Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

Causation, Explanation and Nonmonotonic Temporal Reasoning

P.D. Grünwald

Information Systems (INS)

INS-R9701 February 28, 1997

Report INS-R9701
ISSN 1386-3681

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

Causation, Explanation and Nonmonotonic Temporal Reasoning

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

ABSTRACT

We introduce a new approach to reasoning about action and change using nonmonotonic logic. The approach is arrived at by applying Pearl's theory of causal networks to logical formalizations of temporal reasoning domains. It handles complicated reasoning domains involving concurrent actions, actions with nondeterministic effects, preconditions over several points in time, ramification/qualification constraints and causal chains of events. Our theory comes in two versions: version $\mathbf{S}_0$ that works for logical theories in which causal knowledge is represented explicitly, and version $\mathbf{I}_0$ that works for theories in which this is not the case. We show that $\mathbf{S}_0$ can be seen as a generalization and/or correction of most existing approaches that are explicitly based on causation. Specifically, we prove that on a large class of reasoning domains, $\mathbf{S}_0$ is equivalent to McCain & Turner's theory of Ramifications and Qualifications. For another large class of reasoning domains, $\mathbf{S}_0$ turns out to be equivalent to Baral & Gelfond's $\mathcal{L}_3$ approach. We give examples of reasoning domains that fall outside these classes and for which $\mathbf{S}_0$ yields better results than either McCain & Turner's or Baral & Gelfond's approach. We give strong arguments that $\mathbf{S}_0$ and Lin's recent causal theory are equivalent on all reasoning domains on which both are defined. For the causal approaches of Morgenstern & Stein, Geffner, Haugh and Lifschitz & Rabinov we again provide examples of reasoning domains that $\mathbf{S}_0$ handles better than they do. We also show that two of the most well-known non-causal approaches, namely Baker's account and 'chronological minimization with filter preferential entailment', can be reinterpreted as approximations of $\mathbf{I}_0$. In the case of Baker we again give an equivalence theorem and a counterexample. We thus provide a reinterpretation in terms of causal network theory of much of the work done in nonmonotonic 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

Any approach to common-sense temporal reasoning must address the *frame problem* [24]: how can one properly model the common sense intuition that most facts about the world tend to stay the same over time, unless an action takes place that affects them? Formalizing this notion of *inertia* or *persistence* has turned out to be a notoriously difficult problem to which numerous solutions have been proposed. All these different approaches to formalizing persistence can be classified in a number of families; one of the larger of these consists of those accounts in which causal knowledge (for example, of the form 'A causes B') is represented explicitly, either by a predicate or by extending the logic with a 'causal' connective or modal operator. We will call these approaches the *causal* ones (as opposed to the other, *non-causal* approaches).

On the other hand, in recent years there have been a number of papers on the use of directed acyclic graphs for the modeling of causal relations; the theory of these 'causal networks' is successfully applied in several domains [5, 6, 26, 27]. We will call the view on causation that is implied by these networks 'Pearl's account of Causation' as it is mainly associated with J. Pearl's writings.

In our work, we establish a clear connection between nonmonotonic temporal reasoning and Pearl's account of causation. We apply Pearl's theory to arrive at a nonmonotonic logic that formalizes persistence. It comes in two versions: the first, *model selection criterion* $\mathbf{S}_0$, works for causal logical theories (i.e. theories in which causal knowledge is axiomatized explicitly). It selects only those models of a theory that are subject to persistence, and it does so the way Pearl's theory prescribes that it should be done. The second version, model selection criterion $\mathbf{I}_0$, first infers *what causes what* in a non-causal theory. It then uses the causal relations it found to select the models subject to persistence. Practically all of this is done again simply by applying Pearl's ideas. Now $\mathbf{S}_0$ can be formally compared to the existing causal approaches to formalizing persistence, and $\mathbf{I}_0$ can be compared to the non-causal ones. We have proven that some recent causal and 'action description language'-based approaches are equivalent to various restrictions of $\mathbf{S}_0$ – specifically, we proved an equivalence theorem stating that, for a large, precisely defined class of reasoning domains, $\mathbf{S}_0$ and McCain and Turner's approach [21] permit exactly the same inferences. For another large class of reasoning domains, we proved that $\mathbf{S}_0$ is equivalent to Baral and Gelfond's approach based on the action description language $\mathcal{L}_3$ (an extension of Lifschitz and Gelfond's well-known language $\mathcal{A}$ [3]; also, we have made an informal comparison that makes clear that Lin's recent approach [19, 20] is largely identical to $\mathbf{S}_0$. We also provide examples of reasoning domains for which $\mathbf{S}_0$ is not equivalent to either McCain and Turner's or Baral and Gelfond's approach, and for which it gives better results than these approaches do. For the other existing causal approaches – specifically, those of Morgenstern and Stein [25], Geffner [7] and Lifschitz, Haugh and Rabinov [11, 18] we give other, new examples of reasoning domains where our approach gives better results than they do. For two popular non-causal approaches, we argue that they can be interpreted as implementing an approximation of $\mathbf{I}_0$. For one of them, namely Baker's approach [1, 4, 12], we actually prove this, again by an equivalence theorem and a counterexample. Summarizing,

> *Our approach to nonmonotonic temporal reasoning can be seen as a generalization and/or correction of most existing ones.*

Moreover, the comparisons clearly show that most of the problems with existing approaches can be interpreted as stemming from not treating causation the way it is treated in Pearl's theory.

1.1 Structure of this report

In this report we give an overview of our results on $\mathbf{I}_0$ and $\mathbf{S}_0$. The report does not contain any proofs; a longer report that contains these is in preparation.

We will now first introduce our formalism. We then introduce our basic ideas (section 3) and show how they are connected to Pearl's theory (section 4). This prepares our definition of $\mathbf{I}_0$ and $\mathbf{S}_0$ which is given in section 5. We continue to give more details on $\mathbf{S}_0$ (section 6) and $\mathbf{I}_0$ (section 7). Sections 8,9 and 10 compare $\mathbf{I}_0$ and $\mathbf{S}_0$ to the many other existing approaches. We end the paper with some concluding, somewhat more philosophical remarks. The paper is set up in a more or less modular fashion: the reader mainly interested in $\mathbf{S}_0$ may skip section 7 and section 10. If one is only concerned about $\mathbf{I}_0$, it is possible to skip section 6 and section 9.

2. Our Formalism

We use a many-sorted first order logic. The sorts in our logic are *generalized fluents* (variables of the sort will be denoted by g) and *timepoints* (t). There are two 'subsorts' to generalized fluents: *regular fluents* (f) and *event fluents* (e). A regular fluent denotes a property of the world that may be subject to persistence; an event fluent stands for an event or action. The set of regular fluent constants for our language will be called FC; it will always be finite: $\mathrm{FC} = \{F_1, \ldots, F_{|\mathrm{FC}|}\}$. Similarly, the set $\mathrm{EC} = \{E_1, \ldots, E_{|\mathrm{EC}|}\}$ contains our event fluent constants. Time-points will be equated to the nonnegative integers; additionally, our language contains a finite set of *time name constants* $\mathrm{TNC} = \{T_1, \ldots, T_{|\mathrm{TNC}|}\}$ that will enable us to have names for some of our points in time. The predicate $Ho(g,t)$ will denote that generalized fluent g is the case ('*Ho*lds') at time t. $\mathbf{e}, \mathbf{f}, \mathbf{g}$ and $\mathbf{t}$ will be used as meta-variables that can stand for any ground term of the sort indicated by the letter.

2.1 Causal and Non-Causal Theories

We formalize our domain knowledge in two different manners: non-causal theories T_{NC} , which are intended to formalize knowledge essentially in the same way as most existing non-causal approaches do, and causal theories T_{C} , which are intended to do this essentially in the same way as most existing causal approaches do. We will use a variation of the Yale Shooting Domain [10] to illustrate both of them. In this domain there is a person that can be alive or not; there is a gun that can be loaded or not; there is an action 'load' which loads the gun, an action 'wait' which has no effects at all and an action 'shoot', which, if performed when the gun is loaded, causes the person not to be alive anymore. We further suppose that the person 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. We will first show our non-causal formalization $T_{\text{YSP-NC}}$ of this domain. It contains axioms (1)-(8) :

$$\mathit{Ho}(\mathit{Load}, t) \supset \mathit{Ho}(\mathit{Loaded}, t+1) \tag{1}$$

$$\mathit{Ho}(\mathit{Loaded}, t) \wedge \mathit{Ho}(\mathit{Shoot}, t) \supset \neg \mathit{Ho}(\mathit{Alive}, t+1) \tag{2}$$

$$\mathit{Ho}(\mathit{Alive}, 0) \wedge \neg \mathit{Ho}(\mathit{Loaded}, 0) \tag{3}$$

$$\mathit{Ho}(\mathit{Load}, 0) \wedge \mathit{Ho}(\mathit{Wait}, 1) \wedge \mathit{Ho}(\mathit{Shoot}, 2) \tag{4}$$

Here *Alive* and *Loaded* are regular fluents, while *Load*, *Wait* and *Shoot* are event fluents. (1) and (2) are the *effect axioms* which describe the 'action laws' of our domain. (3) and (4) are our *observation axioms* which are always propositional combinations of *fluent facts*. The latter are defined to be instances of *Ho* that contain no variables. Our language further contains a function *Not* mapping event fluents to event fluents and regular fluents to regular fluents. It occurs in the following two *domain independent axioms* that will be included in any T_{NC} :

$$\mathit{Ho}(g, t) \equiv \neg \mathit{Ho}(\mathit{Not}(g), t) \tag{5}$$

$$\mathit{Not}(\mathit{Not}(g)) = g \tag{6}$$

The restriction of a T_{NC} to its domain independent axioms will be called $T_{\mathrm{NC}}\rfloor_{\mathrm{IND}}$. The effect axioms together with $T_{\mathrm{NC}}\rfloor_{\mathrm{IND}}$ will be called $T_{\mathrm{NC}}\rfloor_{\mathrm{GEN}}$ (the *general axioms* of a T_{NC}). For each T_{NC} , $T_{\mathrm{NC}}\rfloor_{\mathrm{IND}}$ further contains unique-names axioms [22]: for all $F_1, F_2 \in \mathrm{FC}$, $F_1 \neq F_2$ we have:

$$F_1 \neq F_2 \tag{7}$$

For all $E_1, E_2 \in \mathrm{EC}$, $E_1 \neq E_2$ we have:

$$E_1 \neq E_2 \tag{8}$$

2.2 Causal Theories

Causal and non-causal theories share the same domain-independent axioms: $T_{\mathrm{C}}\rfloor_{\mathrm{IND}} = T_{\mathrm{NC}}\rfloor_{\mathrm{IND}} =$ (5)-(8). Thus they are also part of our causal formalization of the shooting domain, $T_{\text{YSP-C}}$. $T_{\text{YSP-C}}$ also contains observation axioms (3) and (4). General axioms (1) and (2) however will be replaced by the following two axioms that make use of an additional predicate $\mathit{Ca}(g, t)$ denoting that g is *Ca*used to hold at time t:

$$\mathit{Ho}(\mathit{Load}, t) \supset \mathit{Ca}(\mathit{Loaded}, t+1) \tag{9}$$

$$\mathit{Ho}(\mathit{Loaded}, t) \wedge \mathit{Ho}(\mathit{Shoot}, t) \supset \mathit{Ca}(\mathit{Not}(\mathit{Alive}), t+1) \tag{10}$$

2.3 On Models

We denote by *Mod*(T) the class of those classical models for any T in which *timepoints* are interpreted as the nonnegative integers. If two models $\mathcal{M}_1$ and $\mathcal{M}_2$ share the same interpretation of *Ho* - which means that they model exactly the same 'history' or 'development' of the domain under consideration - we will say $\mathcal{M}_1$ and $\mathcal{M}_2$ *correspond in terms of Ho*, written as $\mathcal{M}_1 \equiv_{\mathit{Ho}} \mathcal{M}_2$. Formally, $\mathcal{M}_1 \equiv_{\mathit{Ho}} \mathcal{M}_2$ iff for all $\mathbf{g}, \mathbf{t}$ we have $\mathcal{M}_1 \models \mathit{Ho}(\mathbf{g}, \mathbf{t}) \Leftrightarrow \mathcal{M}_2 \models \mathit{Ho}(\mathbf{g}, \mathbf{t})$

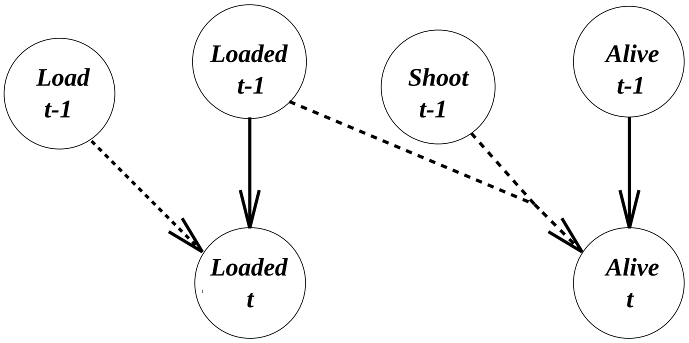

Figure 1: A Yale Shooting Causal Graph

3. Our Two Basic Ideas

The originally proposed method to formalize persistence was to select those models of a theory in which the least changes of fluent values occur [10, 23]. The YSP paper then showed that this method does not work well. One of the first proposed solutions [11, 17] to the YSP was to refine the original model selection criterion to select only the models with the least *uncaused* changes. Indeed, all existing causal approaches [7, 11, 17, 18, 19, 21, 25] implement some variation of this idea. In our own approach we will refine this idea by connecting causation to explanation: informally, if A causes B, then A also provides an explanation for B. But in domains subject to persistence, we may also say that if we have $Ho(f,t)$ for some regular fluent f and nothing happens between t and $t+1$, then this explains $Ho(f,t+1)$ just as well. This leads us to the two key ideas behind our approach:

Basic Principle *From among the classical models, select those with the least (in the subset sense)* unexplained *fluent facts.*

Preliminary Work *We must correctly determine for* each *fluent fact in* each *model whether it is explained or not.*

Intuitively, something is 'explained' in a model if it is normally ('by default') the case given the rest of the model. Thus nonmonotonic model selection becomes an abductive 'search for explanations' consisting of two stages: first, the 'preliminary work' is done to find out what is explained in each model; second, this is used to find the models in which as much as possible is explained, i.e. the models which are the *least surprising*. Now the first stage is where Pearl's account of causation comes in – we use it to do our 'preliminary work'.

