Part III

Reasoning under Uncertainty

Introduction to Part III

The previous parts of this thesis have been concerned with inductive inference. Inductive inference can be seen as an aspect of *reasoning under uncertainty*, the problem of how to make rational inferences and decisions in the presence of incomplete and unreliable information.

Default Reasoning

In part III of this thesis we explore other kinds of reasoning under uncertainty. We will mostly be concerned with *default reasoning*. Default reasoning is concerned with rules that typically (by *default*) hold but that may in some exceptional cases be violated. Below we give an informal introduction to default reasoning. In an epilogue to part III (page 269) we explore connections between default reasoning and MDL.

The prototypical example of default reasoning involves the rule ([56], page 2) 'Typically, birds fly'. If we are told that Tweety is a bird, and we are then asked whether it can fly or not, we would be inclined to say that it can. If, later on, we are told that Tweety is a penguin, we would like to retract our conclusion that it can fly and instead conclude that it cannot fly after all.

Common Sense Reasoning

It has always been one of the major goals of Artificial Intelligence to formalize *common sense reasoning*, the kind of every-day reasoning that we humans perform when making plans or decisions. An important part of common sense reasoning is really default reasoning; we give a little example to illustrate this general point. Imagine a situation in which your friend gives you a phone call and asks whether you can meet him in an hour. You reason as follows: the drive to your friend takes about 45 minutes; you have to drive to the gas station first which will take an additional 10 minutes. Therefore, you say that you will be able to make it.

In your reasoning, you did not take into account any of the following eventualities: your car keys may be lost; your car may be stolen or towed away; the gas station may be closed; the road may be blocked; you may get a traffic accident etc. If you would consider all preconditions that must hold in order for you to arrive in time, you will never be able to conclude that you actually will arrive on time. Yet in practice, you do jump to this conclusion. This means that even though you do not know whether your car has been stolen, you implicitly assume that it has not; even though you do not know whether the road is blocked, you assume it is not etc. These are all default assumptions: *typically* roads are not blocked, cars are not stolen etc., which leads you to assume that your own car has not been stolen, that you will not encounter a road block etc. Assumptions of this kind are ubiquitous in common sense reasoning.

The Frame Problem; Nonmonotonic Temporal Reasoning

The following default assumption is central to common sense reasoning: most properties of the world are *persistent*, that is, they usually do not change over time. The problem of how to properly formalize this 'law of persistence' or equivalently, *inertia*, is known as the 'frame problem' [107]. We have already encountered several instantiations of this law: in the reasoning described above, you implicitly assumed that your car would still be in the same location as where you last parked it; you implicitly assumed that your car keys would still be in the same location as where you last put them; you assumed that your car would still work etc.

Solving the frame problem, that is, formalizing the law of persistence, has turned out to be a surprisingly difficult task. It is one of the main challenges of the subfield of Artificial Intelligence that is known as 'nonmonotonic temporal reasoning' or 'common sense reasoning about action and change' [135]. In the next chapters we present a fairly general theory of nonmonotonic temporal reasoning.

Formalizing Nonmonotonic Reasoning: Nonmonotonic Logic

Default rules cannot be readily expressed in classical (propositional or first-order) logic. The reason is that classical logic is *monotonic*: as your set of knowledge grows, so does the set of conclusions that can be drawn from it. On the other hand, the inferences we made above are *nonmonotonic*; in the 'birds fly' example, if we are only told that Tweety is a bird, we conclude that it can fly, whereas if we are given *additional* knowledge (namely, that Tweety is a penguin) we would like to retract our previous conclusion.

This has led researchers to search for an extension of classical logic that allows for a correct treatment of default rules [106, 109, 124, 56, 21]. The resulting logics are called *nonmonotonic*. Most nonmonotonic logics can be viewed from a unifying perspective if one characterizes them in terms of their semantics [142]: they modify classical logic by introducing a *preference order* on models. In standard logic, when given a theory T, a sentence Γ is entailed by T iff Γ holds in all models of T. In nonmonotonic logic, the models for T are (roughly speaking) ordered according to some partial order and a sentence Γ is entailed by T iff it holds in the models that are the most preferred models in this order.

The order on models induced by such a 'preferential semantics' (a term due to Shoham [142]) is often determined by one or more *abnormality predicates*. We use the 'birds fly' example for illustration. In this case we need only one abnormality predicate Ab, where $Ab(x)$ stands for the fact that x is in some sense 'abnormal', that is, 'atypical' or 'unusual'. The rule 'typically birds fly' would be expressed as follows:

$$\forall x.\ [\ \mathit{Bird}(x) \wedge \neg Ab(x)\ ] \supset \mathit{Can\_Fly}(x) \tag{7.35}$$

($\supset$ standing for standard material implication[2]). In words, if x is a bird and x is not abnormal, then x can fly. Let us consider a logical theory consisting of (7.35) together with the fact that Tweety is a bird:

$$\mathit{Bird}(\mathit{Tweety}) \tag{7.36}$$

According to standard first-order logic, there exist models for this theory in which Tweety is not an abnormal bird and can therefore fly ($\neg Ab(\mathit{Tweety}) \wedge \mathit{Can\_Fly}(\mathit{Tweety})$)

[2]So $A \supset B$ is equivalent to $\neg A \vee B$.

while there also exist models in which Tweety is abnormal and cannot fly. The idea behind 'preferential semantics' is to order the models of the theory according to how 'normal' they are. More precisely, model $\mathcal{M}_1$ is preferred over $\mathcal{M}_2$ if and only if the interpretation of Ab in model $\mathcal{M}_1$ is a proper subset of the interpretation of Ab in model $\mathcal{M}_2$. The set of allowed conclusions is then restricted to those sentences that hold in all most-preferred models. In our example, there exist models $\mathcal{M}$ in which no abnormalities at all are the case ($\mathcal{M} \models \forall x \neg Ab(x)$). Such models will be preferred over all other models. Therefore, if $\mathcal{M}$ is a most preferred model, then $\mathcal{M} \models \neg Ab(\mathit{Tweety})$ and hence, by axiom (7.35), $\mathcal{M} \models \mathit{Can\_Fly}(\mathit{Tweety})$.

Does this work?

Can nonmonotonic logics be successfully applied for any but the simplest reasoning domains? The answer depends on one's point of view.

When nonmonotonic logic was introduced in the late 1970s, it was hoped that a large class of reasoning problems could be successfully handled by variations of the simple mechanism introduced above: first add a large amount of abnormality predicates $Ab_1, Ab_2, \ldots$ to one's language, then express all default rules by axioms of the form $a(x) \wedge \neg Ab_i(x) \supset b(x)$ and then choose the models for the theory with the smallest interpretations of $Ab_i(x)$. However, in the 1980s it gradually became clear that this simple idea of minimizing abnormalities only works for a very restricted class of reasoning domains. One of the most famous counterexamples is Hanks and McDermott's 'Yale Shooting Problem' (YSP) [67], which is concerned with the law of persistence ('things usually stay the same') introduced above. It shows that a straightforward formalization of this law using abnormalities will lead to counterintuitive results for all but the simplest reasoning domains. To make things worse, the YSP turned out to be just the first member of a long chain of problematic examples (this will be discussed at length in the next chapter). When taken together, these examples strongly suggest that the original goal of nonmonotonic reasoning - employing a single, simple extension to standard logic for handling a wide variety of default rules - is impossible to obtain. In this sense nonmonotonic logic turned out to be a failure (incidentally causing one of its originators, D. McDermott, to be so disappointed that he quit the field altogether [108]). On the other hand, if one views nonmonotonic formalisms merely as *technical tools* and uses them to extend classical logic in various different ways, not just by rules of the form (7.35), then they turn out to be reasonably successful, as the slow but steady progress in the field of nonmonotonic temporal reasoning shows [151, 12, 104].

