. From sublemma 4.1 on page 46 it follows that F_i is an effect of a_t in σ_t, so $F_i \in E^+_{a_t}(\sigma_t)$. □

The only-if direction of (4) can be proven in a manner completely analogously to the if-direction, again using the claim above. We omit the details.

This finishes the proof of lemma 4.4. □

The Actual Proof Before actually giving the proof, we have to introduce yet one more proposition which says that the set of event-time pairs holding in a model always corresponds to the set of abnormalities in the model.

Proposition 4.1 *For any T_C for an instance of our language and any domain description D for a language $\mathcal{L}_3$ such that $T_C \sim D$, let $\mathcal{M}$ be any model with $\mathcal{M} \models_c T_C$. For such an $\mathcal{M}$ we have:*

$$\mathcal{M}[\![Ab_1]\!] = \{(e,t) \mid \mathcal{M} \models Ho(e,t); e \in \mathbf{E}; t \in \mathtt{N}_0\}$$

Proof: Since $T_C \sim D$, it follows from def. 8.5 that CONS contains no axioms mentioning $Do(e,b,t)$ for any e, b, t. Hence the circumscription of Do in CONS will ensure that in all models $\mathcal{M}$ with $\mathcal{M} \models_c T_C$ we have $\mathcal{M} \models \forall e,b,t\ .\ \neg Do(e,b,t)$. The proposition then follows immediately from axiom (15) and the fact that all models of T_C interpret all elements of **E** and **F** as themselves (proposition 7.1). □

We now first prove the first part of theorem 8.2, which is restated below.

Theorem 4.2 (part I) *For any theory T_C for our language and any domain description D for a language $\mathcal{L}_3$ of Baral and Gelfond's such that a) $T_C \sim D$ and b) D is not* ambiguous*, we have:*

$M = (\Psi, \Sigma)$ *is a model for D such that $\Psi(\Sigma(s_N))$ is defined* $\Rightarrow$
there exists an $\mathcal{M}$ with $M \cong \mathcal{M}$ such that $\mathcal{M}$ is a preferred model of T_C

Proof: By lemma 4.2 and lemma 4.3 together we have that there exists an $\mathcal{M}$ with $\mathcal{M} \models_c T_C$ that is causally consistent to M. Clearly $\mathcal{M} \cong M$. Suppose that $\mathcal{M}$ is *not* preferred, i.e. that there is an $\mathcal{M}^* \models_c T_C$ with $\mathcal{M}^*[\![Ab_1]\!] \subset \mathcal{M}[\![Ab_1]\!]$. By lemma 4.4, this means that there is an $M^* = (\Psi^*, \Sigma^*)$ such that M^* and $\mathcal{M}^*$ are causally consistent and hence condition **C1** of the definition of modelhood (def. 4.4) is satisfied for M^*. Applying lemma 4.2 (first part) we find that for such an M^*, condition **C2** of that definition is satisfied too. Since $\mathcal{M}^*[\![Ab_1]\!] \neq \mathcal{M}[\![Ab_1]\!]$, we have by proposition 4.1 and the fact that $\mathcal{M} \cong M$ and $\mathcal{M}^* \cong M^*$ that $\Sigma^*(s_N) \neq \Sigma(s_N)$. We are assuming M is a model of D, so we have by condition **C3** in def. 4.4 that it is *not* the case that $\Sigma^*(s_N) \leq \Sigma(s_N)$. Then it is not the case either that $a_t^* \subseteq a_t$ for all t (where a_t is defined as in def. 8.6, and a_t^* stands for the a_t belonging to Σ^*). Since $M \cong \mathcal{M}$ and $M^* \cong \mathcal{M}^*$, it follows by definition 8.6 and proposition 4.1 that it is not the case that $\mathcal{M}^*[\![Ab_1]\!] \subset \mathcal{M}[\![Ab_1]\!]$ and we have arrived at a contradiction. □

We now restate and prove the second part of theorem 8.2.

Theorem 4.2 (part II) *For any theory T_C for our language and any domain description D for a language $\mathcal{L}_3$ of Baral and Gelfond's such that a) $T_C \sim D$ and b) D is not* ambiguous*, we have, for all $\mathcal{M}$:*

$$\mathcal{M} \text{ is a preferred model of } T_C \Rightarrow \text{there exists a model } M \text{ for } D \text{ such that } M \cong \mathcal{M}$$

Proof: Clearly, if $\mathcal{M}$ is a preferred model of T_C then also $\mathcal{M} \models_c T_C$. By lemma 4.4 we have that there exists an M that is causally consistent to $\mathcal{M}$ and that therefore satisfies condition **C1** of the definition of modelhood (def. 4.4). Clearly, $M \cong \mathcal{M}$. By lemma 4.2 (first part) we further have that M also satisfies condition **C2** of definition 4.4. We now prove that M is a model of D by showing that assuming that condition **C3** of def. 4.4 does *not* hold for M leads to a contradiction.