4. Pearl's Account of Causation

Here we will only treat Pearl's theory in a very informal manner, specifically geared to enable us to quickly introduce $\mathbf{I}_0$ and $\mathbf{S}_0$. Given a domain with some variables, Pearl's theory uses 'causal graphs' to depict which variables have causal relationships to each other and which *direction* these relationships take. A causal graph G together with a description D of the functional relations between its variables is called a *causal network* (G,D). In order to give a rough idea of how this works, fig. 1 gives a causal graph of the YSP domain. For example, it shows that (performing) a *Load*-action at time $t-1$ has a *causal influence* on (whether we have) *Loaded* at time t. Also, *Loaded* at time $t-1$ has a causal influence, but of a different kind, on *Loaded* at time t. The parents of a node (connected by either dashed or solid arrows) in the network are called its 'direct causes'. Here is the fundamental idea of Pearl's theory [5]:

The Sufficient Cause Principle If an *action* that sets the truth value of a proposition P is performed successfully, then the truth values of all propositions in the causal network that are not descendants of P become independent of P.

In graphical terms this means that we *delete* all arrows of variables going into the node that represents P. In our example: *normally*, $Ho(Loaded, t)$ implies $Ho(Loaded, t-1)$. But if a *Load*-action takes place at $t-1$, then $Ho(Loaded, t)$ does not tell us anything anymore on whether we had $Ho(Loaded, t-1)$. Pearl's theory also tells us that for every causal network (G, D) the functions in D must be such that the following principle holds [6], which for example makes sure that there are no dependencies between variables which are not connected by any path in the graph:

The Causal Independence Principle Once all unexplainable facts and the direct causes of a proposition P are known, the truth value of P becomes independent of all propositions in the causal network that are not descendants of P .

5. Model Selection Criteria $\mathbf{I}_0$ and $\mathbf{S}_0$

In the following, we assume that T_{NC} 's are defined for a language that also contains the *Ca*-predicate and a new predicate *JE* ('*J*ust *E*xplained' - the name will become clear later). But T_{NC} 's themselves never contain axioms about either *Ca* or *JE*.

$\mathbf{I}_0$ is defined on the next page as a five-step procedure that receives as input a theory T_{NC} and that outputs a set of models $\mathbf{I}_0(T_{\mathrm{NC}}) \subset Mod(T_{\mathrm{NC}})$. Steps 1-4 of $\mathbf{I}_0$ implement our 'preliminary work' (section 3). Step 5 uses the results of steps 1-4 to apply our 'basic principle' (section 3). The 'preliminary work' can further be decomposed into two stages: in the first stage (steps 1-3 in definition 1), $\mathbf{I}_0$ will *build a complete causal network for* T_{NC} . It will do this by using Pearl's theory and some precisely stated additional assumptions about the 'physics' of the domain under consideration. This is meant in the following sense: given theory T_{NC} and the additional assumptions, there is only one specific causal network for T_{NC} which does not contradict the properties that causal networks must have according to Pearl's theory. In the second stage (step 4 in the definition), $\mathbf{I}_0$ will apply Pearl's semantics to the causal network found to determine for each fluent fact in each model whether it is explained or not.

This means that steps 1-3 (i.e. the first stage) should not rule out any particular interpretation of *Ho*; rather, they will determine for each interpretation of *Ho* a single interpretation of *Ca* and *JE* belonging to it. This interpretation of *Ca* and *JE* *encodes* a causal graph in the following sense: if, after performing the first stage of $\mathbf{I}_0$, we have $Ca(\mathbf{g}, \mathbf{t})$ in a model this will mean that there is a *sufficient* cause for the fluent $\mathbf{g}$ to hold at time $\mathbf{t}$ in the model (i.e. there is an *action* that successfully makes $\mathbf{g}$ true); in graphical terms, there is a *dashed* arrow pointing into node $(\mathbf{g}, \mathbf{t})$. $JE(\mathbf{g}, \mathbf{t})$ in a model will mean that there is a causal influence on $\mathbf{g}$ which, *if no intervention takes place* would *normally* make it hold at time t; graphically, a solid arrow points into $(\mathbf{g}, \mathbf{t})$. We will now go through the five steps of definition 1 in detail, using the YSP as an example as we go along .

Step 1 Here we determine the causal relations that hold in T_{NC} . This is the most complicated step by far, and for the time being we will content ourselves by explaining the *outcome* of this step: after step 1, a model $\mathcal{M}$ with $\mathcal{M} \models Ca(\mathbf{g}, \mathbf{t})$ will be selected if there is a sufficient cause within the model for $\mathbf{g}$ to hold at $\mathbf{t}$. For example, for $T_{\mathrm{YSP\text{-}NC}}$, all models remaining after step 1 will have $Ca(Loaded, 1)$. Also, among all models for $T_{\mathrm{YSP\text{-}NC}}$ with $Ho(Loaded, 2)$ all those selected in step 1 will also have $Ca(Not(Alive), 3)$. A detailed explanation of how step 1 works follows in section 7 .

Step 2 introduces persistence; if something holds at time t, there will be an explanation for it to hold at time $t+1$ (condition 1). Condition 2 expresses that if something holds at time $t = 0$, our initial point in time, then it is explained *anyway*, independent of any other fluent fact. Condition 3 makes sure that 'normally we do not expect events to happen'. For example, in $T_{\mathrm{YSP\text{-}NC}}$, all models further selected will have $JE(Not(Shoot), 2)$. Among all models for $T_{\text{YSP-NC}}$ with $Ho(Alive, 2)$, those further selected will also have $JE(Alive, 3)$.

Step 3 We have determined in step 1-2 all places in the causal graph where there *must* be a causal influence and we *must* have an arrow. Step 3 embodies a *domain closure* of *Ca* and *JE*. It effectively minimizes the interpretations of *Ca* and *JE* for every interpretation of *Ho*, while it does not rule out any interpretation of *Ho* itself. This closure follows from the *physical assumption* that $T_{\text{NC}}\rfloor_{\text{GEN}}$ and the persistence assumption of step 2 of $\mathbf{I}_0$ are *all* the laws of nature to which our domain is subjected. Also, this step forces the functional relations between the variables in the domain to obey Pearl's *principle of causal independence* (sec.4)!

In $T_{\text{YSP-NC}}$, all models further selected will have $\forall t.\neg Ca(Alive, t)$. Also, among those models in $\text{I}_{0,2}$ that have $Ho(Alive, 2)$, all those further selected will also have $\neg JE\ (Not(Alive))\ , 3)$.

Step 4 We now use Pearl's *sufficient cause principle* to determine for each model $\mathcal{M}$ the set of unexplained fluent facts $\mathbf{Ab}^1_g(\mathcal{M}), \mathbf{Ab}^2_f(\mathcal{M}), \mathbf{Ab}^2_e(M)$ - first, a fluent fact $(\mathbf{g}, \mathbf{t})$ in a model $\mathcal{M}$ is unexplained (i.e. $(\mathbf{g}, \mathbf{t}) \in \mathbf{Ab}^1_g(\mathcal{M})$) iff there is a sufficient cause for $(\mathbf{g}, \mathbf{t})$ in the model, while $(\mathbf{g}, \mathbf{t})$ does not hold; i.e. an action has not had its regular effect. *Regular* fluent fact $(\mathbf{f}, \mathbf{t})$ is in $\mathbf{Ab}^2_f(\mathcal{M})$ if a) there is an explanation for it not to hold, b) there is no *sufficient* cause for it to hold (i.e. the sufficient cause principle does not apply), but c) in fact, it does hold. This means there is an unexplained *breach of persistence*. *Event* fluent fact $(\mathbf{e}, \mathbf{t})$ is in $\mathbf{Ab}^2_e(\mathcal{M})$ if event $\mathbf{e}$ takes place at time $\mathbf{t}$ without being caused by something else in the model. *Ca* always overrules *JE*: if we have $JE(Alive, 3)$ but $Ca(Not(Alive), 3)$ in all models, we prefer models with $Ho(Not(Alive), 3)$: *Alive* was 'just' explained, *Not*(*Alive*) had a *sufficient* cause.

The intended model $\mathcal{M}$ of $T_{\text{YSP-NC}}$ will have empty $\mathbf{Ab}^1_g(\mathcal{M})$ and empty $\mathbf{Ab}^2_f(\mathcal{M})$, while $\mathbf{Ab}^2_e(\mathcal{M}) = \{(Load, 0), (Wait, 1), (Shoot, 2)\}$. The models $\mathcal{M}'$ in which the gun gets unloaded at $t = 1$ will all have $(Not(Loaded), 2) \in \mathbf{Ab}^2_f(\mathcal{M}')$.

Step 5 We first select the models with the least violations of causal laws; we then further select those with the least breaches of persistence. Finally, we rule out as many unexplained events as possible. We perform this step as the latest because we regard unexplained events as less 'surprising' than unexplained breaches of persistence. In all selected models for $T_{\text{YSP-NC}}$, we have $\neg Ho(Alive, 3)$, so we solve the YSP.

5.1 And what about $\mathbf{S}_0$*?*

Reconsider a causal theory such as our $T_{\text{YSP-C}}$. From a Pearlian viewpoint, we can see that what is actually formalized by axioms like (9) and (10) are the *sufficient causes* for fluents to hold at some time. This means that what $\mathbf{S}_0$ really must do is to perform steps 2, 3, 4 and 5 of $\mathbf{I}_0$ as step 1 has already been done by the person who formalized the domain knowledge! A formal definition can be found in the box (def. 2).

6. More About $\mathbf{S}_0$

Here we will briefly show how $\mathbf{S}_0$ handles some of the well-known problems in nonmonotonic temporal reasoning: *ramifications* and *actions with non-deterministic effects.* We will see in section 9.1 how $\mathbf{S}_0$ handles causal chains of events.

6.1 Ramifications

We may enhance the class of reasoning domains expressible in our language by introducing another kind of axiom: *constraint axioms* that impose further restrictions on what may hold at what time in our domains. As first noticed by Ginsberg, [16] sometimes such a constraint axiom is meant to yield indirect effects (ramifications) while sometimes it should yield action preconditions (qualifications).

Definition 1 *The model selection criterion* $\mathbf{I}_0$ *from theories* T_{NC} *to sets of models* $\mathrm{M} \subseteq$ *$Mod(T_{\mathrm{NC}})$ is defined by the following procedure:*

Step 1 Perform the following two sub-steps:

1. First derive 'what causes what' in T_{NC} using the procedure of definition 3 (next page).
2. Now select a model $\mathcal{M} \in Mod(T_{\mathrm{NC}})$ iff for all ϕ, ψ, Φ (as in def. 3) we have that

   if $\mathcal{M} \models \Phi \wedge \phi$
   and ϕ **s–causes** ψ **under circumstances** Φ
   then $\mathcal{M} \models Ca(\mathbf{g}_\psi, \mathbf{t}_\psi)$

   Let $\mathrm{I}_{0,1}$ be the set of models thus selected.