Connection to Probability and MDL; what do defaults stand for?

From the subjectivist (Bayesian) view on probability theory, probabilities can be used to model 'degrees of belief' [77, 38]: if an agent (for example, a human or a robot) assigns a high probability to proposition A, this means he (it) has a high degree of belief in A. This leads to the identification of 'abnormalities' with 'events having small probability'. Whether probability theory should or should not be used to model default reasoning is still a matter of debate (see [25] for a review of the arguments for and against probability; see also [56, 106, 116]). The appropriateness of probability is related to the question of what exactly one is trying to model when one is using default rules. In one view, nonmonotonic reasoning systems should be used by (implemented in) 'agents' (for example, robots) which have some knowledge about the world (given by axioms) and which receive additional information through their sensors. One assumes that this information is preprocessed so that it can also be encoded as axioms. The agent can then use its nonmonotonic reasoning system to arrive at sensible conclusions from its given knowledge. In this view, 'abnormalities' can be quite literally seen as 'things that happen with small probability': if the robot sees that a car is in a parking lot at time t, it should be able to infer that it will still be there at time $t = t + 1$, since the probability that this will be the case is very high (at least if the unit of time is chosen appropriately). In this view, default rules should be seen as being analogous to laws of physics: they describe how the world works. The only difference to what we usually call the 'laws of physics' is that the description is at a different level of abstraction and deals with different phenomena.

According to another point of view, defaults may also stand for *conventions* [121, 1] that humans implicitly use in communication: if one reads a story in which a bird is mentioned and no clues are given as to whether or not it can fly, the reader will typically conclude that it can fly. If one regards defaults as conventions, then it is not so clear whether or not they should be treated probabilistically.

In a very influential book [116], J. Pearl argues that uncertainty should be handled probabilistically (even when one is modeling conventions rather than 'physical laws'), and he shows how many of the problems associated with a probabilistic interpretation of defaults can be overcome. However, he concedes that in practical applications, it may often be useful to work with nonmonotonic logic instead of probability theory. He then proposes to use nonmonotonic logic after all, but with a semantics that is consistent with probability theory in the sense that it can be given a probabilistic interpretation.

In our own work on nonmonotonic temporal reasoning, to be presented in chapters 9 and 10, we implicitly take a similar (yet different) view: though all nonmonotonicity will be dealt with in entirely non-probabilistic terms, most of it can in principle be interpreted probabilistically. The reason that a probabilistic interpretation is not available for *all* our uses of nonmonotonicity will be explained in the Epilogue to part III of this thesis. There we shall argue that when nonmonotonic logic is used to model *defaults*, it should indeed have a probabilistic interpretation. However, we shall also argue that in applied work, nonmonotonicity is often used only to arrive at concise descriptions of a domain; in such cases, a probabilistic semantics is inappropriate. In the Epilogue we shall review our own theory of chapters 9 and 10 and show that indeed, all *defaults* used there can be interpreted probabilistically. We shall also demonstrate that the connection between defaults and probability arises most naturally if probabilities are treated from an MDL perspective, thereby connecting part III to parts I and II of this thesis.

Chapter 8

Introduction to Nonmonotonic Temporal Reasoning

8.1 Introduction

In this chapter we give a short overview of the field of Nonmonotonic Temporal Reasoning (NMTR). We will discuss the central problems, the main goals and the methodology of the field. The main purpose is to provide the necessary background for the following two chapters 9 and 10, in which we will develop our own theory of NMTR.

We start (Section 8.2) by introducing the field by means of an historical overview. In Section 8.3 we give a critical discussion of the current research methodology and the main research goals of the field. In Section 8.4 we discuss our own view on these issues and indicate how some of the problems raised in Section 8.3 can be overcome. We end with a summary of challenges faced by the field. For more details about both the history and the problems of the field we refer to Sandewall and Shoham's overview article [135] and Shanahan's recent book [139].

8.2 Introduction to and History of the Field

Nonmonotonic Temporal Reasoning, also called *common-sense reasoning about action and change* is a subfield of Artificial Intelligence (AI). The original goal of this field was to come up with a logic that allows for common-sense reasoning about action and change. This would be part of the grander goal of developing a general 'logic of common-sense reasoning', the central goal in the logicist view of AI [108]; see also Section 8.3. Before we start our historical overview of the field, we introduce a formalism for representing statements about action and change as sentences in first-order logic.

The Language We use a many-sorted first-order language $\mathcal{L}$. $\mathcal{L}$ contains three sorts: *fluents* (variables of the sort denoted by f), *events* (e) and *time points* (t). Fluents and events correspond to the basic entities of our domains. Fluents are those properties of a domain which usually do not change over time. Fluents can *hold* ('be the case') or not at any point in time t. If fluent f holds at time t we write $Ho(f, t)$. Actions or *events* influence the (truth) value of fluents[1]. If an event e takes place between time t and $t+1$, we will also say that e 'holds' at time t and write $Ho(e, t)$. Time-points are identified with the integers. Apart from Ho, $\mathcal{L}$ contains the following functions and predicates: '+', '=' and Ab. '+' and '=' will receive their usual interpretation. Ab is defined over fluent-time pairs. $Ab(f, t)$ denotes that there is something 'abnormal' about fluent f at time t. We will assume all formulas to be implicitly universally quantified.

The Frame Problem The field got started in 1969 when McCarthy and Hayes [107] first formulated the *frame problem*. This is the problem of how to properly formalize the common-sense law of *inertia* or *persistence*, which says that most properties of the world do not change over time, unless there is some action that affects them. This notion turns out to be surprisingly difficult to model.

As an example of what we mean by 'persistence', consider a very simple reasoning domain consisting only of a person who can either be walking or not (denoted by fluent *Walking*) and who can be alive or not (denoted by fluent *Alive*). Only one action may take place in the domain, which is to sit down (denoted by event *Sit_down*). The effect of sitting down is that the person is not *Walking* any more. This can be formalized as

$$Ho(Sit\_down, t) \supset \neg Ho(Walking, t+1) \tag{8.1}$$

Such an axiom is usually called an *effect axiom*. Let it further be given that (1) the person is alive and walking at $t = 0$, and (2) the person takes a seat at time $t = 2$; no other events take place. This is formalized as

$$Ho(Walking, 0) \wedge Ho(Alive, 0) \tag{8.2}$$

$$Ho(Sit\_down, 2) \wedge (t \neq 2 \supset \neg\exists e. Ho(e, t)) \tag{8.3}$$

These axioms are usually called *observation axioms*.

We want our domain to be subject to persistence. Hence we would like to be able to infer that the person remains *Walking* until $t = 2$, remains $\neg$*Walking* thereafter and remains *Alive* throughout. Clearly, this cannot be inferred from axioms (8.1)–(8.3) alone. We could enforce persistence by adding the following so-called *frame axioms*:

$$Ho(Alive, t) \equiv Ho(Alive, t+1) \tag{8.4}$$

$$\neg Ho(Sit\_down, t) \supset [Ho(Walking, t) \equiv Ho(Walking, t+1)] \tag{8.5}$$

However, handling persistence in this way leads to theories that are unwieldy and brittle (see [139], page 10): *unwieldy* in the sense that for each fluent, we have to

[1] In the simple domains we will encounter in this chapter we need not make any distinction between 'actions' and 'events'.

specify the complete list of actions that do *not* interfere with its persistence. In the present case this was achieved by the simple axioms (8.4) and (8.5), but when we move on to domains consisting of thousands of actions and fluents, it will make our theories inconveniently large. Our theories would also be *brittle* in two ways: first, if we want to add knowledge to a pre-existing theory, we may have to change most axioms (rather than just adding new ones). Second, our theories would become *inconsistent* if the observations would contradict persistence – as will sometimes happen in the real world (see Example 8.4, page 191). If a theory is inconsistent then all reasoning breaks down; we would prefer a system that concludes that something unexpected has happened and keeps making rational inferences about other aspects of a domain.