So assume this condition does not hold for M. Then there exists $M' = (\Psi', \Sigma')$ that satisfies **C1** and **C2** of def. 4.4 with $\Sigma'(s_N) \neq \Sigma(s_N)$ and $\Sigma'(s_N) \leq \Sigma(s_N)$ and $\Psi'(\Sigma'(s_N))$ is defined. By lemma 4.3 and lemma 4.2 (second part) together we have that there is an $\mathcal{M}' \models_{\mathbf{c}} T_{\mathrm{C}}$ with $\mathcal{M}' \cong M'$. We will now 'stretch' $\mathcal{M}'$ into another $\mathcal{M}'' \models_{\mathbf{c}} T_{\mathrm{C}}$ which will have the property that $\mathcal{M}''[\![Ab_1]\!] \subset \mathcal{M}[\![Ab_1]\!]$. In $\mathcal{M}''$ exactly the same events and changes take place as in $\mathcal{M}'$, but the time in between two events/changes may be larger. We now prepare the construction of $\mathcal{M}''$:

Let $\Sigma(s_N) = [a_0, \ldots, a_{last}]$ and $\Sigma'(s_N) = [a'_0, \ldots, a'_{last'}]$. Notice first that, since $\Sigma'(s_N) \neq \Sigma(s_N)$ and $\Sigma'(s_N) \leq \Sigma(s_N)$, there is a strictly increasing sequence of time points $x_0, \ldots, x_{last'}$ such that for all $0 \leq t \leq last'$ we have $\{e : \mathcal{M}' \models Ho(e,t)\} \subseteq \{e : \mathcal{M} \models Ho(e, x_t)\}$ where for at least one t, the inclusion is proper. Also, for all $t > last'$, $\{e : \mathcal{M}' \models Ho(e,t)\}$ is empty.

Now define $\mathcal{M}''$ as follows: for all $0 \leq t \leq last'$, the extension in $\mathcal{M}''$ of Ho, Ab_1 and Do at time x_t is defined to be equal to its extension in $\mathcal{M}'$ at time t; i.e. for all x: $\mathcal{M}' \models Ho(x,t) \Leftrightarrow \mathcal{M}'' \models Ho(x, x_t)$ and analogously for Do and Ab_1. We now have to 'fill up the gaps' in $\mathcal{M}''$, i.e. define the interpretations of Ho and Do for the time points in $\mathcal{M}''$ that are not equal to any of the $x_0, \ldots, x_{last'}$. For this, we set $x_{-1} := -1$. For all $0 \leq t \leq last'$, for all $x_{t-1} < t' < x_t$ and $t' > x_{last'}$, define:

- for all $e \in \mathbf{E}$: $\mathcal{M}'' \models \neg Ho(e,t') \wedge \neg Ab_1(e,t')$.
- for all $x \in \mathbf{E} \cup \mathbf{F}$: $\mathcal{M}'' \models \neg Do(x, \textsc{true}, t') \wedge \neg Do(x, \textsc{false}, t')$.
- for all $f \in \mathbf{F}$: $\mathcal{M}'' \models Ho(f,t') \Leftrightarrow \begin{cases} \mathcal{M}' \models Ho(f,t) & \text{if } t' < x_t \\ \mathcal{M}' \models Ho(f, last'+1) & \text{if } t' > x_{last'} \end{cases}$

Clearly, all predicates in the language for $\mathcal{M}''$ have now been defined for all t, f, e, b (notice that Ab_1 only exists for event-time pairs, not for fluent-time pairs, see section 7.4).