Step 2 Among the remaining models $\mathcal{M} \in \mathrm{I}_{0,1}$, we select exactly those $\mathcal{M}$ with
1) $\mathcal{M} \models \forall f, t.\ Ho(f,t) \supset JE(f, t+1)$ and
2) $\mathcal{M} \models \forall f.\ JE(f,0) \equiv Ho(f,0)$ and
3) $\mathcal{M} \models \forall e, t.\ JE(Not(e), t)$.
Let the set of resulting $\mathcal{M}$ be $\mathrm{I}_{0,2}$.

Step 3 Further select any $\mathcal{M} \in \mathrm{I}_{0,2}$ iff $\mathcal{M}$ has a minimal extension of Ca within
$\{\mathcal{M}' \mid \mathcal{M}' \in \mathrm{I}_{0,2}$ and $\mathcal{M} \equiv_{Ho} \mathcal{M}'\}$
and $\mathcal{M}$ has a minimal extension of JE within
$\{\mathcal{M}' \mid \mathcal{M}' \in \mathrm{I}_{0,2}$ and $\mathcal{M} \equiv_{Ho} \mathcal{M}'\}$

Let the set of remaining models be $\mathrm{I}_{0,3}$.

Step 4 We determine for each model $\mathcal{M} \in \mathrm{I}_{0,3}$ the set of unexplained fluent facts:

$$\begin{aligned}
\mathbf{Ab}^1_g(\mathcal{M}) &= \{(\mathbf{g},\mathbf{t}) | \mathcal{M} \models \neg Ho(\mathbf{g},\mathbf{t}) \wedge Ca(\mathbf{g},\mathbf{t})\}\\
\mathbf{Ab}^2_f(\mathcal{M}) &= \{(\mathbf{f},\mathbf{t}) | \mathcal{M} \models JE(Not(\mathbf{f}),\mathbf{t}) \wedge Ho(\mathbf{f},\mathbf{t}) \wedge \neg Ca(\mathbf{f},\mathbf{t})\}\\
\mathbf{Ab}^2_e(\mathcal{M}) &= \{(\mathbf{e},\mathbf{t}) | \mathcal{M} \models JE(Not(\mathbf{e}),\mathbf{t}) \wedge Ho(\mathbf{e},\mathbf{t}) \wedge \neg Ca(\mathbf{e},\mathbf{t})\}
\end{aligned}$$

Step 5 Perform the following steps in order:

1. Select any $\mathcal{M} \in \mathrm{I}_{0,3}$ iff $\mathbf{Ab}^1_g(\mathcal{M})$ is minimal over $\mathrm{I}_{0,3}$, i.e.. for no $\mathcal{M}' \in \mathrm{I}_{0,3}$, $\mathbf{Ab}^1_g(\mathcal{M}')$ is a proper subset of $\mathbf{Ab}^1_g(\mathcal{M})$. Let the set of remaining models be $\mathrm{I}_{0,4}$.
2. Select any $\mathcal{M} \in \mathrm{I}_{0,4}$ iff $\mathbf{Ab}^2_f(\mathcal{M})$ is minimal over $\mathrm{I}_{0,4}$. Let the set of remaining $\mathcal{M}$ be $\mathrm{I}_{0,5}$.
3. Select any $\mathcal{M} \in \mathrm{I}_{0,5}$ iff $\mathbf{Ab}^2_e(\mathcal{M})$ is minimal over $\mathrm{I}_{0,5}$.

$\mathbf{I}_0(T_{\mathrm{NC}})$ *is now defined to be the set of models remaining after step 5.3.*

Definition 2 $\mathbf{S}_0$ *is the procedure that results from setting* $\mathrm{I}_{0,1} := Mod(T_{\mathrm{C}})$ *and jumping into the procedure of* $\mathbf{I}_0$ *at the beginning of step 2.*

We will now give an example (inspired by McCain & Turner [21]). Let $T_{\mathrm{YSP\text{-}CR}}$ be the union of $T_{\mathrm{YSP\text{-}C}}$ and the following two axioms:

$$\neg Ho(Alive, t) \supset Ca(Not(Walking), t) \tag{11}$$

$$Ho(Get\_up, t) \supset Ca(Walking, t+1) \tag{12}$$

Here (12) is another effect axiom, while (11) is a constraint axiom. (11) makes sure that in all selected models for the $T_{\mathrm{YSP\text{-}CR}}$ we have $Ho(Walking, t) \supset Ho(Alive, t)$ while we do *not* necessarily have $Ho(\ Walking, t) \supset Ca(Alive, t)$. This means that $Ho(Not(Alive), \mathbf{t})$ would *block* performing *Get_up* at time $\mathbf{t}$: the reader may check that any $\mathcal{M}$ with $Ho(Get\_up, \mathbf{t})\ \wedge Ho(Not(Alive), \mathbf{t})$ would have the abnormality $(Alive, \mathbf{t}+1) \in \mathbf{Ab}^2_f(\mathcal{M})$. On the other hand, $Ho(Walking, \mathbf{t})$ can *never* block performing a *Shoot*-action at time $\mathbf{t}$: we have models M for $T_{\mathrm{YSP\text{-}CR}}$ with $\mathcal{M} \models Ho(Walking, \mathbf{t}) \wedge Ho(Alive, \mathbf{t}) \wedge Ho(Shoot, \mathbf{t})$ but without any abnormalities.

It is easy to show that $\mathbf{S}_0$ also deals properly with 'pure' ramification and qualification constraints and with ramification constraints involving several preconditions (as in the 'Emperor Problem' and the 'Two Switches Problem' [21]).

6.2 Nondeterministic Effects

If we toss a coin, we would like to infer that after the *Toss* event, we can have either *Heads* or *Not*(*Heads*), independent of whether we had *Heads* or *Not*(*Heads*) before the tossing (thus there is no persistence). This can be expressed by the following axiom :

$$Ho(Toss, t) \supset [Ca(Heads, t+1) \equiv \neg Ca(Not(Heads), t+1)]$$

In this way, if there is *Toss*-event at time $\mathbf{t}$, then in all selected models there will either be a cause for *Heads* but not *Not*(*Heads*) to hold, or for *Not*(*Heads*) but not *Heads*, but there never is a cause for both (which would inevitably lead to an abnormality) or neither (which would let the old value of *Heads* persist). Note that this is a way of directly representing nondeterministic effects and thus considerably simpler than most earlier approaches to this problem [29], in which *Toss* is interpreted as a set of deterministic actions $Toss_1$ with effect *Heads* and $Toss_2$ with effect *Not*(*Heads*) .

7. More About $\mathbf{I}_0$

In this section we explain step 1 of $\mathbf{I}_0$ in more detail. We need to find out for each $(\mathbf{g}, \mathbf{t})$ in each model whether there is a sufficient cause for it in the same model. We will do this by first generating the *causal laws* that must minimally hold in all models of our theory T_{NC} . To find them, we need to temporarily restrict it to $T_{\mathrm{NC}}\rfloor_{\mathrm{GEN}}$ – we will see that we must not let our *general* causal relations be influenced by *contingent* observations! We see in definition 3 that we can infer
ϕ **s–causes** ψ **under circumstances** Φ (**s-causes** is to be read as 'is a sufficient cause of') if, given that Φ holds, ϕ *always* implies ψ (step 1 of definition 3) and (step 2), that ψ is actually *brought about* by ϕ. Step 3 in definition 3 tells us that the causal influence only goes one way. Steps 1, 2 and 3 are necessary if we want our causal relations to stand for causal networks in Pearl's sense – steps 1 and 2 follow from the sufficient cause principle, step 3 from the directedness of the causal graph! We can never be sure that ϕ is a sufficient cause for ψ by step 1 alone; ϕ could then still 'cause' ψ in the models with Ψ if Ψ implied both ϕ and ψ. Step 2 rules out this possibility. Step 4 adds another assumption about the physics of our domains, which simply says that performing an action at time $\mathbf{t}$ can never change anything about the world at an earlier point in time $\mathbf{t}'$.

For example, in $T_{\mathrm{YSP\text{-}NC}}$ we will get '$Ho(Shoot, 2)$ **s-causes** $Ho(Not(Alive)\ , 3)$ **u.c.** $Ho(Loaded, 2)$'. Now consider what would have happened if we had not restricted ourselves to $T_{\mathrm{NC}}\rfloor_{\mathrm{GEN}}$: we would have had $Ho(Shoot, 2)$ in *all* models; therefore we could never have checked whether it is really the *shooting* that brought about *Not*(*Alive*). For this, we need to be able to look at a counterfactual model $\mathcal{M}_c$ where *no* shooting takes place at time 2! We must then check whether a) we can have $Ho(Alive, 3)$ in $\mathcal{M}_c$ (step 2), and b) whether, in *all* models with $Ho(Shoot, 2)$ and $Ho(Loaded, 2)$, we must have $Ho(Not(Alive), 3)$ (step 1) .

Definition 3 *Suppose a theory* T_{NC} *is given. Let* $\mathrm{M}_{\mathrm{GEN}} = \mathit{Mod}(T_{\mathrm{NC}}\rfloor_{\mathrm{GEN}})$. *Let* $\phi = \mathit{Ho}(\mathbf{e}_\phi, \mathbf{t}_\phi)$, $\psi = \mathit{Ho}(\mathbf{g}_\psi, \mathbf{t}_\psi)$ *be any (event and generalized, respectively) fluent facts, and* Φ *be any propositional combination of fluent facts. We say that* ϕ **s–causes** ψ **under circumstances** Φ *iff all of the following hold :*

1. *For all* $\mathcal{M} \in \mathrm{M}_{\mathrm{GEN}}$: $\quad \mathcal{M} \models \Phi \Rightarrow \mathcal{M} \models \phi \supset \psi$.
2. *There is an* $\mathcal{M} \in \mathrm{M}_{\mathrm{GEN}}$ *with* $\mathcal{M} \models \Phi \wedge \phi$.
   There also exists a 'counterfactual model' $\mathcal{M}_c \in \mathrm{M}_{\mathrm{GEN}}$ *of* $\mathcal{M}$ *w.r.t.* ϕ, ψ *and* Φ,
   i.e. $\mathcal{M}_c \models \Phi \wedge \neg\phi \wedge \neg\psi$.
3. *There exists an* $\mathcal{M} \in \mathrm{M}_{\mathrm{GEN}}$ *with* $\mathcal{M} \models \Phi \wedge \neg\phi \wedge \psi$.
4. *For all time points* $\mathbf{t}_\Phi$ *occurring in* Φ*:* $\mathbf{t}_\Phi \leq \mathbf{t}_\phi < \mathbf{t}_\psi$.

8. Comparisons of $\mathbf{S}_0$ and $\mathbf{I}_0$ - the Harvest

We will now compare $\mathbf{S}_0$ and $\mathbf{I}_0$ to various existing approaches. We first compare $\mathbf{S}_0$ to existing causal and Action-Description-Language (ADL) approaches (section (9). In section 10 we compare $\mathbf{I}_0$ to two non-causal approaches, i.e. Baker's method and 'chronological minimization'.

We feature two kinds of comparisons. The first kind is the 'equivalence proof'. We have defined several '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. All approaches we considered have as their basic objects regular 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 generalized fluents hold at what time iff B selects a model M_B with the same interpretation of 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 second kind of comparison will always be an example of a simple reasoning domain for which we claim that our approach works better than the one we are comparing it to.

The General Pattern For all our examples it turns out that if an approach behaves badly on them, this is always either because our 'basic principle' is violated (the approach does not or not only select the least surprising models) or because our 'preliminary work' is not done well (what is explained and what is not is determined in a way not consistent with causal network theory).

9. $\mathbf{S}_0$ and the others

9.1 Causal Model Selection Criteria

MAT The philosophy behind Motivated Action Theory (MAT) [25] is quite similar to ours: while they prefer models with the least number of 'unmotivated actions', we prefer those with the least number of 'uncaused events'. However, if one looks at the formal definitions, it turns out that in the MAT model preference criterion a model $\mathcal{M}_1$ is preferable over another model $\mathcal{M}_2$ if all events occurring in $\mathcal{M}_1$ *that do not occur in* $\mathcal{M}_2$ are motivated (i.e. have a cause within the model). This means that if in two models the same events happen at the same time, the MAT criterion does not reject the one in which there is less 'motivation' for these events in terms of (regular) fluent values, and thus *MAT violates our basic principle.* This is illustrated in the following example for a reasoning domain involving persons A and B. Person A can be *Steady* or not, can *Fall* or not, and can lie

On_Floor or not. Person B can *Push* or not. We present the domain axioms in our language here but the translation into MAT is straightforward.

$$Ho(Fall, t) \supset Ca(On\_Floor, t+1)$$
$$[\neg Ho(Steady, t) \wedge Ho(Push, t)] \supset Ca(Fall, t+1)$$
$$Ho(Push, 0) \wedge \neg Ho(On\_Floor, 1) \wedge Ho(On\_Floor, 2)$$

In this case, both approaches will yield the models in which $\neg Ho(Steady, 0) \wedge Ho(Fall, 1)$. But whereas for our approach these are the only ones, MAT also selects the models with $Ho(Steady, 0) \wedge Ho(Fall, 1)$!