It seems a better idea to express the law of persistence by a single mechanism, to which only axioms of the form (8.1) and (8.3) but *not* of the form (8.4)-(8.5) have to be supplied. The obvious idea is to use a simple form of nonmonotonic reasoning (page 184) that allows one to infer that fluents do not change value whenever there is no evidence to the contrary. In terms of a preferred model semantics, this amounts to selecting only those models for a theory in which the smallest number of changes take place. However, in their classic 1986 paper, Hanks and McDermott introduced the Yale Shooting Problem (YSP) which essentially shows that this idea fails for all but the simplest reasoning domains. In order to discuss the YSP, we first need to extend our formalism so that it can handle nonmonotonicity.

Models and Preferred Models A structure $\mathcal{M}$ for our language $\mathcal{L}$ consists of universes for all the sorts and interpretations for all function and predicate constants in $\mathcal{L}$. For a function or predicate constant K, we write $\mathcal{M}[K]$ to denote the interpretation of K in $\mathcal{M}$ (which is then a function or a set, respectively). A model of a set of sentences Γ is any structure $\mathcal{M}$ such that Γ is true in $\mathcal{M}$, where truth of sets of sentences is defined as usual. If $\mathcal{M}$ is a model of Γ, we write $\mathcal{M} \models \Gamma$. To keep things simple we will only consider models in which all time points are interpreted as integers and all fluent and event constants as themselves. Moreover, we will assume that only one event happens at a time (the latter restriction will be dropped in later chapters). Our nonmonotonic formalism is simply to prefer the models of a theory T with the fewest (in the subset sense) abnormalities. Summarizing:

Definition 8.1 *A* candidate model $\mathcal{M}$ *for a theory* T *is a model satisfying (1)* $\mathcal{M} \models T$*, (2)* $\mathcal{M}$ *interprets time points as the integers and all event and fluent constants as themselves, and (3) (*$\exists!$ *denoting 'there exists exactly one):*

$$\mathcal{M} \models \forall t \exists! e.\ Ho(e,t)$$

Definition 8.2 *A model* $\mathcal{M}$ *is a* preferred model *for theory* T *iff* $\mathcal{M}$ *is a candidate model for* T *and there is no other candidate model* $\mathcal{M}'$ *for* T *with* $\mathcal{M}'[Ab] \subsetneq \mathcal{M}[Ab]$.

In what follows, we assume the set of fluent and event constants to contain just those elements that are mentioned in the relevant axioms. When we write '$\mathcal{M}'$ has fewer abnormalities than $\mathcal{M}$' we mean 'fewer' in the subset sense, i.e. $\mathcal{M}'[Ab] \subsetneq \mathcal{M}[Ab]$.

8.2.1 The Yale Shooting Problem

Using the formalism just introduced, a straightforward formalization of the law of persistence would look as follows:

$$\forall f, t \,.\, \neg Ab(f,t) \supset [Ho(f,t) \equiv Ho(f,t+1)] \tag{8.6}$$

In words, (8.6) says that if property f holds at time t and there is nothing abnormal about f at time t, then it will still hold at time $t+1$. Combining this with axioms (8.1)-(8.3) and using Definition 8.2 to pick the preferred models, we essentially choose the models with the fewest possible changes. We will call the combination of Definition 8.2 with (8.6) the *naive persistence formalization.*

Example 8.3 (Yale Shooting Problem) Assume again there is a person (in this context typically named Fred) that can be either *Alive* or not. Furthermore someone points a gun at Fred that can either be *Loaded* or not. There are three possible actions in the domain: *Load*, *Wait* and *Shoot*. A *Load* action loads the gun; a *Wait* action has no effects whatsoever; a *Shoot* action kills Fred if it is performed when the gun is loaded. We are given the following information: initially, Fred is alive and the gun is unloaded. At $t = 0$, the gun is loaded. At $t = 1$, nothing happens (a *Wait* action takes place) and at $t = 2$ Fred is shot at. We will assume our domain to be subject to persistence; this means that intuitively, the gun remains loaded during the *Wait* action; therefore, Fred is shot at with a loaded gun and will therefore not be alive any more at time $t = 3$.

The following theory $T_{\text{H-M}}$ is a straightforward formalization of our domain. $T_{\text{H-M}}$ is the union of Axiom (8.6) (persistence) together with axioms (8.7)–(8.10) below.

$$Ho(Load,t) \supset Ho(Loaded,t+1) \tag{8.7}$$

$$Ho(Loaded,t) \wedge Ho(Shoot,t) \supset \neg Ho(Alive,t+1) \tag{8.8}$$

$$Ho(Alive,0) \wedge \neg Ho(Loaded,0) \tag{8.9}$$

$$Ho(Load,0) \wedge Ho(Wait,1) \wedge Ho(Shoot,2) \tag{8.10}$$

In their YSP paper Hanks and McDermott showed that the naive persistence formalization does not handle this domain correctly: the set of *preferred* models does not coincide with the set of *intended* models . By an 'intended' model we mean a model that is intuitively admissible. Intuitively we would conclude that Fred is not alive any more at $t = 3$. Hence for all intended models $\mathcal{M}$ for $T_{\text{H-M}}$ we have $\mathcal{M} \models \neg Ho(Alive,3)$. There indeed exists a preferred model $\mathcal{M}_1$ for which this is the case. But there is also another preferred model $\mathcal{M}_2$ with $\mathcal{M}_2 \models Ho(Alive,3)$. In the model $\mathcal{M}_1$ everything happens as we would intuitively expect. It contains two abnormalities, $Ab(Loaded,0)$ (the gun becomes *Loaded* at time $t = 0$) and $Ab(Alive,2)$ (Fred dying at $t = 2$). $\mathcal{M}_2$ also contains two abnormalities: $Ab(Loaded,0)$ and additionally $Ab(Loaded,1)$: the gun mysteriously becomes unloaded again during the waiting. Therefore, Axiom (8.8) does not apply and Fred remains alive at $t = 3$. Notice that there is nothing in our axioms which is in contradiction with the spurious change of *Loaded*; moreover, both $\mathcal{M}_1$ and $\mathcal{M}_2$ have two abnormalities and, as is easy to show, there is no model for $T_{\text{H-M}}$ with only one abnormality. Hence $\mathcal{M}_1$ and $\mathcal{M}_2$ are both preferred and the intended conclusion $\neg Ho(Alive,3)$ cannot be drawn.

The YSP is not confined to the specific nonmonotonic formalism we introduced above; Hanks and McDermott showed that in most (but not all; see Section 8.2.3) standard nonmonotonic logics, a naive formalization of the law of persistence in terms of minimizing change leads to the same problem [67, 135]. Also, more sophisticated representations of time than the integer-valued time points we used here do not help [67, 135].