One easily verifies that an $\mathcal{M}''$ as defined above indeed exists, and that from the fact that $\mathcal{M}' \models_{\mathbf{c}} T_{\mathrm{C}}$, it follows that also $\mathcal{M}'' \models_{\mathbf{c}} T_{\mathrm{C}}$ (simply check all the axioms in CONS and EQ; we omit the details). We see that $\{(e,t) | \mathcal{M}'' \models Ho(e,t)\} \subseteq \{(e,t) | \mathcal{M} \models Ho(e,t)\}$ and hence by proposition 4.1 $\mathcal{M}''[\![Ab_1]\!] \subseteq \mathcal{M}[\![Ab_1]\!]$. Since $\Sigma'(s_N) \neq \Sigma(s_N)$, by construction of $\mathcal{M}''$ and proposition 4.1 we also have $\mathcal{M}''[\![Ab_1]\!] \neq \mathcal{M}[\![Ab_1]\!]$. This means $\mathcal{M}''$ is preferred over $\mathcal{M}$ and hence $\mathcal{M}$ is not a preferred model of T_{C}, which is in contradiction with our assumption. □

References

1. J. Amsterdam. Temporal reasoning and narrative conventions. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 15–21, Cambridge, MA, 1991.
2. A.B. Baker. Nonmonotonic reasoning in the framework of situation calculus. *Artificial Intelligence*, 49:5–23, 1991.
3. 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.
4. C. Baral, M. Gelfond, and A. Provetti. Reasoning about actual and hypothetical occurrences of concurrent and non-deterministic actions. In *Proceedings AAAI Spring Symposium*, 1995.
5. C. Baral, M. Gelfond, and A. Provetti. Representing actions: Laws, observations and hypotheses. *Journal of Logic Programming*, 12:1–44, 1997.
6. 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.
7. A. Darwiche and J. Pearl. Symbolic causal networks. In *Proceedings AAAI-94*, 1994.
8. 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.
9. H. Geffner. Causality, constraints and the indirect effects of actions. In *Proceedings IJCAI '97*, pages 555–560, Nagoya, Japan, 1997.
10. M. Gelfond and V. Lifschitz. Representing actions in extended logic programming. In K. Apt, editor, *Logic Programming: Proceedings Tenth Conference*, pages 559–573, 1992.
11. E. Giunchiglia and V. Lifschitz. Dependent fluents. In *Proceedings IJCAI-95*, pages 1964–1969, 1995.
12. M. Goldszmidt and J. Pearl. Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence*, 84(1-2):57–112, 1996.
13. 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.

14. 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.
15. 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.
16. S. Hanks and D. McDermott. Default reasoning, non-monotonic logics and the frame problem. In *Proceedings AAAI-86*, pages 328–333, 1986.
17. H.A. Kautz. The logic of persistence. In *Proceedings AAAI-86*, pages 401–405, 1986.
18. V. Lifschitz. Computing circumscription. In *Proceedings IJCAI-85*, pages 121–127, 1985.
19. V. Lifschitz. Frames in the space of situations. *Artificial Intelligence*, 46:365–376, 1990.
20. V.L. Lifschitz and A. Rabinov. Miracles in formal theories of action. *Artificial Intelligence*, 38:225–237, 1989.
21. F. Lin. Embracing causality in specifying the indirect effects of actions. In *Proceedings IJCAI-95*, 1995.
22. F. Lin. Embracing causality in specifying the indeterminate effects of actions. In *Proceedings AAAI-96*, 1996.
23. N. McCain and H. Turner. A causal theory of ramifications and qualifications. In *Proceedings IJCAI-95*, 1995.
24. N. McCain and H. Turner. Causal theories of action and change. In *Proceedings AAAI-97*, 1997.
25. 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.
26. J. McCarthy. Circumscription – a form of non-monotonic reasoning. *Artificial Intelligence*, 13(1,2):27–39,171–172, 1980.
27. J. McCarthy and P.J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Mitchie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969.
28. L. Morgenstern and L.A. Stein. Why things go wrong: A formal theory of causal reasoning. In *Proceedings AAAI-88*, pages 5–23, 1988.
29. J. Pearl. Graphical models, causality, and intervention. *Statistical Science*, 8(3):266–273, 1993.
30. J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–709, 1995.
31. 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.
32. J. Pearl. Causation, action and counterfactuals. In Y. Shoham, editor, *Proceedings TARK-VI*, pages 51–73. Morgan Kaufmann, 1996.
33. 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.
34. 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.
35. L.K. Schubert. Explanation closure, action closure and the Sandewall test suite for reasoning about change. *Journal of Logic and Computation*, 4(5):679–700, 1994.

36. L.A. Stein and L. Morgenstern. Motivated action theory: A formal theory of causal reasoning. *Artificial Intelligence*, 71(1):1–42, nov 1994.

37. M. Thielscher. The logic of dynamic systems. In *Proceedings IJCAI-95*, pages 1956–1962, Montreal, 1995.

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