We add here without going into detail that **Geffner's causal model selection** [7] also gives the unintended model in the example above.

9.2 ADL Approaches

From our point of view, both the recent ADL approaches and Sandewall's 'systematic' approach [28] are to be classified among the causal approaches, since they too, are based on explicitly distinguishing between what is caused to hold and what actually holds in a model (though the *word* 'caused' is not always employed). Next we compare $\mathbf{S}_0$ to two ADL approaches: Baral and Gelfond's $\mathcal{L}_3$ [3] and McCain and Turner's [21] Theory of Ramifications and Qualifications. $\mathcal{L}_3$ is an extension of the well-known language $\mathcal{A}$ [8] that can deal with concurrent actions, incomplete specification of actions and incomplete knowledge on the order of the observations in time; it cannot at all deal with ramifications. McCain and Turner (MT)'s approach however is specifically geared to handle these, but cannot handle the other domains mentioned for $\mathcal{L}_3$. Our comparisons thus cover a rather wide area of problem domains.

McCain and Turner's approach and $\mathbf{S}_0$

MT consider theories that are triplets (S, E, C), defined for a propositional language where each atom stands for a (regular) fluent. The 'state of the world' S is an interpretation for all fluents, denoted by a maximal consistent set of literals. 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 determination relations' that determine the ramifications of effects. MT define a function $Res^4_C(E, S)$ such that it gives the set of possible states of the world after an action with effects E has taken place in state S. For the definition of $Res^4_C(E, S)$ we refer to [21]; here we just give an example of its use :

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

In this case, $Res^4_C(E, S) = \{\{\neg Alive, \neg Walking\}\}$ 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 $Res^4_{C'}(E', S')$ would have been empty: '$\Rightarrow$' has a function similar to our '*Ca*', enabling changes of right-hand side fluent values given changes of left-hand side values, but *not* the other way around (compare this to section 6.1). 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 (i.e. something which is *Ca*used) may be a disjunction of two fluents. This cannot be expressed by theories T_{C} (notice that a caused *disjunction* is not the same as a disjunction of instances of *Ca*). But it is exactly here that MT can give counterintuitive results; to see this, consider the following domain about a person that may or may not have a) the flue, b) a headache, c) nausea. If she catches the flue, she will also get either a headache or nausea. Suppose that two events happen, one causing the flue and the other causing a headache :

$$
\begin{aligned}
S_2 &= \{\neg \mathit{Flue}, \neg \mathit{Headache}, \neg \mathit{Nausea}\} \\
E_2 &= \{\mathit{Flue}, \mathit{Headache}\} \\
C_2 &= \{\mathit{Flue} \Rightarrow (\mathit{Headache} \vee \mathit{Nausea})\}
\end{aligned}
$$

Proposition 1 $Res^4_{C_2}(E_2, S_2) = \{\{\mathit{Flue}, \mathit{Headache}, \neg \mathit{Nausea}\}\}$

Proof: Follows directly from definition 4 in [21]. □

We have the rather unintuitive conclusion that our patient is guaranteed not to get any nausea! Assuming that an additional event happened which directly causes a headache allows us to rule out the possibility of the person getting nauseous. We would say that our *preliminary work* (section 3) is violated here - according to the sufficient cause principle of causal network theory, an effect causing $\mathit{Headache} \vee \mathit{Nausea}$ would mean that the value of '$\mathit{Headache} \vee \mathit{Nausea}$' would become true, and also, more importantly, the fact whether we have *Headache* or *Nausea* after the effect would *not depend anymore* on whether *Headache* and/or *Nausea* held before the effect. The problem with the example above then cannot occur any more. We have not extended our own approach to represent causes of disjunctions as it does not seem to be of too much practical importance.

Apart from causes of disjunctions, it turns out that MT and $\mathbf{S}_0$ agree on all problem domains that can be represented in the languages of both approaches. In definitions 4 and 5 syntactic ($\sim$) and semantical ($\cong$) correspondence for $\mathbf{S}_0$ and MT are defined. $\sim$ has been defined such that causes of disjunctions cannot occur in corresponding theories. $\cong$ is defined such that $\mathcal{M} \cong (S, S')$ if 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. Here is our equivalence theorem :

Theorem 1 *If we have any theory T_{C} and any domain description $T_{\mathrm{MT}} = (S, E, C)$ such that a) $T_{\mathrm{C}} \sim T_{\mathrm{MT}}$ and b) there is an $\mathcal{M} \models T_{\mathrm{C}}$ with $\mathbf{Ab}^1_g(\mathcal{M}) = \mathbf{Ab}^2_f(\mathcal{M}) = \emptyset$ then*

1. $\mathcal{M} \in \mathbf{S}_0(T_{\mathrm{C}}) \Rightarrow Res^4_C(E, S) = \{S'\},\ M \cong (S, S')$
2. $S' \in Res^4_C(E, S) \Rightarrow$ *there exists an* $\mathcal{M}$ *with*
   $\mathcal{M} \in \mathbf{S}_0(T_{\mathrm{C}}), \mathcal{M} \cong (S, S')$.

Baral and Gelfond's $\mathcal{L}_3$ and $\mathbf{S}_0$

Baral and Gelfond's (BG) approach [3, 2] which is based on their ADL $\mathcal{L}_3$, consists of *domain descriptions D* written in $\mathcal{L}_3$. We compared our theories T_{C} to domain descriptions D in exactly the same way as we compared them to theories T_{MT} of MT's approach: we will first give an example of a reasoning domain where $\mathbf{S}_0$ 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:

Review of $\mathcal{L}_3$ Baral and Gelfond'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 generalized fluents hold at what time. $\mathcal{L}_3$ consists of the sets of symbols $\mathcal{F}$ (fluents), $\mathcal{A}$ (unit actions) and $\mathcal{S}$ (situations). $\mathcal{S}$ 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

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

We denote by $\mathbf{f}$ a fluent literal in a T_{C} . Whenever we write $\mathbf{f}'$, we mean the corresponding literal in MT's language: if $\mathbf{f} = \mathit{Not}(\mathbf{f}''), \mathbf{f}'' \in \mathrm{FC}$, then $\mathbf{f}' = \neg\mathbf{f}''$; otherwise, $\mathbf{f}' = \mathbf{f}$.

Definition 4 *For any T_{C} of our approach and $T_{\mathrm{MT}} = (S, E, C)$ of MT's approach we define $T_{\mathrm{C}} \sim T_{\mathrm{MT}}$ ('T_{C} corresponds to T_{MT} ') to be true iff all of the following hold:*

1. $\mathrm{EC} = \emptyset$; $\mathrm{TNC} = \emptyset$; *$F_i \in \mathrm{FC}$ iff F_i is an atom in the propositional language for which T_{MT} is defined .*

2. *$T_{\mathrm{C}}\rfloor_{\mathrm{IND}}$ consists of axioms* (5)-(8) *.*

3. *T_{C} contains the axiom $\mathit{Ho}(\mathbf{f}_1, 0) \wedge \ldots \mathit{Ho}(\mathbf{f}_{|\mathrm{FC}|}, 0)$ iff $S = (\mathbf{f}'_1, \ldots, \mathbf{f}'_{|\mathrm{FC}|})$.*
   Here $\mathrm{FC} = \{F_1, \ldots, F_{|\mathrm{FC}|}\}$; each $\mathbf{f}_i$ stands for either F_i or $\mathit{Not}(F_i)$.

4. *T_{C} contains an axiom of the form*
   $$\mathit{Ca}(\mathbf{f}_1, 1) \wedge \ldots \wedge \mathit{Ca}(\mathbf{f}_n, 1) \qquad (n \geq 1)$$
   iff E contains formula $\mathbf{f}'_1 \wedge \ldots \wedge \mathbf{f}'_n$. Here each $\mathbf{f}_i$ can stand for any F or $\mathit{Not}(F)$ with $F \in \mathrm{FC}$.

5. *T_{C} contains an axiom of the form $(n \geq 1, m \geq 1)$*
   $$\mathbf{H}(\mathbf{f}_1, \ldots, \mathbf{f}_n, 1) \supset \mathit{Ca}(\mathbf{f}_{n+1}, 1) \wedge \ldots \wedge \mathit{Ca}(\mathbf{f}_{n+m}, 1)$$
   iff C contains formula $\mathbf{H}'(\mathbf{f}_1, \ldots, \mathbf{f}_n) \Rightarrow \mathbf{f}'_{n+1} \wedge \ldots \wedge \mathbf{f}'_{n+m}$.
   Here each $\mathbf{f}_i$ can stand for any F or $\mathit{Not}(F)$ with $F \in \mathrm{FC}$. $\mathbf{H}(\mathbf{f}_1, \ldots, \mathbf{f}_n, 1)$ is a propositional combination of fluent facts, each of the form $\mathit{Ho}(\mathbf{f}_i, 1)$ with $1 \leq i \leq n$. $\mathbf{H}'(\mathbf{f}_1, \ldots, \mathbf{f}_n)$ is the same propositional formula but with each $\mathit{Ho}(\mathbf{f}_i, 1)$ replaced by $\mathbf{f}'_i$.

6. *All axioms in T_{C} are of one of the forms occurring in conditions 2-5 above. All formulas in E are of the form occurring in condition 4 and all formulas in C are of the form occurring in condition 5.*

Definition 5 *Given a model $\mathcal{M}_C$ for a T_{C} and a pair of states (S, S') for a T_{MT} we say that $\mathcal{M}_C$ corresponds to (S, S') iff for all $\mathbf{f} \in \mathrm{FC}$ we have*

$$\mathcal{M}_C \models \mathit{Ho}(\mathbf{f}, 0) \Leftrightarrow \mathbf{f}' \in S \quad \textit{and}\ \mathcal{M}_C \models \mathit{Ho}(\mathbf{f}, 1) \Leftrightarrow \mathbf{f}' \in S'$$

where a is a compound action and $f, p_1, \ldots, \_n$ $(n \geq 0)$ 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 $bf occurs\_at$ 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[1].

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 S to sequences of actions, such that 1) $\Sigma(s_0) = []$ ($[]$ denotes the empty sequence) and 2) for every $s_i \in S$, $\Sigma(s_i)$ is a prefix of $\Sigma(s_n)$. 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.

If an interpretation M is consistent with the description D, it is called a *model* of D. For a more detailed definition of modelhood and for illustrating examples we refer to Baral and Gelfond's article [3].

A Distinguishing Example For most domain descriptions D, the models M for D correspond to the models $\mathbf{S}_0$ selects for the theory T_{C} corresponding to D (Notice again that in $\mathcal{L}_3$ we can only express a small subset of what is expressible in our language). There is only one class of domains where BG and $\mathbf{S}_0$ give different results, and again we claim that for this class, BG gives the less intuitive ones. We are talking about domains involving *specificity*, which can best be illustrated using the following example: suppose that if you lift a bowl of soup with either your left or your right hand, but not both, then you spill the soup and the table gets wet. If you lift the bowl with both hands however, then you do not spill the soup (this is in fact the standard example illustrating the need for 'specificity' [3] – see below). BG formalize this using the following domain description D:

$$\{\mathit{Lift\_left}\} \quad \textbf{causes} \quad \mathit{Wet} \tag{13}$$

$$\{\mathit{Lift\_right}\} \quad \textbf{causes} \quad \mathit{Wet} \tag{14}$$

$$\{\mathit{Lift\_left}, \mathit{Lift\_right}\} \quad \textbf{causes} \quad \neg\mathit{Wet} \ \textbf{if} \ \neg\mathit{Wet} \tag{15}$$

BG then use the rule of *specificity* in determining the models for their domain description, which says that 'more specific information about actions overrides less specific information': if the preconditions of (15) hold, then rules (13) and (14) are ignored when determining the effects of *Lift_left* and *Lift_right.* 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:

$$\mathit{Turn\_on\_sprinkler} \ \textbf{causes} \quad \mathit{Wet} \tag{16}$$

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 is undefined: suppose we have the additional facts

$$\neg\mathit{Wet} \ \textbf{at} \ s_0 \quad \wedge \quad [\{\mathit{Turn\_on\_sprinkler}, \mathit{Lift\_left}, \mathit{Lift\_right}\}] \ \textbf{occurs\_at} \ s_0 \tag{17}$$

[1] 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.

In the following, we will sometimes refer to fluent literals p which are syntactical objects of the language $\mathcal{L}_3$. A fluent literal can be any p such that $p \in \{f \mid f \in \mathcal{F}\} \cup \{\neg f \mid f \in \mathcal{F}\}$. Whenever we write p', we will mean a fluent literal for our language: $p' = Not(f)$ if $p = \neg f, f \in \mathcal{F}$; $p' = p$ otherwise.