Solutions to the YSP The YSP has provoked a very large number of solutions [143, 83, 95, 96, 68, 7, 116, 4, 136]. In some of these a different pre-existing nonmonotonic logic was used for which the problem did not occur after all; in other approaches, more complicated persistence axioms involving new predicates were introduced and/or new and more complicated minimization policies were proposed. We mention *chronological minimization* [83, 96, 143], *causal minimization* [95, 68, 112, 98, 51], *logic programming* [4, 45], *counterfactual* (also known as 'state-based') [7, 8, 81] and *explanation closure* [136, 101] approaches; there are many more. Apart from these approaches, which were all developed within the nonmonotonic reasoning community, a fundamentally different approach based on Bayesian networks and Pearl's ideas on causality started to evolve in the probabilistic reasoning community [116, 51, 33].

The Cyclic Pattern All these approaches had two things in common: (1) they solved the original YSP, but (2) they were soon confronted with their own 'counterexamples', that is, examples of very simple reasoning domains for which they gave counterintuitive results. This led to a cyclic pattern, where counterexamples triggered new proposals, which in turn triggered new counterexamples etc. To give the reader a feel of what happened, we will briefly sketch two of the better known approaches that were proposed: *chronological minimization* and *logic programming* approaches.

8.2.2 Chronological Minimization

The first three proposals to solve the YSP [83, 143, 96] were all based on essentially the same idea: reasoning should follow the direction of time, that is, the default assumptions should be applied in temporal order. In our semantical setting, this 'chronological minimization' (a term due to Shoham [143]) amounts to the following: instead of simply choosing the models with the fewest abnormalities, we select the set of our preferred models in a 'pointwise' fashion: from the class of all classical models consistent with a theory T, we first select the set with the fewest abnormalities at time $t = 0$. Among the remaining models, we select those with the fewest abnormalities at time $t = 1$; we then further restrict ourselves to those with the fewest abnormalities at $t = 2$ etc. The reader may (easily) check that this minimization policy indeed solves the YSP, preferring only the model $\mathcal{M}_1$ with $\mathcal{M}_1 \models \neg Ho(Alive, 3)$. However, it fails to give the intended models for almost every other reasoning problem proposed thereafter. Here is a simple example:

Example 8.4 [Stolen Car Problem] In Kautz' Stolen Car scenario [83], a car is left behind in a parking lot at the initial point in time $t = 0$. The car driver leaves the lot and then returns at some later point in time, say, $t = 10$, to find that the car is not there any more (it might have been stolen, or towed away for example). We model this domain using a theory consisting of persistence axiom (8.6) and the following observation axiom:

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

If we determine our preferred models according to chronological minimization as defined above, we find that in all selected models we have $Ho(Car\_In\_Lot, t)$ for all $1 \leq t < 10$: we can infer that the car disappears as late as is consistent with the observations, which clearly does not correspond to our intuitive understanding of the situation.

Sandewall [131] proposed to overcome the Stolen Car and similar problems by supplying chronological minimization with a trick called 'filtered preferential entailment' [135]. This yielded the intended models for the Stolen Car scenario but it still failed on some other domains. For this reason, another concept called 'occlusion' was invented and put on top of chronological minimization and filtered preferential entailment. At that point, the development took an interesting turn to which we shall return later.

8.2.3 Logic Programming Approaches

In this example we assume the reader to be familiar with the basics of logic programming; see for example [3]. Most approaches to Logic Programming use *negation as failure* which is an inherently nonmonotonic construct; see [10]. However, the nonmonotonicity is achieved in a manner quite different from our simple minimization of abnormality; we may therefore wonder what happens when temporal reasoning domains are expressed as logic programs.

Interestingly, as noted independently by Apt and Bezem [4], Elkan [43] and Evans [45], if the Yale Shooting *domain* is formalized as a logic program then the Yale Shooting *problem* does not occur. We now show the logic program P, a variation of Apt and Bezem's formalization that has been adjusted to our use of integer time. We adopt PROLOG syntax: capitals stand for variables; `:-` stands for implication and `not` stands for negation-as-failure. `s(T)` denotes the successor of `T`, i.e. `T + 1`. `ho_f` denotes the *Ho*-predicate for fluents; `ho_e` denotes the *Ho*-predicate for events.

```
ho_f(F,s(T))       :- ho_f(F,T), not ab(F,T).

ho_f(loaded,s(T)) :- ho_e(load,T).
ho_f(dead,s(T))    :- ho_e(shoot,T), ho_f(loaded,T).
ab(alive,T)        :- ho_e(shoot,T), ho_f(loaded,T).

ho_f(alive,0).
ho_e(load,0).    ho_e(wait,s(0)).  ho_e(shoot,s(s(0))).
```

Anyone familiar with PROLOG will see that the answer to either of the queries `ho_f(alive,s(0))` and `ho_f(alive,s(s(0)))` will be 'yes' while the answer to the query `ho_f(alive,s(s(s(0))))` will be 'no'. Apt and Bezem showed that the program P also solves the shooting problem from a semantical point of view: several different ways of defining the semantics of logic programs all lead to the adoption of the same unique model $\mathcal{M}$ for P. This is the intended model, satisfying $\mathcal{M} \models \texttt{ho\_f}(\texttt{alive}, \texttt{s}(\texttt{s}(0)))$ and $\mathcal{M} \not\models \texttt{ho\_f}(\texttt{alive}, \texttt{s}(\texttt{s}(\texttt{s}(0))))$.

As formulated above, all negation is handled as 'negation as failure' in this approach. This is undesirable in general; for example, it is not immediately clear how to represent versions of the YSP in which it is not known whether the gun is initially loaded or not. To see why, note that in the program P listed above the answer to the query `ho_f(loaded,0)` will be 'no' though we have not specified anything about `loaded` at time 0 explicitly.

To address this kind of problems, Gelfond and Lifschitz [53] introduced 'extended logic programs'. These are capable of dealing with both negation-as-failure and classical negation. However, the class of reasoning domains that could be handled by a basic formalization of persistence in an extended logic program was still quite restricted (for example, it had problems with the Stolen Car domain). To be able to handle larger classes of reasoning domains, the logic programming community has gradually adopted a different methodology altogether, as we shall see in Section 8.3.1.

8.2.4 The Ramification Problem

We have just seen two typical examples of the kinds of solutions that were proposed, and one typical example of the kind of counterexamples (the Stolen Car problem) this led to. As the cycle between solutions and counterexamples continued, it became increasingly clear that the most challenging of these counterexamples were all interrelated: they were all instances of the *ramification problem* [135]. This is the problem of how to properly deal with the side effects of actions. Sometimes it is viewed as part of the frame problem, sometimes it is seen as a problem in its own right [135]. It occurs in domains where an action or event has only a small number of *direct* effects while it can have a very large number of *indirect effects*. For example, if a person is shot at, he/she will stop being alive (a direct effect), but he/she will also stop walking, breathing, eating etc.; these are indirect effects. Theories in which all such indirect effects of an action are stated explicitly would quickly become unmanageably large; one would prefer to state them in more general and implicit ways. Here we will give just one example (due to Baker [7]) of the ramification problem - there are many more, all of them highlighting different aspects of it; some of these will be discussed in the next chapter, Section 9.5.

Example 8.5 [The Walking Turkey - Ramification] Imagine a turkey that can be either *Alive* or not and that can be either *Walking* or not. If the turkey is shot at (we assume a gun that is always loaded here) it will stop being alive. Initially, the turkey is alive. Then somebody shoots at the turkey. We can formalize this as follows:

$$Ho(Alive, 0) \wedge Ho(Shoot, 0) \tag{8.11}$$

$$Ho(Shoot, t) \supset \neg Ho(Alive, t+1) \tag{8.12}$$

We now want to encode the additional rule that if the turkey stops being alive, it will also stop walking. In light of the discussion above, we do *not* want to add a rule of the form $Ho(Shoot, t) \supset \neg Ho(Walking, t+1)$; such a methodology would lead to overly large and inconvenient theories in domains where there are several actions with direct effect $\neg Ho(Alive, t)$ and/or where $\neg Ho(Alive, t)$ has many indirect effects. Instead, it seems more reasonable to try the following *state constraint axiom*:

$$\neg Ho(Alive, t) \supset \neg Ho(Walking, t) \tag{8.13}$$

Unfortunately this will not work in general. We note first that our naive persistence formalization will lead to counterintuitive results, as the reader may (easily) check: let T be the union of (8.11)–(8.13) and persistence axiom (8.6). By axioms (8.6), (8.11) and (8.12), the preferred models for T all have abnormality $Ab(Alive, 0)$. There is only one model $\mathcal{M}$ with no further abnormalities, therefore, $\mathcal{M}$ is the single preferred model. By (8.13), $\mathcal{M} \models \neg Ho(Walking, 1)$. By persistence, $\mathcal{M} \models \neg Ho(Walking, 0)$. This means that, while nothing is known about whether the Turkey is initially *Walking* or not, we can infer that it is not, clearly an unintended (overly strong) result. The two other approaches introduced here also have problems with this example, but we shall not discuss these further.

It seems we should look for a solution in which the fluent *Walking* is somehow 'released' that is, not subjected to persistence, whenever *Alive* is affected by an action. This kind of solution was indeed proposed [135], but was later found to be inadequate by several authors [99, 103]. Briefly, the reason is that there may be other fluents in a domain for which the domain constraint $\neg Ho(A, t) \supset \neg Ho(B, t)$ should also hold in all intended models, but for which B should *not* be released from persistence when A is affected by an action. In the next chapter, Section 9.5 this will be discussed in detail.

8.2.5 Ramification and the Resurgence of 'Causal' Approaches

Let us now return to the historical development. The ramification problem was identified as a problem shortly after the YSP was presented (1986). From that time onwards, reasoning domains involving it have been gradually accumulating. A few early solutions were proposed [7, 97], but none of them of any reasonable generality. Then, around 1995, several authors, apparently independently, succeeded in defining relatively simple and elegant formalisms that where sufficiently general to account for a substantial number of examples (including the Walking Turkey) [99, 103, 66]. These formalisms were all closely related to one another; in the following we concentrate on one of them, namely Lin's [99] approach. It is based on the intuition that some effects of actions *cause* other effects while for other effects, this is not the case. Lin introduces an additional predicate *Caused*. To give an example, in Lin's formalization, the fact that $\neg Walking$ is an indirect effect of $\neg Alive$ would be expressed as:

$$\neg Ho(Alive, t) \supset Caused(Walking, \text{FALSE}, t) \tag{8.14}$$

to be read as '$\neg$*Alive causes* $\neg$*Walking*'. This proposal is related to the early *causal minimization approaches* to the Yale Shooting Problem. These used a causal predicate and a philosophy similar to Lin's. However, there were a few small but essential differences which allowed Lin to solve a much wider class of domains than these forerunners. In Chapter 10 we shall treat Lin's proposal in detail.

Interestingly, around the same time as Lin's approach was published, McCain and Turner [103] independently made a very similar proposal. Moreover, several of the originally non-causal approaches to the frame problem seemed to develop naturally, as it were, into approaches that were closely related to Lin's approach. We already mentioned that in the *chronological minimization* school, the concept of 'occlusion' had been introduced. Later, Gustafsson and Doherty [66] showed that one can also treat persistence and ramification based *only* on occlusion *without* combining it with chronological minimization. They showed that the resulting theories are closely related to Lin's. In the *logic programming* community, people gradually came to adopt a different methodology which we shall describe in Section 8.3.1. The methodology formalizes domains in some 'high level action description language'. Such languages invariably have a primitive **causes** which functions in a way similar to Lin's *Caused*-predicate [54, 57, 12]. In Shanahan's*event calculus-based approach* [139], a construct *Initiates* is used in way similar to Lin's use of *Caused* ([139], Chapter 16). Finally, Lin's own approach really had developed as an extension of the *explanation closure approaches* of Schubert [136] and Lin and Reiter [101] rather than one of the original causal approaches.

The Present Situation Summarizing , many different research groups nowadays use a construct similar to Lin's *Caused*. Approaches to persistence developed in the probabilistic community are all based on some version or other of Pearl's ideas about causation. It seems that among all the originally proposed approaches to persistence, the causal approach is the clear winner.

However, *Caused* is just a predicate with a certain semantics. It has been called '*Caused*' because it can often successfully be used in contexts where $\mathcal{M} \models \mathit{Caused}(f,t)$ seems to correspond to the statement 'fluent f is caused to hold at time t'. Unfortunately, causality is a notoriously difficult concept and it seems quite unclear what such a statement really means. Gustafsson, Doherty [66] and Shanahan [139] do not interpret their *Caused*-like constructs as being necessarily related to causation. Thielscher [151] even shows that the *Caused* predicate can sometimes be fruitfully used in a context where it does *not* correspond to the use of 'causes' in natural language (see the next chapter, Section 9.5.3). At the time of writing this thesis, one of the main concerns of the field is the question whether the causal road is really the one to take.

8.3 Research Goals, Criticisms and Methodology

Having introduced the field by giving examples of the problems it is dealing with, we now stand back a little and discuss the major research goals and the methodology of the field. Both of these have been subject to much criticism [108, 135]. Recall that the original goal was to develop a theory of action and change as part of a grand 'logic of common-sense reasoning'. Later, when it was realized that this goal was hard or even impossible to achieve, many researchers scaled down their ambition.

The Research Goal The goal now became to find an *engineering solution*: to develop a formalism (preferably, but not necessarily a logic) capable of handling a reasonably large class of reasoning domains involving action and change. Here 'handling' is meant in the following sense: the domains are represented in a *clear* and *concise* manner, the inferences that can be made from a representation of a domain are all *correct* (in the sense that they agree with our intuitive understanding of the domain) and *efficient* (in the sense that relevant inferences can be made in a small number of computation steps). The system developed could then be installed as a specific component of the reasoning- and planning system of some intelligent agent (e.g. a robot). Even the achievability of this more modest goal can be questioned. The main problem is that for sufficiently complex reasoning domains, it is really not clear what a 'correct' representation of such a domain would be. We discuss two reasons:

Clashes of Intuitions A 'correct' representation of a domain should allow us to make those inferences that we consider *intuitively* correct. Unfortunately, different researchers often make incompatible statements about what are the admissible (intuitive) common-sense inferences for a given example. In the next chapter (page 218 and 221) we shall see examples of reasoning domains where different people think that different, conflicting inferences are admissible. More in general, one sometimes gets the feeling that the community really is trying to formulate a logic that permits those inferences that are judged admissible by the members of the community itself (and not necessarily by others) (see [156]).

Arbitrariness of Translations Formalizing a domain always involves a step where our intuitive understanding of the domain is translated into a formal representation of it. This step *itself* is, however, necessarily informal: there is no formally defined translation function that, when input a description of a domain in natural language, would output a description of the same domain in the many-sorted first-order language we use in this chapter. The same holds of course for any other formalism. The inherent informality makes it unclear what parts of a problem are really solved by a clever translation and what parts are solved by the formalism itself. It then, once more, becomes very hard to judge whether a given formalism can be considered adequate or not: if somebody dismisses it on the grounds that it leads to counterintuitive results, somebody else may always claim that the formalism can handle the task after all; one merely has to express the domain knowledge in a different way.