Definition 6 *For any theory T_{C} in our language 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) $a \in \mathrm{EC}$ iff $a \in \mathcal{A}$; 2)) $f \in \mathrm{FC}$ iff $f \in \mathcal{F}$; and 3) $s \in \mathrm{TNC}$ iff $s \in \mathcal{S}$.*

2. d.i. axioms *$T_{\mathrm{C}}\rfloor_{\mathrm{IND}}$ consists of all domain independent axioms* (5)-(8) *plus the additional axiom $s_0 = 0$.*

3. effect laws *D contains an expression of the form '$\{a^1, \ldots, a^n\}$* **causes** *f* **if** *$p_1, \ldots, p_m$' iff T_{C} contains the following axiom:*

$$\begin{aligned} \forall t \ . \ [\quad & Ho(p_1', t) \wedge \ldots \wedge Ho(p_m', t) \ \wedge \\ & Ho_{ev}(a^1, t) \wedge \ldots \wedge Ho_{ev}(a^n, t) \ ] \quad \supset Ho(f', t+1) \end{aligned} \tag{BG18}$$

Notice that t is an object language variable while the p_i', f' and a^i are meta variables.

4. fluent facts *D contains an expression of the form 'f* **at** *s', iff T_{C} contains the axiom '$Ho(f', s)$'.*

5.occurrence facts *D contains an expression of the form '$[a_1, \ldots, a_n]$* **occurs_at** *s', where $n > 0$, $a_i = a_{i1} \mid \ldots \mid a_{im_i}$, $a_{ij} = \{a_{ij}^1, \ldots, a_{ij}^{k(i,j)}\}$, $a_{ij}^{k(i,j)} \in \mathcal{A}$ iff T_{C} contains the following axiom:*

$$\begin{array}{rcl} \mathbf{H}(a_{11}, s) \vee & \ldots & \vee\ \mathbf{H}(a_{1m_1}, s) \ \vee \\ \mathbf{H}(a_{21}, s+1) \vee & \ldots & \vee\ \mathbf{H}(a_{2m_2}, s+1) \ \vee \\ & \vdots & \\ \mathbf{H}(a_{n1}, s+n-1) \vee & \ldots & \vee\ \mathbf{H}(a_{nm_n}, s+n-1) \end{array} \tag{BG19}$$

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

6. precedence facts *D contains an expression of the form 's_1* **precedes** *s_2', iff T_{C} contains the axiom '$s_1 < s_2$'.*

7. facts *A* fact *in $\mathcal{L}_3$ is a propositional combination of the three sorts of 'atomic' facts mentioned above. D contains such a propositional combination iff T_{C} contains the corresponding propositional combination of the translations of the constituent atomic facts.*

8. closure *All axioms in T_{C} are either in $T_{\mathrm{C}}\rfloor_{\mathrm{IND}}$ or a propositional combination of constituents that can be either of the form (BG18) or of the form $Ho(f, s)$ or of the form (BG19) or of the form $s_1 < s_2$.*

Consider an interpretation $M = (\Psi, \Sigma)$ of D in $\mathcal{L}_3$. $\Sigma(s_N)$ can be written as a sequence of compound actions:

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

Notice that for all s occurring in D, $\Sigma(s)$ is a proper prefix of $\Sigma(s_n)$ and specifically $\Sigma(s_0) = []$. We now define $A(t)$ to be the t-th element of $\Sigma(s_N)$ for $t \leq last$ and to be the empty set for $t > last$. Similarly, we can write $\Psi(\Sigma(s_N))$ as follows:

$$\Psi(\Sigma(s_n)) = [\sigma_0, \sigma_1, \ldots, \sigma_{last}]$$

We define $F(t)$ to be the t-th element of $\Psi(\Sigma(s_n))$ for $t \leq last$ while for $t > last$, $F(t) = F(last)$.

Definition 7 *Suppose that $T_{\mathrm{C}} \sim D$ for some T_{C} and D. A model $\mathcal{M}$ for T_{C}* corresponds *to a model M of D (written as $\mathcal{M} \cong M$) iff the following three conditions hold for all a, f and s mentioned in D and all $t \in Time\_Point\_Symbols$ (which, on the left hand sides, are interpreted as plain integers):*

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

Proposition 2 *There is no model M for $D' = D \cup (16) \cup (17)$.*

Proof: Follows directly from definition 3.7 in [3]. □

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 current version of $\mathbf{S}_0$. The real problem lies in the inappropriate use of **causes** in (15): there is no *cause* for not getting wet if you lift the soup bowl with two hands. Again, our *preliminary work* (Pearl's semantics) is performed incorrectly, and we arrive at counterintuitive results.

Because we do not feature specificity, we would have to formalize (13) as

$$Ho(Lift\_left, t) \wedge \neg Ho(Lift\_right, t) \supset Ho(Wet, t+1) \tag{20}$$

and (14) accordingly. (15) would then simply disappear.

Correspondence and Equivalence If we want to prove equivalence between BG's approach and $\mathbf{S}_0$, we first 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 *If a domain description D in the language $\mathcal{L}_3$ contains two effect laws*

a **causes** f **if** $p_1, \ldots, p_n$ and
a' **causes** $\neg f$ **if** $p'_1, \ldots, p'_m$

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

The correspondence relations '$\sim$' and '$\cong$' are introduced in definitions 6 and 7 In these definitions we will deviate a little from our usual notation: we will denote tokens for event, fluent and time name constants using regular (non-bold,non-capital) symbols as is done in [2].

Definition 9 *The method* $\mathbf{M}_0$ *applied to a domain theory* T_{C} *consists of the following steps:*

1. *Circumscribe* Ca *in* T_{C} *with* Ho *(and all other functions and predicates) fixed and let the resulting theory be* T'_{C} *:*
$$T'_{\mathrm{C}} = Circum(T_{\mathrm{C}}\ ; Ca)$$
2. *Now add the causation and persistence axioms* (M21), (M22) and (M23) *to* T'_{C} *thereby obtaining a theory* T''_{C} *.*
3. *Circumscribe* Ab^1_g*,* Ab^2_f *and* Ab^2_e *in* T''_{C} *with all other predicates and functions allowed to vary, giving* Ab^1_g *a higher priority than* Ab^2_f *and* Ab^2_f *a higher priority than* Ab^2_e*. Let the resulting theory be* $\mathbf{M}_0(T_{\mathrm{C}}\ )$*:*
$$\mathbf{M}_0(T_{\mathrm{C}}\ ) = Circum(T''_{\mathrm{C}}\ ; Ab^1_g > Ab^2_f > Ab^2_e; \langle \textit{ all other non-logical symbols } \rangle)$$

Theorem 2 *If we have any theory* T_{C} *for our language and domain description* D *for a language* $\mathcal{L}_3$ (D) *of Baral and Gelfond's such that a)* $T_{\mathrm{C}} \sim D$*, b)* D *is not* ambiguous*, and c) a model* $\mathcal{M} \models T_{\mathrm{C}}$ *with* $\mathbf{Ab}^1_g(\mathcal{M}) = \mathbf{Ab}^2_f(\mathcal{M}) = \emptyset$ *exists, then*

1. $\mathcal{M} \in \mathbf{S}_0(T_{\mathrm{C}}\ ) \Rightarrow$ *there exists a model* M *for* D *such that* $\mathcal{M} \cong M$*.*
2. M *is a model for* D $\Rightarrow$ *there exists an* $\mathcal{M} \in \mathbf{S}_0(T_{\mathrm{C}}\ )$ *such that* $\mathcal{M} \cong M$*.*

9.3 Circumscriptive Causal Approaches and $\mathbf{M}_0$

Before we give any more comparisons of $\mathbf{S}_0$, we will first rephrase it as an equivalent method $\mathbf{M}_0$ that achieves nonmonotonicity by using *circumscription* [22, 23]. This will make comparisons to other circumscriptive approaches such as causal minimization (section 9.4) and Lin's 'embrace of causality' (section 9.5) much easier, as these approaches have a large 'surface similarity' to $\mathbf{M}_0$: just looking at the definitions of the approaches will show how related to $\mathbf{M}_0$ (and therefore, $\mathbf{S}_0$) they are.

An additional advantage is that with $\mathbf{M}_0$ we express some of the causal reasoning within our domain theories themselves, thus implicitly showing that we do not have to do everything semantically (as we do in $\mathbf{S}_0$ and $\mathbf{I}_0$).

The Reasoning Method $\mathbf{M}_0$

In the following we assume that our language contains the three additional predicates $Ab^1_g(g,t)$, $Ab^2_f(f,t)$ and $Ab^2_e(e,t)$. Our theories T_{C} are not allowed to mention these three predicates however. In the definition of $\mathbf{M}_0$ we will use some additional axioms. First, there is a *causation axiom*:

$$\neg Ab^1_g(g,t) \supset (Ca(g,t) \supset Ho(g,t)) \tag{M21}$$

The following is our *persistence axiom part I*:

$$\begin{aligned}&\neg Ab^2_f(f,t) \supset \\ &\qquad [\neg Ca(f,t) \wedge \neg Ca(Not(f),t)] \supset [Ho(f,t-1) \equiv Ho(f,t)]\end{aligned} \tag{M22}$$

(M22) says that *one can prove that a fluent persists iff one can prove that there is no cause for it to change*. Here is our *persistence axiom part II*:

$$\neg Ab^2_e(e,t) \supset \quad [\neg Ca(e,t) \supset Ho(Not(e),t)] \tag{M23}$$

$\mathbf{M}_0$ is introduced in definition 9 as a circumscriptive two-step procedure (we use the notation of [15]). To see that this procedure is really almost equivalent to $\mathbf{S}_0$, one should realize first that $\mathbf{S}_0$ still selects the same models if we change it in the following way: we skip its second step and the part of the third step involving *JE*. We then replace all occurrences of *JE* in the definitions of $\mathbf{Ab}_f^2$ and $\mathbf{Ab}_e^2$ (step 4 of $\mathbf{S}_0$, see[2] definition 2) by the instances of *Ho* to which they are equivalent. For example, step 2 and step 3 make $JE(\mathbf{f},\mathbf{t})$ equivalent to $Ho(\mathbf{f},\mathbf{t}-1)$ for all $\mathbf{t} > 1$. We can thus replace $JE(Not(\mathbf{f}),\mathbf{t})$ in the definition of $\mathbf{Ab}_f^2$ by $Ho(Not(\mathbf{f}),\mathbf{t}-1)$. The definition of *JE* in step 2 and 3 of $\mathbf{S}_0$ implies that there can never be any elements $(\mathbf{f},0)$ in $\mathbf{Ab}_f^2$ anyway so the special case $\mathbf{t} = 0$ is of no concern either. We can replace $JE(\mathbf{e},\mathbf{t})$ analogously. We have now modified $\mathbf{S}_0$ such that it looks different but still selects the same models. Note that the minimization of *Ca* (step 3 of $\mathbf{S}_0$) has *not* been modified. This minimization is achieved in $\mathbf{M}_0$ by circumscribing *Ca* with *Ho fixed*: in that way, it is made sure that no particular extension of *Ho* is ruled out. After this minimization of *Ca*, the causation axiom (M21) is added, giving the predicate Ab_g^1 the same 'extension' as we give to the set $\mathbf{Ab}_g^1$ in step 4 of $\mathbf{S}_0$. The persistence axiom (M22) does not look too similar to the definition of $\mathbf{Ab}_f^2$. Indeed, one would rather expect the following:

$$\begin{aligned}&\neg Ab_f^2(f,t) \supset\\ &\qquad \neg Ca(f,t) \supset (Ho(Not(f)),t-1) \supset Ho(Not(f),t))\end{aligned} \tag{M24}$$

Interestingly, however, we have the following proposition (the proof of which is trivial):

Proposition 3 $[\forall f,t.\neg Ab_g^1(f,t) \wedge \neg Ab_f^2(f,t)] \supset [\ ((\text{M21}) \wedge (\text{M24})) \ \Leftrightarrow\ ((\text{M21}) \wedge (\text{M22}))\ ]$

In other words, though (M24) and (M22) are not equivalent, they *become* equivalent in the presence of the causation axiom and if additionally, nothing unexpected happens. Considering all this, it is not surprising that we can prove equivalence between $\mathbf{S}_0$ and $\mathbf{M}_0$ as long as we do not have any abnormalities. For this, we first need some more definitions. We first define $\mathcal{M}_1 \equiv_{Ca} \mathcal{M}_2$, $\mathcal{M}_1 \equiv_{Ab_g^1} \mathcal{M}_2$, $\mathcal{M}_1 \equiv_{Ab_f^2} \mathcal{M}_2$ and $\mathcal{M}_1 \equiv_{Ab_e^2} \mathcal{M}_2$ analogously to $\mathcal{M}_1 \equiv_{Ho} \mathcal{M}_2$ (see section 2.3). This leads us to the following definition:

Definition 10 *Suppose* M *is some class of models for our language extended with predicates* Ab_g^1, Ab_f^2 *and* Ab_e^2*. We will write '*$\mathcal{M}\rfloor_{HCA} \in \mathrm{M}$*' as an abbreviation for*

$$\begin{array}{lll}\exists \mathcal{M}' \quad \mathcal{M}' \in \mathrm{M} & \text{and} & \mathcal{M}' \equiv_{Ho} \mathcal{M} \text{ and } \mathcal{M}' \equiv_{Ca} \mathcal{M}\\ & \text{and} & \mathcal{M}' \equiv_{Ab_g^1} \mathcal{M} \text{ and } \mathcal{M}' \equiv_{Ab_f^2} \mathcal{M} \text{ and } \mathcal{M}' \equiv_{Ab_e^2} \mathcal{M}\\ & \text{and} & \forall \mathbf{g},\mathbf{t} \ :\ (\mathbf{g},\mathbf{t}) \in \mathbf{Ab}_g^1(\mathcal{M}') \Leftrightarrow \mathcal{M}' \models Ab_g^1(\mathbf{g},\mathbf{t})\ \}\\ & \text{and} & \forall \mathbf{f},\mathbf{t} \ :\ (\mathbf{f},\mathbf{t}) \in \mathbf{Ab}_f^2(\mathcal{M}') \Leftrightarrow \mathcal{M}' \models Ab_f^2(\mathbf{f},\mathbf{t})\ \}\\ & \text{and} & \forall \mathbf{e},\mathbf{t} \ :\ (\mathbf{e},\mathbf{t}) \in \mathbf{Ab}_e^2(\mathcal{M}') \Leftrightarrow \mathcal{M}' \models Ab_e^2(\mathbf{e},\mathbf{t})\ \}\end{array}$$

We are now ready for our theorem:

Theorem 3 *If we have any* T_C *for our language extended with* Ab_g^1, Ab_f^2 *and* Ab_e^2 *such that a)* T_C *does not mention JE nor any* $\mathbf{t} \in \mathrm{TNC}$*, and b) There is an* $\mathcal{M} \in \mathbf{S}_0(T_\mathrm{C})$ *with* $\mathbf{Ab}_g^1(\mathcal{M}) = \mathbf{Ab}_f^2(\mathcal{M}) = \emptyset$*, then we have for all* $\mathcal{M}$ *for our language that*

$$\mathcal{M}\rfloor_{HCA} \in \mathbf{S}_0(T_\mathrm{C}) \Leftrightarrow \mathcal{M}\rfloor_{HCA} \in Mod(\mathbf{M}_0(T_\mathrm{C}))$$

[2] We number the steps in $\mathbf{S}_0$ as in $\mathbf{I}_0$, i.e. we call the first step of $\mathbf{S}_0$ 'step 2'.

9.4 Haugh, Lifschitz and Rabinov's Causal Minimization

Causal Minimization was introduced independently by Lifschitz [17] and Haugh [11]. We focus on Lifschitz' and Rabinov's approach [18], in which some of the problems of the two earlier methods are overcome. Lifschitz and Rabinov (LR) represent time as in the situation calculus [24]. The basic sort in their logic is a *situation*, which stands for the state of the world at a particular point in time. For any situation s, the function $\mathit{Result}(a,s)$ denotes the situation that would result if action a were performed in s. LR use truth-valued fluents; the function $\mathit{Value}(f,s)$ gives the value of fluent f in situation s (thus $\mathit{Value}(f,s) = \text{TRUE}$ corresponds to our $\mathit{Ho}(f,t)$.) The predicate $\mathit{Causes}(a,f,v)$ is used to indicate that action a may, under some conditions, cause fluent f to take on value v (TRUE or FALSE). $\mathit{Success}(a,s)$ is true iff all preconditions for action a to actually cause something in situation s are true. $\mathit{Miracle}(f,s)$ will mean that a miracle (something unexplainable) happened to the fluent f in situation s. We will now quote the two basic domain independent axioms of the approach. We changed the syntax somewhat to make the axioms easier to understand within our context. The given formulas are equivalent to the original ones, however. First, there is the *Law of Change*:

$$\begin{aligned}&\neg \mathit{Miracle}(f,\mathit{Result}(a,s)) \supset\\ &\quad [\mathit{Success}(a,s) \wedge \mathit{Causes}(a,f,v)] \supset \mathit{Value}(f,\mathit{Result}(a,s)) = v \end{aligned} \tag{LR25}$$

Similarly, there is the *Law of Inertia*:

$$\begin{aligned}&\neg \mathit{Miracle}(f,\mathit{Result}(a,s)) \supset\\ &\quad [\neg \mathit{Success}(a,s) \vee (\neg \mathit{Causes}(a,f,\text{FALSE}) \wedge \neg \mathit{Causes}(a,f,\text{TRUE}))] \supset\\ &\qquad \mathit{Value}(f,\mathit{Result}(a,s)) = \mathit{Value}(f,s)\end{aligned} \tag{LR26}$$

In simple theories in which actions do not have any preconditions, we may assume that $\mathit{Success}(a,s)$ holds for every a and s – we will not go into to the particular way in which preconditions are handled in LR's approach, as the approach can already go wrong easily in theories without them. The interested reader may check in [18] that in theories without preconditions $\mathit{Success}(a,s)$ will indeed always be true.

LR start with theories T_{LR} that contain the two axioms above, some other domain-independent axioms (which are not relevant for the discussion here) and some domain-dependent knowledge. They then achieve persistence by circumscribing Ca and $\mathit{Miracle}$ in T_{LR} with Causes given a higher priority than $\mathit{Miracle}$, and Value allowed to vary.

Now compare (LR25) with our causation axiom (M21) and (LR26) with our persistence axiom (M22), and notice that they are *very* similar! Clearly, $\mathit{Miracle}$ corresponds to $\mathbf{Ab}^1_g$ and $\mathbf{Ab}^2_f$, Causes corresponds to Ca and Value to Ho.

A Distinguishing Example If the two basic axioms are so similar, why then do LR and $\mathbf{S}_0$ get different results in so many of the well-known examples of reasoning domains? Before answering this question, we will introduce a new simple example where things go wrong for LR while $\mathbf{M}_0$ (and thus $\mathbf{S}_0$) handle it correctly. Suppose we have a theory T_{LR} containing all domain-independent axioms for LR's approach and the following domain dependent ones:

$$\begin{gathered}\mathit{Causes}(\mathit{Shoot},\mathit{Alive},\text{FALSE})\\ \mathit{Value}(\mathit{Alive},S_0) = \text{TRUE}\\ \mathit{Value}(\mathit{Alive},s) = \text{FALSE} \supset \mathit{Value}(\mathit{Walking},s) = \text{FALSE}\end{gathered}$$

Proposition 4 *Using Lifschitz' and Rabinov's circumscription, we can now miraculously deduce:*

$$\mathit{Value}(\mathit{Walking},S_0) = \text{FALSE} \quad ! \tag{LR27}$$

Proof: Follows directly from the definitions in [18]. □

Intuitively, this is because with the law of change (LR25) one can deduce $\mathit{Value}(\mathit{Alive}, \mathit{Result}(\mathit{Shoot}, S_0)) =$ FALSE, but as the circumscription of *Causes* yields $\neg \mathit{Causes}(\ \mathit{Shoot}, \mathit{Walking}, v)$, one can then deduce (LR27) thru the law of inertia (LR26). There may be something to say for a solution in which $\mathit{Value}(\mathit{Alive}, \mathit{Result}(\mathit{Shoot}, S_0)) =$ TRUE: we did not specify the domain constraint in a causal form (any way, it is impossible to represent causal rules that do not contain actions in Lifschitz and Rabinov's system); so the constraint may be interpreted as giving an implicit precondition for the *Shoot*-action to be successful. But we do not get that solution either.

The reason that things go wrong here is that LR do not realize that, when they circumscribe *Causes*, they are in the business of generating causal laws (or constructing a causal graph, if you like). This generation of laws should of course not have any influence on what holds when! Only once we know these laws (the graph) can we use them to decide what should hold when. But in LR's approach, *Value* is allowed to vary during the circumscription of *Causes* and thus possible extensions of *Value* (i.e. *Ho*, in our approach) are spuriously ruled out.

We can see that our 'basic principle' and our 'preliminary work' are *mixed up* here: models with certain fluent facts are already ruled out while it has not yet been established whether these fluent facts are explained in these models or not.

9.5 Lin's Embrace of Causality

Lin [19, 20] has recently introduced a new method for reasoning about action that is based on a version of the situation calculus [24]. In the situation calculus, time is represented by *situations* s rather than timepoints. The sort *actions* corresponds to our event fluents; however, in Lin's approach symbols of the sort are not necessary constants (i.e. they can be n-ary function symbols rather than only 0-ary). Similarly, fluents in situation calculus correspond to our regular fluents, and again, Lin allows fluent symbols to be n-ary. For any action e and situation s, the function $Do(e, s)$ 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 *Ca* and *Poss*.

It turns out that Lin's method is *very* similar to ours[3]. Lin's *Ca*-predicate is ternary: $Ca(f, v, s)$ is true if the fluent f is caused to have the truth value v in situation s. On page 20 we give a (somewhat extended) quote of [20] 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 now first need the following:

Proposition 5 *If we exchange step 1 and step 2 in Lin's procedure on page 20 we will arrive at an equivalent final theory* T'''_{LIN} .

Proof: If we rewrite (L28) and (L29) as disjunctions, then *Ca* appears only in negated form in them. It then easily follows from the model-theoretic characterization of circumscription [15] that the theory obtained by adding (L28) and (L29) before the circumscription has the same models as the theory obtained by adding them after the circumscription. □

But with step 1 and step 2 interchanged and step 4 left out, Lin's approach starts to look suspiciously similar to $\mathbf{M}_0$! While $\mathbf{S}_0$ and $\mathbf{M}_0$ do not handle the qualification problem, Lin's approach does not handle failure of causation and unexpected breaches of persistence (i.e. it has no abnormality predicates). Neither can it represent concurrent events and events causing other events to take place. Otherwise, as is hopefully clear from just looking at the respective definitions, Lin's approach is almost identical to $\mathbf{S}_0$. Though we have not proven it formally, we conjecture the two approaches to be equivalent on the set of reasoning domains where both are defined.

[3] We should add here that a large part of Lin's papers is about restricted variants of his approach which are computationally tractable. We do not restrict ourselves to these variants here.

Lin's Procedure Start with a theory T_{LIN} that consists of unique names axioms, all domain-dependent axioms and no further axioms.

1. Add the following basic axioms for the predicate *Ca* to T_{LIN} :

$$Ca(f, \text{TRUE}, s) \supset Ho(f, s) \tag{L28}$$
$$Ca(f, \text{FALSE}, s) \supset \neg Ho(f, s) \tag{L29}$$

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

$$Poss(a, s) \supset \{(\neg Ca(f, \text{TRUE}, Do(a, s)) \wedge \neg Ca(f, \text{FALSE}, Do(a, s))) \supset [Ho(f, Do(a, s)) \equiv Ho(f, s)]\} \tag{L30}$$

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

We end this section with the following very important remark. From the point of view of our philosophy, one should add causation and persistence axioms only *after* circumscribing *Ca*. Remember that we interpret the circumscription/minimization of *Ca* in a T_{C} as a step in the determination of a causal graph belonging to T_{C} . The causation and persistence axioms are meant to infer what holds and what does not *using an* already existing *causal graph*, i.e. an existing full extension of *Ca*. Clearly, we should not be using these axioms to determine the extensions of *Ca* themselves!

Now because *Ca* appears only negatively in the causation axioms, nothing changes in *practice* if one adds them before circumscribing *Ca*. However if one adds the persistence axiom before, both Lin's and our method would break down: for example, in a model with $Ho(Alive, t)$ and $\neg Ho(Alive, t+1)$ and $\neg Ab^2_f(Alive, t+1)$ we would automatically infer $Ca(Not(Alive), t+1)$, thus introducing a spurious sufficient cause.

Interestingly, in Lin's papers neither the choice to keep *Ho* fixed during circumscription of *Ca*, nor the choice to add the persistence axiom only after this first circumscription is motivated in terms other than 'if you do not do it, you get counterintuitive results'. Causal network theory thus provides an external motivation for these choices.

10. $\mathbf{I}_0$ AND THE OTHERS

We will now see that two popular non-causal approaches, namely Baker's [1] and 'Chronological Minimization' [13, 30] can be reinterpreted as an approximation of $\mathbf{I}_0$ - specifically, these approaches try to generate the sufficient causes within their models, but sometimes fail to do so – so once again, our 'preliminary work' is not done well. While our comparison to chronological minimization remains completely informal, in Baker's case we again provide an equivalence theorem and a counterexample.