A third criticism is directed against the methodology of the field rather than its goals:

Chaotic Methodology Even *if* the overall goal of the field can be achieved, it is very doubtful whether we get any closer to it by accumulating more and more specific examples of problematic domains and adjusting existing formalisms to cope with them. All past proposals were soon confronted with simple counterexamples (page 191). This suggests that if an approach handles a bunch of specific examples correctly, this does not give any guarantee that it really handles a 'wide class' of domains correctly.

In reaction to these and similar criticisms, an alternative methodology, often called 'systematic' [135] has developed in recent years. As we shall see, this methodology partly (but only partly) resolves the issues raised above.

8.3.1 The Systematic Methodology

Originally proposed around 1992 [132, 102, 54], the systematic methodology nowadays underlies the work of a substantial part of the researchers in this field. There are two main variations of this methodology, one developed by Sandewall [132] and one by Gelfond and Lifschitz [54]. We will concentrate on the latter variation, which has mainly been developed in the context of the logic programming approaches but which by now has been applied to many other approaches as well [80]. Roughly, it works as follows: instead of directly formalizing domain knowledge as axioms in some logic, one uses a high-level 'action description language' (ADL) [54] with a restricted syntax. The idea is that by using a restricted, tailor-made language rather than a full first-order language, it is possible to endow such a language with very simple and clear semantics. We shall see an example of such a language, the language $\mathcal{L}_3$, in Chapter 10, Section 10.4.

ADLs together with their semantics are regarded as *definitions* of a class of reasoning domains. The language defines what reasoning domains belong to the class (namely, precisely those that can be represented by the language). The semantics determine for every element of the class what the correct inferences for that element are. For any given ADL+semantics and any given logical formalism L, one can define a formal translation function π mapping domain descriptions written in the ADL to domain descriptions written in the language of L. The ADL+its semantics + a translation function 'resolve' all three issues raised above: there can be no clash of intuitions about what is correct and what not, since this is *defined* by the semantics of the ADL. There is no informal translation step any more, for translations are done by a precisely defined function π. Finally, one can often formally *prove* that a proposal L in combination with π is correct for a whole *class* of reasoning domains by proving that for every domain description D written in the ADL, the models assigned to D (by the semantics of the ADL) correspond to the models assigned to $\pi(D)$ (by the semantics of L).

Begging the Question Unfortunately, this solution evades all the difficulties rather than faces them: they all re-occur one level higher, in the question of how to define the proper semantics for an ADL; but this question is simply moved outside the scope of investigation. Nevertheless, the idea of ADLs, in our opinion, has two merits: first, by the restricted syntax of an ADL, it becomes much easier (though still hard) to judge whether it assigns the 'correct' semantics to all the problems that can be represented in it. Second, the idea of formally translating descriptions of approach A to descriptions of approach B is, we think, fundamental to any advances of the field. It allows us, when introducing a new approach, to show exactly in what ways it is related to previous approaches.

8.4 An Alternative View: NMTR as Modeling

In the previous section we saw some serious criticisms of the goals and methods of our research field and we presented a methodology that partially answers them. As we see it, at least part of the criticisms can be better met by a more radical shift of the focus and goals of the field. In this new view, the theories to be developed should be *models* about the *physics* of action and change. To explain what we mean by this and why it is really different from the research goal as stated before, we need to say a little about modeling in general.

8.4.1 Two Aspects of Modeling

In our view, the field of NMTR is basically concerned with finding good Models for scenarios concerning action and change. A logical theory (i.e. a set of axioms) together with a nonmonotonic semantics can be considered as a *Model* of the domain of interest (from now on we write the word 'Model' with capital M to indicate that it is used not in the logician's, but rather in the general sense of a 'model of some phenomenon'. When written uncapitalized, it denotes the logician's notion of 'mathematical structure in which all axioms of some theory hold'). A Model of a domain can be specified by giving the list of laws to which the basic entities in the domain are subject (or, more correctly, *would be* subject if the Model where in any sense 'true'). However, for many interesting domains, this list will be very long. Yet it is often possible to specify a Model in a concise manner, by using some precisely defined convention about how the short specification can be translated into the full list of laws corresponding to it. In such a case, there are two fundamentally different aspects to the process of modeling:

1. Finding a good Model for a domain, that is, identifying a list of 'laws' to which the domain is subject.
2. Finding a method to specify such a Model in a compact manner.

Much of the NMTR literature can be characterized by a confusion of these two issues. Yet (at least in our view) they should be clearly separated in order to make any real progress in the field. Perhaps the best example of this confusion is given by what we called the 'naive persistence formalization', that is persistence axiom (8.6) together with the minimization of Ab. Let us stand back a little and analyze what rôle the researchers who proposed it had intended it to play. We see that it was meant to capture some quite different things at the same time:

a. *Modeling* persistence per se: if at time t no action takes place that affects fluent f, then the value of f usually remains unchanged at time $t+1$. For example, this says that if Fred is alive and none of the actions that can kill Fred take place at time t, then he will usually still be alive at time $t+1$.

b. *Expressing* that actions affect no more things in the world than those we explicitly state they affect. For example, if a *Load* action takes place, then this should have no effects on *Alive* ; if a *Wait* action takes place, this should neither have an effect on *Loaded* nor on *Alive*, etc.; yet none of this is stated explicitly. These laws about non-effects are assumed to be enforced by the persistence axiom and the minimization of Ab.

(a) can be seen as a kind of 'physical law'. It is part of a Model of the domain as referred to in item (1) above. (b) is an attempt to express some of these laws (namely, the laws involving the non-effects of actions) in a very concise manner. As such it is related to item (2) above. Item (1) is about determining which things are generally the case in our domain of interest; item (2) is about determining how to specify these things.

8.4.2 Models, models, and Two Uses of Nonmonotonicity

To clarify our distinction we will give an alternative, more precise characterization of what we mean by a Model and by a 'concise representation' of it.

We first define the notion of 'Model' in a simplistic way; a more realistic definition follows later. For now, a (modeler's) *Model* of a domain is *defined* as a *set* of (logician's) models. Intuitively, each model $\mathcal{M}$ in the Model $\mathbf{M}$ can be identified with a single instantiations of the domain. For example, the 'intended' model $\mathcal{M}_1$ with $\mathcal{M}_1 \models \neg Ho(Alive, 3)$ (page 190) should be an element of any reasonable Model of the Yale Shooting domain.

Let $\mathbf{M}$ be a Model of some domain defined in this way. For a sentence Γ we write $\mathbf{M} \models \Gamma$ if for each model $\mathcal{M} \in \mathbf{M}$ we have $\mathcal{M} \models \Gamma$.

The Contingent and the General; Generic Models Now note that the problem domains we deal with in NMTR can usually have several instantiations: typically there are *general laws* (that hold in all possible instantiations of the domain) and *contingent facts* (whose truth value differs from instantiation to instantiation). For example, in the Yale Shooting domain, a sentence like $Ho(Loaded, 0)$ is a contingent fact. A sentence like

$$Ho(Load, t) \supset Ho(Loaded, t+1)$$

is a general law.

For simplicity, we will only consider 'generic' Models of domains. These are Models in which none of the contingent facts have been instantiated (for example, a 'generic' Model of the Yale Shooting domain should contain both a model $\mathcal{M}_a$ with $\mathcal{M}_a \models Ho(Alive, 0)$ and a model $\mathcal{M}_b$ with $\mathcal{M}_b \models \neg Ho(Alive, 0)$).

Theories as Specifications of Models Generic Models can be characterized by a set of logician's models, but, at least if they are not too complicated, also by a logical theory. We call a theory T a *sound specification* of a Model $\mathbf{M}$ if, for all sentences Γ,

$$T \models \Gamma \Rightarrow \mathbf{M} \models \Gamma \tag{8.15}$$

We call a theory T a *complete specification* of a Model $\mathbf{M}$ if, for all sentences Γ,

$$\mathbf{M} \models \Gamma \Rightarrow T \models \Gamma \tag{8.16}$$

We call T an incomplete specification of $\mathbf{M}$ iff it is not a complete specification of $\mathbf{M}$.

One of the main goals of NMTR is to specify the Models of our reasoning domains in a compact manner (recall that we wanted to avoid using all the 'frame axioms' (page 188)). This is almost always done by giving a theory T that is a sound but incomplete specification of the Model $\mathbf{M}$ we have in mind. This theory is then turned into a complete specification by using some nonmonotonic entailment relation $\approx$ ($T \approx \Gamma$ denoting that Γ holds in all most-preferred models of T). Hence, in the practice of NMTR one uses a theory T to specify $\mathbf{M}$ which is such that (8.15) holds, (8.16) does not hold, but it *does* hold for all sentences Γ that

$$T \approx \Gamma \Leftrightarrow \mathbf{M} \models \Gamma \tag{8.17}$$

There would be nothing wrong with this, were it not for the fact that a good Model for the kind of domains we are actually interested in is usually not so simple that it can be identified with a plain set of logician's models. We often want such a Model to have a preference structure over the possible instantiations (logician's models) of the domain. In such a case, a Model $\mathbf{M}$ becomes an *ordered* set of logician's models rather than a plain set. For example, the generic Model of the Stolen Car domain (page 191) should contain a model $\mathcal{M}$ with $\mathcal{M} \models \forall t.\neg Ho(Car\_In\_Lot, t)$ among its most-preferred models; but, among the less preferred models, it should also contain a model $\mathcal{M}'$ with $\mathcal{M}' \models Ho(Car\_In\_Lot, 0) \wedge \neg Ho(Car\_In\_Lot, 10)$: after all, if we want to model the fact that persistence *may* be broken without discernible reason, then $\mathcal{M}'$ *should* be a possible instantiation of the domain.

Hence, more realistic Models $\mathbf{M}$ can be identified with a set of logician's models together with a preference ordering over these models. Let us call the preference ordering of this kind that belongs to a Model $\mathbf{M}$ the *internal preference structure of* $\mathbf{M}$. In contrast, we will call the nonmonotonic mapping performed by the relation $\approx$ of Equivalence 8.17 a *completion mapping*, since it is meant to turn theories that are incomplete specifications of a Model $\mathcal{M}$ into complete specifications of $\mathcal{M}$.

The Confusion Both completion mappings and internal preference structures are nonmonotonic constructs. In many approaches to NMTR, they are implemented by the same, single, mechanism. We already saw the example of the naive persistence formalization, where all nonmonotonicity is handled by minimization of Ab. We claim the same thing happens in chronological minimization, early causal approaches and in state-based approaches [83, 96, 143, 95, 7]). In our view, the problematic behaviour of these approaches is partly due to their treating both uses of nonmonotonicity by a single mechanism. For these two uses are really very different: the nonmonotonicity inherent in the completion mapping is concerned with how the Model of the domain is to be specified, while the internal preference structure (which is defined relative to a model $\mathbf{M}$) is *part of that Model itself.*

Further Uses of the Distinction In the next chapter, we shall present our own theory of nonmonotonic reasoning. In this theory the two different uses of nonmonotonicity will be strictly separated. In the conclusion to the next chapter (page 234) this will be verified and some examples will be given. In the Epilogue to part III, we shall further see that the use of nonmonotonicity to make incomplete specifications complete is *not related to default reasoning.* Finally, we shall point out there how the distinction between the two uses helps in understanding the relationship between probability theory and nonmonotonic logic.

8.4.3 A New Research Goal

The distinction introduced above leads to a different view of the goals of the field. The challenge now becomes:

Alternative Research Goal First, to find good Models for domains of action and change; and second, to find concise and computationally efficient representations of these Models.

A Model should constrain the possible instantiations of a domain so that they obey some set of 'laws of nature'. A concise and efficient representation of a model **M** MAY BE a logical theory T that incompletely specifies $\mathcal{M}$ together with a completion preference ordering $\models_{\approx}$, but we see no reason why to constrain ourselves in this way.

The 'laws of nature' implied by a model should be transparent: it should be clear what assumptions one has to make about the world in order for them to hold. For example, the persistence law that we will formulate in our own theory of NMTR (Section 9.6) can be read as follows:

> If a fluent f holds at time t and there are no interventions in the domain at time t that can influence the value of f, then with high probability f still holds at time $t+1$.

We call this a 'physical law' since it describes what really happens in the world if it is regarded at a certain level of abstraction. To clarify, we give an example of a 'law' used in the existing literature that clearly can*not* be viewed as a physical law:

> A fluent f changes value as late as possible.

This is just 'chronological minimization[2]' discussed in Section 8.2.2, page 191. Whereas our own formalization of persistence applies for many properties f and events e that exist in the real world, the second formalization, at least in our view, does not.

If, as in chronological minimization, one does not separate the issues of concise representation and Modeling, then the preference ordering over models one comes up with needs to serve two purposes at once; therefore, even if a single preference ordering of this kind exists that leads to intuitive results, it will probably not correspond in any way to 'what happens in the real world'.

The 'modeling' view of the field is adopted not only in our own work. It is also implicit in the works of some of the other researchers in the field; we mention Sandewall's 'trajectory semantics' [133] and Thielscher's work on causality [151]. The work on action and change done in the probabilistic community [33, 34, 52] invariably takes a 'modeling' perspective.

[2] Of course, the inventors of chronological minimization did not believe that fluents change as late as possible in the real world. Rather, they wanted to formalize the idea that when people *reason* about action and time, they sometimes do so in temporal order. However, the law that fluents change as late as possible *is* a direct consequence of the formal definition of chronological minimization.

8.4.4 The Clash of Intuitions Revisited

In Section 8.3 the field was criticized because of the inherent difficulty in determining whether a proposed formalism is really 'correct', the reason being that different people may have conflicting intuitions about what is and what is not an admissible inference for some given domain.

We think these 'clashes of intuitions' are better understood, and can be partially alleviated, if one adopts the alternative research goal as stated on the previous page. In our view, there are really two different kinds of 'clashes of intuitions':

1. Clashes concerning the question 'what is a *good Model* for a given domain', which is the central question of the first part of the research goal.

2. Clashes concerning the question 'what is a *concise yet intuitive way to specify* a Model of a domain?', which is a question related to the second part of the research goal.

The clashes of intuitions of kind (1) are problematic, but they seem not much worse than clashes of intuitions that appear when Models are developed in fields different from NMTR, i.e. economics, biology. As in these fields, one can use, at least in principle, empirical evidence to determine what a good Model is: when in doubt, one can think of a robot, equipped with the Model under investigation. The question of whether a Model is 'correct' then becomes: will the robot draw sensible conclusions and function sensibly, in the domains we intend to place it in?

The clashes of kind (2) are more serious; they are clashes of a kind that simply does not seem to occur in other fields concerned with modeling (at least in statistics, Models are *always* described by a *complete* specification). In our view, there are *bound* to be differing intuitions here. For this reason, we think it is a good idea to drop the requirement of intuitive specifications: in our view, in a proper approach to the field, the mapping from concise representations to Models should just be *simple* and *precisely defined*, but not necessarily intuitive - simplicity and precision are all we can ask for.