10.1 $\mathbf{I}_0$ *and Baker*

Baker's approach is based on the situation calculus [24] where time is represented by *situations s* rather than timepoints; there still is a finite set of regular fluent constants FC and of events EC, but *Ho* is now defined only for regular fluents f and situations s. For any e and s, $Result(e, s)$ stands for the situation that results when performing e in s. Theories T_{B} for Baker's approach all contain domain independent axioms (5), (6) (with g replaced by f and t by s) and (8). Furthermore they contain the fluent domain-closure axiom (31) and the *frame axiom* (32) :

$$f = F_1 \vee \ldots \vee F_{|\mathrm{FC}|} \tag{31}$$

$$\neg Ab(f, e, s) \supset (Ho(f, Result(a, s)) \equiv Ho(f, s)) \tag{32}$$

Finally, there are *existence of situations* ('EOS') axioms (the precise form of which is given in def. 11, page 23) that make sure that for each set of fluents $\mathrm{F} \subseteq \mathrm{FC}$ there is at least one situation in which they hold. Baker's method consists of circumscribing [22] the *Ab*-predicate in a T_{B} with *Ho* fixed and *Result* allowed to vary. Roughly, this means that we select those models of T_{B} in which we have $Ab(f, e, s)$ (and thus, by (32), no persistence) only if f is *forced* to take on another value if e is performed in s.

It now turns out that we can reinterpret $Ab(f, e, s)$ as saying 'e is a sufficient cause for f in situation s'; similarly, circumscribing *Ab* is very similar to the inference of $\mathbf{s-causes}$ in step 1 of $\mathbf{I}_0$. To see this, one should realize that in the absence of any sufficient cause for a fluent $\mathbf{f}$ to take on a specific value at time $\mathbf{t}$, we have persistence in all our preferred models: $Ho(\mathbf{f}, \mathbf{t}) \equiv Ho(\mathbf{f}, \mathbf{t}-1)$. It now follows that if in a preferred model we have a change of fluent value from time $\mathbf{t}-1$ to $\mathbf{t}$, this means we *must also have a sufficient cause in our model* for the fluent to take on a specific value at time $\mathbf{t}$. Now notice that we have $Ab(f, e, s)$ only if there is an *event* e, that, when performed in *circumstances* s, brings about a *change* in the value of f, i.e. the event e generates a sufficient cause for f!

The EOS- axioms can now be interpreted as making sure that, in determining the instances of *Ab*, one looks at all possible situations, including *counterfactual* ones in which the contingent observations do not hold! But EOS only creates counterfactual situations for all possible combinations of *regular* fluent values, while step 1 of $\mathbf{I}_0$ takes into account *all* counterfactual *models* consistent with the effect axioms. This means that in Baker's approach, contingent observations can sometimes still influence the sufficient causes found. We illustrate this by an example that shows that in counterfactual situations/models a) one also needs to include event fluents; b) one also has to consider timepoints earlier than the time at which an action is performed.

Example We now feature a variation of the YSP, in which a new action *Point* is introduced: we suppose that if we *Shoot*, we have to *Point* first in order not to miss. We formalize this in a theory $T_{\mathrm{B\text{-}POINT}}$ for Baker's approach, containing axioms (5),(6),(8) & (31)-(34) :

$$Ho(Loaded, Result(Point, s)) \supset [\ \neg Ho(Alive, Result(Shoot, Result(Point, s)))\ ] \tag{33}$$

$$Ho(Alive, S_0) \wedge Ho(Loaded, S_0) \tag{34}$$

Suppose $\mathbf{s} = Result(Shoot, Result(Point, S_0))$.

Proposition 6 *Baker's procedure does not prefer models for* $T_{\mathrm{B\text{-}POINT}}$ *with* $\neg Ho(Alive, \mathbf{s})$ *over models with* $Ho(Alive, \mathbf{s})$.

Proof: It is easy to show by the technique used in [1] that this is because *no counterfactual situation exists in which the Point-action does not take place.* □

Note also that in a non-causal theory $T_{\mathrm{NC\text{-}POINT}}$ that encodes the same domain knowledge, the problem does not occur: step 1 of $\mathbf{I}_0$ will make sure that any model $\mathcal{M}$ with $\mathcal{M} \models Ho(Loaded, \mathbf{t}) \wedge Ho(Point, \mathbf{t}-1) \wedge Ho(Shoot, \mathbf{t})$ will also have $Ca(Not(Alive), \mathbf{t}+1)$, and thus this $\mathcal{M}$ will only be selected in step 5 of $\mathbf{I}_0$ if it also has $Ho(Not(Alive, \mathbf{t}+1)$.

All other problems with Baker's approach we are aware of, like the 'extended stolen car problem', the 'two-tank problem' [4] and the general problem of handling actions with nondeterministic effects [12] can be seen to stem from either the narrow definition of counterfactual situations or from the fact that Baker uses a single predicate *Ab* to denote both causation (our *Ca*) and surprise (our $\mathbf{Ab}^1_g$ and $\mathbf{Ab}^2_f$).

If one restricts $\mathbf{I}_0$ such that our 'circumstances' (in def.3) coincide with Baker's 'situations', and furthermore one allows only theories for which there are models without surprises about regular fluents, then the two methods become the same! Formally, we define $\mathbf{I}_0^{\mathrm{B}}$ to be the same as $\mathbf{I}_0$ except that a) in definition 3, Φ must only contain fluent facts of the form $Ho(\mathbf{f}, \mathbf{t}_\phi)$, and b), step 5.3 is changed such that only models in which exactly one event happens at a time are selected.

In def. 11 we define $T_{\mathrm{NC}} \sim T_{\mathrm{B}}$ to hold for theories T_{NC} of our approach and theories T_{B} of Baker's approach which intuitively encode the same domain knowledge. The class of T_{NC} for which $T_{\mathrm{NC}} \sim T_{\mathrm{B}}$ is defined seems rather restricted; still, most specific examples of reasoning problems given in the literature on Baker's approach fit in this format. We want to prove that Baker's method and $\mathbf{I}_0^{\mathrm{B}}$ select the same models in terms of what holds at what time. However, one model for a theory in situation calculus corresponds to a whole *set* of models in our language, namely, one specific model for each possible sequence of events. We call such a set of models a 'branching time (b.t.) model set' (def.12). In def.13 we define semantical correspondence: $\mathcal{M}_{\mathrm{B}} \cong \mathrm{M}_{bt}$ for models $\mathcal{M}_{\mathrm{B}} \in Mod(T_{\mathrm{B}})$ and b.t. model sets $\mathrm{M}_{bt} \subset Mod(T_{\mathrm{NC}})$ will hold iff $\mathcal{M}_{\mathrm{B}}$ models exactly the same fluent histories as the elements of M_{bt} .

Theorem 4 *If we have any T_{NC} and T_{B} such that a) $T_{\mathrm{NC}} \sim T_{\mathrm{B}}$, and b) there is a b.t. model set $\mathrm{M}_{bt} \subseteq \mathbf{I}_0^{\mathrm{B}}(T_{\mathrm{NC}})$ with for all $\mathcal{M}_{\mathrm{NC}} \in \mathrm{M}_{bt}$, $\mathbf{Ab}_f^2(\mathcal{M}_{\mathrm{NC}}) = \emptyset$ then for all $\mathcal{M}_{\mathrm{B}}$ and M_{bt} :*

1. *Baker's procedure selects $\mathcal{M}_{\mathrm{B}}$ $\Rightarrow$ There exists an M_{bt} with $\mathrm{M}_{bt} \cong \mathcal{M}_{\mathrm{B}}$ and $\mathrm{M}_{bt} \subseteq \mathbf{I}_0^{\mathrm{B}}(T_{\mathrm{NC}})$.*
2. *$\mathrm{M}_{bt} \subseteq \mathbf{I}_0^{\mathrm{B}}(T_{\mathrm{NC}})$ $\Rightarrow$ There exists an $\mathcal{M}_{\mathrm{B}}$ preferred by Baker's procedure with $\mathrm{M}_{bt} \cong \mathcal{M}_{\mathrm{B}}$.*

10.2 $\mathbf{I}_0$ *and Chronological Minimization*

In *chronological minimization* [30, 13] (CM) one selects a model with a change of fluent value at time t only if there is no model which is the same (in terms of fluent facts) up until time t but in which no fluent changes value at time t. Again, this can be reinterpreted as approximating step 1 of $\mathbf{I}_0$. To see this, note that it is equivalent to

> selecting a model $\mathcal{M}$ with a change of fluent value at time $\mathbf{t}$ only if that change is *forced* to happen, given the fact that the past cannot be altered by any event.

Here 'forced' means that there are no models in which the same event(s) take(s) place at time $\mathbf{t}$ while the change at $\mathbf{t}$ does not take place, but which are the same in terms of *Ho* up until time $\mathbf{t}$. Now in the absence of any sufficient cause for a fluent to take on another value at time $\mathbf{t}$, the fluent value will persist preferred models. This means that the fluent will only change value if there is a sufficient cause for it. Thus selecting a model with a change of fluent value $\mathbf{f}$ at $\mathbf{t}$ only if the change is forced to happen can be seen as trying to find the models where there is either persistence of $\mathbf{f}$ *or* a sufficient cause for $\mathbf{f}$ to take on another value, and this is already quite similar to $\mathbf{I}_0$. The condition above that 'the past cannot be altered by any event' embodies the same physical assumption as the one we make in step 4 of definition 3.

It should then come as no surprise that CM does not work very well unless enhanced by *filter preferential entailment* [29] which modifies CM by first removing all regular fluent facts from a theory, then performing CM on the models of this 'general' theory and only then further select those models in which the regular fluent facts hold .

11. Conclusion and Discussion

From all the above we conclude that our work unifies many different approaches to nonmonotonic temporal reasoning. We end this paper by addressing two more or less philosophical issues that this conclusion raises: a new view on events and on the assessment of the *validity* of existing theories .

11.1 Events and Explanations

To us, nonmonotonic model selection is simply a search for the models in which as much as possible is explained. Such models typically contain chains of explanations: A is explained by B, B by C,

We call a vector $\vec{\mathbf{f}} = (\mathbf{f}_1, \ldots, \mathbf{f}_{|\mathrm{FC}|})$ where each $\mathbf{f}_i \in \{F_i, \mathit{Not}(F_i)\}\})$ a *full fluent vector.* We abbreviate
'$\mathit{Ho}(\mathbf{f}_1, t) \wedge \ldots \wedge \mathit{Ho}(\mathbf{f}_{|\mathrm{FC}|})$' to '$\mathit{Ho}(\vec{\mathbf{f}}, t)$' .

Definition 11 *For any two theories T_{NC} for our language and T_{B} for a language of Baker's approach, we define $T_{\mathrm{NC}} \sim T_{\mathrm{B}}$ to hold iff all of the following hold:*

1. *The two languages share the same* FC *and* EC .

2. *$T_{\mathrm{NC}}\rfloor_{\mathrm{IND}}$ consists of all domain independent axioms* (5)-(8). *T_{B} contains axioms* (5),(6),(8),(31),(32) .

3. *$T_{\mathrm{NC}}\rfloor_{\mathrm{GEN}}$ contains an axiom of form*
$\mathit{Ho}(\mathbf{f}_1, t) \wedge \ldots \mathit{Ho}(\mathbf{f}_n, t) \wedge \mathit{Ho}(\mathbf{e}, t) \supset \mathit{Ho}(\mathbf{f}_r, t+1)$
iff T_{B} contains the axiom
$\mathit{Ho}(\mathbf{f}_1, s) \wedge \ldots \mathit{Ho}(\mathbf{f}_n, s) \supset \mathit{Ho}(\mathbf{f}_r, \mathit{Result}(\mathbf{e}, s))$ (B0)

4. *$T_{\mathrm{NC}}\rfloor_{\mathrm{GEN}}$ contains an axiom of form*
$\mathit{Ho}(\mathbf{f}, t) \supset \mathit{Ho}(\mathbf{f}', t)$ (B1)
iff T_{B} contains the same axiom (where t is taken to be a situation constant) .

5. *T_{B} contains a propositional combination of fluent facts, each of the form $\mathit{Ho}(\mathbf{f}, \mathbf{s})$, with*
$\mathbf{s} = \mathit{Result}(\mathbf{e}_t, \mathit{Result}(\mathbf{e}_{t-1}, \ldots, \mathit{Result}(\mathbf{e}_0, S_0) \ldots))$
iff T_{NC} contains the corresponding propositional combination where each $\mathit{Ho}(\mathbf{f}, \mathbf{s})$ is replaced by $[\mathit{Ho}(\mathbf{e}_0, 0) \wedge \ldots \wedge \mathit{Ho}(\mathbf{e}_t, t) \supset \mathit{Ho}(\mathbf{f}, t+1)]$.

6. *T_{B} contains* existence of situations-*axioms of the form $\exists s \mathit{Ho}(\vec{\mathbf{f}}, s)$ for all $\vec{\mathbf{f}}$ whose existence is consistent with all axioms in T_{B} of the form* (B0) *and* (B1) *(i.e. $\mathit{Ho}(\vec{\mathbf{f}}, s) \not\vdash \neg(\bigwedge \text{(B0)} \wedge \bigwedge \text{(B1)}\,))$.*

7. *Each axiom in T_{B} is of one of the forms occurring in conditions 2-6 above. Each axiom in T_{NC} is of one of the forms occurring in conditions 2-5 above.*

. . . but these chains have to stop somewhere. In our view, 'events' are just those things that we are willing to accept as the end of a chain of explanations. Even for an event we might prefer models where it is explained by some other event (see the example in section 9.1); but usually, in the common sense world, an event that takes place without a cause (like *Shoot* in the models for $T_{\mathrm{YSP\text{-}C}}$) is less surprising than a fluent that takes on a value without a cause (e.g. *Alive* becoming *Not*(*Alive*) without a *Shoot*). In some contexts we might not even want to look for an explanation of the events that are known to take place – for example, if we perform them ourselves. In such a case, we would rather speak of 'actions' than 'events' and model knowledge about them rather as $\mathit{Ca}(\mathbf{e}, \mathbf{t})$ than $\mathit{Ho}(\mathbf{e}, \mathbf{t})$.

11.2 Validity

How can we assess the validity of our approach? We have proven equivalences between our approach and ADL–ones. Actually, such approaches were partially developed as a touchstone for other, intuitively less clear ones: if one could prove those equivalent to an ADL–approach, then this would show that they are correct, in a certain sense. We have given such proofs here, but on the other hand, we have shown that even the ADL-approaches can sometimes give counterintuitive results.

What does this say about the status of either our approach or the ADL-approaches? We will end this paper by trying to answer this question, thereby implicitly providing a view on applied nonmonotonic reasoning in general, not just in the temporal domain.

In the following $\mathit{Ho!}(\mathbf{e}, t)$ will be an abbreviation for $\mathit{Ho}(\mathbf{e}, t) \ \wedge \ \left[\bigwedge_{1\leq i\leq|}\mathrm{EC}_{|,E_i\neq\mathbf{e}} \neg\mathit{Ho}(E_i, t)\right]$.

Definition 12 *A* branching time (b.t.) model set M_{bt} *is any set of models for a theory* T_{NC} *such that :*

1. *For all finite sequences of events* $[\mathbf{e}_0, \ldots, \mathbf{e}_t]$ $(t \geq 0)$
   - M_{bt} *contains an* $\mathcal{M}$ *with*
     $\mathcal{M} \models \mathit{Ho!}(\mathbf{e}_0, 0) \wedge \ldots \wedge \mathit{Ho!}(\mathbf{e}_t, t)$.
   - *There is a sequence of full fluent vectors* $[\vec{\mathbf{f}}_0, \ldots \vec{\mathbf{f}}_{t+1}]$ *such that for all* $\mathcal{M}$ *in* M_{bt}*:*
     $\mathcal{M} \models \left[\bigwedge_{0\leq i\leq t} \mathit{Ho!}(\mathbf{e}_i, i)\right] \supset \left[\bigwedge_{0\leq i\leq t+1} \mathit{Ho}(\vec{\mathbf{f}}_i, i)\right]$
2. *For all* $\mathcal{M}$ *in* M_{bt}, $\mathcal{M} \models \forall t \exists e \ . \ \mathit{Ho!}(e, t)$

Definition 13 *Given an* $\mathcal{M}_{\mathrm{B}}$ *for a* T_{B} *and a b.t. model set* M_{bt} *of models for a* T_{NC} *with* $T_{\mathrm{NC}} \sim T_{\mathrm{B}}$ *we write* $\mathcal{M}_{\mathrm{B}} \cong \mathrm{M}_{bt}$ *iff for all finite sequences of events* $[\mathbf{e}_0, \ldots, \mathbf{e}_t]$ $(t \geq 0)$ *and for all* $\mathbf{f}$*:*

$$\begin{array}{l} \mathcal{M}_{\mathrm{B}} \models \mathit{Ho}(\mathbf{f}, \mathit{Result}(\mathbf{e}_t, \ldots, \mathit{Result}(\mathbf{e}_0, S_0)\ldots)) \\ \qquad\qquad \textit{iff for all } \mathcal{M} \in \mathrm{M}_{bt} \textit{ we have} \\ \qquad \mathcal{M} \models [\bigwedge_{0\leq i\leq t} \mathit{Ho!}(\mathbf{e}_i, i)] \supset \mathit{Ho}(\mathbf{f}, t+1) \end{array}$$

11.3 The Strange Loop in Applied Nonmonotonic Reasoning

In applied nonmonotonic reasoning, we want to express knowledge we have about the 'physics' of some reasoning domain in a set of axioms and some nonmonotonic formalism. The axioms and the formalism are always precisely defined, but our 'translation step' from common-sense knowledge (as represented in natural language) to its axiomatic representation certainly is not.

Now over the last ten years, we have seen the following circular pattern in applied nonmonotonic temporal reasoning:

1. A simple formalization of domain knowledge into axioms of a logical theory is proposed. A nonmonotonic reasoning mechanism is provided along with it. The reasoning mechanism gives the right inferences on all or most examples of reasoning domains that have been considered in the literature so far.

2. Somebody comes up with a counterexample: a simple reasoning domain for which the proposed axiomatization+reasoning mechanism leads to the inference of *unexpected*, *surprising* results.

3. In reaction to this, somebody proposes a new axiomatization+inference mechanism. The procedure is re-entered at step 1.

The hope seems to have been that, by repeating this procedure over and over again, we would one day end up with a theory of NMTR that would really work well. But after a few years, people started to realize that this method does not lead to much progress, and the new paradigm of ADL-languages was introduced: the idea was to formalize the physics of a domain using sentences in some formal language and provide a simple semantics for the language. The language should be defined such as to enable a *clear* semantics, one for which it is intuitively clear that it really corresponds to the common-sense world. Other, intuitively less clear approaches could then be validated by proving that they lead to the same inferences as an existing ADL-based approach.

But our work makes clear that ADL approaches do not escape from the *strange loop* either! Once action description languages become rich enough to admit descriptions of complex reasoning domains, their semantics may again lead to unintuitive inferences, as is clearly shown in this paper! (section 9.2 gives two examples of reasoning domains where ADL-approaches lead to the wrong conclusions) Again, one is tempted to patch up a faulty ADL-approach and reenter the loop at step 1.

Our Paradigm Now take a look at step 2 of the loop again: the proposed approaches suffer from *unexpected, surprising* conclusions. Now what if we take a meta-level stance *and formalize this notion of 'unexpectedness' or 'surprise' explicitly*? Indeed, this is exactly what we do in this paper, as stated by our 'basic principle' (section 3): we want to know for *every* contingent fact in *every* model whether it is explained (= expected, = not-surprising) or not, given the rest of the model.

Indeed, wasn't the original purpose of nonmonotonic reasoning to end up with the conclusions that *normally* hold, that we *expect*, that do not *surprise* us? So why not formalize domain knowledge in these terms in the first place? (!!!)

One may ask what kind of new nonmonotonic principle this will lead to. Actually, if one thinks about it carefully, this approach is not so new after all: *it is essentially probabilistic.*

Statements like 'if we have fact A in a model, this means that B is also explained in the model', can be interpreted as statements about conditional probabilities

$$P(B|A) = 1 - \epsilon \tag{35}$$

where ϵ is some small number[4]. In this way, we can translate all our qualitative statements about explanations into statements about probabilities. The set of probability statements that results defines a probability distribution on all models of our theory; selecting models with the least nr. of abnormalities then amounts to selecting the model with the highest probability.

...but wait a minute? What about al these arguments that probability theory should not be used in default reasoning because of the inherent complexity of probabilistic inference? And isn't it the case that statements like (35) above only *partially* specify a probability distribution?

Both questions have a single answer: probabilistic inference *is* complex, and probabilities on models indeed *are* only partially specified by statements like (35) unless... one uses *conditional independence*, the fundamental idea behind Bayesian Networks and other graphical statistical models [14]. Here, one assumes that all facts A and B for which no statement of the form (35) can be deduced, are independent: knowing A does not change the probability of B and vice versa, i.e. $P(B|A) = P(B)$ and $P(A|B) = P(A)$. With this additional assumption, we do end up with a single probability distribution over all models. In $\mathbf{I}_0$ and $\mathbf{S}_0$, the conditional independence assumption is enforced by the domain closure that is performed in step 3.

We thus end this paper, somewhat provocatively, by stating a conviction that we have gradually adopted while trying to make sense of all the existing approaches to NMTR: if one is to apply nonmonotonic reasoning in the real world, using a formalism compatible with probability theory may be *inevitable* if one wants to escape from the 'strange loop' that has plagued NMTR for so long.

References

1. A.B. Baker. Nonmonotonic reasoning in the framework of situation calculus. *Artificial Intelligence*, 49:5–23, 1991.
2. C. Baral and M. Gelfond.
   Reasoning about actual and hypothetical occurrances of concurrent and non-deterministic actions. Available via ftp://ftp.newcastle.edu.au/pub/papers/nrac95/baral.final.ps, 1995.
3. 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.
4. J.M. Crawford and D.W. Etherington. Formalizing reasoning about change: a qualitative reasoning approach. In *Proceedings AAAI-92*, pages 577–583, 1992.

[4] See Goldtszmidt & Pearl [9] for a nonmonotonic logic explicitly based on probabilities.

5. A. Darwiche and J. Pearl. Symbolic causal networks. In *Proceedings AAAI-94*, 1994.
6. 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.
7. H. Geffner. Causal theories for nonmonotonic reasoning. In *Proceedings AAAI-90*, pages 524–530, 1990.
8. M. Gelfond and V. Lifschitz. Representing actions in extended logic programming. In K. Apt, editor, *Logic Programming: Proceedings Tenth Conference*, pages 559–573, 1992.
9. M. Goldszmidt and J. Pearl. Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence*, 84(1-2):57–112, 1996.
10. S. Hanks and D. McDermott. Default reasoning, non-monotonic logics and the frame problem. In *Proceedings AAAI-86*, pages 328–333, 1986.
11. B.A. Haugh. Simple causal mimimizations for temporal persistence and projection. In *Proceedings AAAI-87*, pages 218–223, 1987.
12. G.N. Kartha. Two counterexamples related to Baker's approach to the frame problem. *Artificial Intelligence*, 69:379–391, 1994.
13. H.A. Kautz. The logic of persistence. In *Proceedings AAAI-86*, pages 401–405, 1986.
14. S.L. Lauritzen. *Graphical Models.* Oxford Science Publications, 1996.
15. V. Lifschitz. Computing circumscription. In *Proceedings IJCAI-85*, pages 121–127, 1985.
16. V. Lifschitz. Frames in the space of situations. *Artificial Intelligence*, 46:365–376, 1990.
17. V.L. Lifschitz. Formal theories of action. In M. L. Ginsberg, editor, *Readings in Nonmonotonic Reasoning.* Morgan Kaufmann Publishers, Los Altos, CA, 1987.
18. V.L. Lifschitz and A. Rabinov. Miracles in formal theories of action. *Artificial Intelligence*, 38:225–237, 1989.
19. F. Lin. Embracing causality in specifying the indirect effects of actions. In *Proceedings IJCAI-95*, 1995.
20. F. Lin. Embracing causality in specifying the indeterminate effects of actions. In *Proceedings AAAI-96*, 1996.
21. N. McCain and H. Turner. A causal theory of ramifications and qualifications. In *Proceedings IJCAI-95*, 1995.
22. J. McCarthy. Circumscription – a form of non-monotonic reasoning. *Artificial Intelligence*, 13(1,2):27–39,171–172, 1980.
23. J. McCarthy. Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence*, 26(3):89–116, 1986.
24. 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.
25. L. Morgenstern and L.A. Stein. Why things go wrong: A formal theory of causal reasoning. In *Proceedings AAAI-88*, pages 5–23, 1988.
26. J. Pearl. Graphical models, causality, and intervention. *Statistical Science*, 8(3):266–273, 1993.
27. J. Pearl. Causation, action and counterfactuals. In Y. Shoham, editor, *Proceedings TARK-VI*, pages 51–73. Morgan Kaufmann, 1996.
28. E.J. Sandewall. *Features and Fluents.* Oxford University Press, 1994.
29. 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.

30. Y. Shoham. Chronological ignorance: experiments in nonmonotonic temporal reasoning. *Artificial Intelligence*, 36:279–331, 1988.