Let us explain this a bit further. We think that the actual clashes of intuitions thus far encountered in the literature are really mainly about clashes of kind (2). Recall that, in order to arrive at a compact representation, the general approach of the field has been to specify only *some* of the laws that should hold according to the Model. To make this work, one needs precisely defined rules (for example, a nonmonotonic semantics) to map an incompletely specified theory to the Model it is supposed to correspond to. Something similar occurs in natural language, where it is commonplace to leave out certain things from a description. Most speakers of a language use approximately the same *narrative conventions* [1] to (unconsciously) determine what somebody really means if he/she gives an incomplete description of a situation. But when an incomplete description of a Model is expressed in the language of first-order logic (i.e. as a logical theory), it seems quite strange to assume that different users of this (artificial) language will share the same intuitions about how incomplete descriptions should be mapped to Models: first-order logic is a recent, man-made invention. It would be really surprising if a set of 'logical narrative conventions' existed that would be judged intuitive by most or all people. This is especially so since, as we shall see in the next chapter, the rule 'if an effect of an action is not specified, then it is not an effect', is not enough; as domains get more complicated, much more complicated conventions are needed, at least if concise theories (descriptions) are what is asked for.

Summarizing, we think that by simply not attempting to do the compact specification of Models in an intuitive manner, the clashes of intuitions are reduced to those occurring for complete specifications; but there, they will not be much more serious than in other fields, and, at least in principle, there are empirical means (by testing a Model or an actual robot equipped with it in actual domains) to help resolve such clashes.

8.5 Summary; Main Challenges

We end this chapter with a summary and an overview of what we regard the main challenges of the field. We have seen that NMTR is based on an elusive goal, the 'logic of common-sense'. Many problematic reasoning domains show that this goal is not at all easy to achieve. Most modern approaches dealing with these problems are interrelated in that they are based on explicitly expressing causal relationships. One of the main current challenges of the field is therefore, in our view:

- Find out the exact relations between all these approaches. Gain a better understanding of the use they make of causality.

More generally one may question some of the assumptions behind the goals and methods of the field. This leads to what we personally regard as another fundamental challenge (though, unlike the first, most researchers seem not to be too concerned about this issue):

- Search for a better methodology; gain a better understanding of the possibilities and limitations of the field.

In Section 8.4 we proposed a different, more *modeling*-oriented view of the field, based on separating the modeling of a domain and finding a concise representation of a Model. In the next chapter, we shall present a theory of NMTR that adheres to this separation. Its treatment of causality is based on a successful approach developed outside the field. Hence, our approach will have implications for both challenges mentioned here.

Chapter 9

Causal Theories for Nonmonotonic Temporal Reasoning

In this chapter we present our own theory of Nonmonotonic Temporal Reasoning. Our theory is based on an approach to causality that has been developed in a statistical rather than an NMTR context. Nevertheless it turns out to be closely related to several recent approaches developed within the NMTR community. Before presenting our theory, we give an introduction in which we discuss what we wish to achieve with our theory and what its basic features are.

9.1 Introduction

By and large, the most successful existing approaches to NMTR are 'causal' in the sense that they use a predicate or logical connective to explicitly express causal relationships. [103, 105, 104, 99, 151, 12, 52]; we discussed this at length in the previous chapter (page 194). Causal approaches seem required especially in domains involving the ramification problem. It has even been argued [151] that the explicit modeling of causal relationships is a necessary ingredient of any solution to the ramification problem. Still, witness the frequently heated debates on the issue taking place during workshops and conferences, it seems that many researchers remain skeptical about the power and applicability of causal approaches. A reason for this may be that hardly any of these approaches has addressed the basic issue of causality directly: what has to be the case in the world in order for the assertion 'A causes B' to be valid? In other words, it is not attempted to define causal relationships in non-causal terms. Instead, causal domain knowledge is formalized explicitly, typically as rules of the form 'A causes B' – and the approaches rely on the assumption that their semantics will yield the right inferences from such assertions.

On the other hand, J. Pearl has provided an empirical semantics for causation that has met with considerable success in statistical applications [117, 118, 119, 120]. However, as will be explained below, Pearl's theory as such cannot be directly applied to common-sense reasoning about action and change. The theory we present in this chapter extends Pearl's theory in a straightforward way so as to overcome this hindrance. In this way we arrive at a causal theory that has the advantage that it can also be understood in non-causal terms, namely as a theory that gives a certain semantics to 'actions'. The theory will be capable of handling wide classes of reasoning domains, specifically involving *persistence*, (complicated) *ramifications* (page 193), actions with *disjunctive and non-deterministic effects*, domains with incomplete specification of the events that occur and *concurrent events*, furthermore *unknown fluent dependencies*, *causal chains of events* and *surprises.* All these terms will be explained in due course. In this chapter we will show that the theory gives intuitive results for specific examples of domains with all these features; in the next chapter we shall provide evidence that it deals well with complete *classes* of reasoning domains involving most of these features.

As we will see, the theory turns out to be closely related to, yet different from several other existing approaches. The theory will therefore allow us to interpret these approaches (and specifically, the use they make of causality), in a new way. Summarizing:

Research Goal Our aim is to develop a theory that allows us to

1. gain a better understanding of the rôle of causality in NMTR by showing its relation to Pearl's theory of causation; specifically, we will show *how causal notions can be understood in non-causal terms.*

2. deal with a wide range of domains involving action and change; moreover, to provide evidence that the way we deal with these domains is, in some sense 'correct'.

3. show how some of the existing approaches are interrelated and explain their successes and failures.

These goals will be pursued in the context of the more general research goal we formulated in the previous chapter (Chapter 8, Section 8.4): we separate the issues of finding a good model for a domain and compactly specifying such a model.

Structure of this Chapter

In the present chapter we focus on the development of our theory. We show how it arises as an extension of Pearl's theory and we illustrate its use by applying it to several reasoning domains. At the end of the chapter we shall return to the goals mentioned above and see whether they have been achieved. A truely detailed comparison to other approaches will be the subject of the next chapter.

In the next sections we will gradually build up our theory. In Section 9.2 we provide an informal introduction to Pearl's semantics of causation. We then introduce 'causal theories' which are a simple, propositional extension of Pearl's causal theories. In Section 9.4 we reach our first 'milestone' when we instantiate our theories to domains involving persistence. The resulting theories are capable of representing and dealing with most aspects of the ramification problem as well as the problem of handling disjunctive and indeterministic effects of actions. We illustrate this by applying the theories to several examples in Section 9.5.

In order to represent an arbitrary number of time points, our causal theories must be extended to the first-order case. This allows us to treat a much wider variety of domains. Section 9.6 discusses the extension to the first order case; it is shown how this allows us to handle the original YSP and related problems, 'dependent' fluents, concurrent events, causal chains of events and surprises. This is followed by a conclusion. The chapter ends with an appendix in which we show formally how our theory arises as an extension of Pearl's.

The End Before the Beginning

Since we will develop our theory gradually, starting with a simple version and then adding more and more features as we encounter more and more complicated examples, it may be useful to first give an idea of what we will end up with. Our general theory will be instantiated to several 'causal theories'. The most sophisticated version of causal theories will be given in Definition 9.27 on page 232. In the box on page 208 we provide an outline of how these sophisticated causal theories will look like.

9.2 Informal Introduction to Causal Theories

9.2.1 Actions Remove Some Dependencies, but Keep Others Intact!

Causal theories are about domains of variables whose values may be influenced by *interventions*. Each variable X in the domain depends in some deterministic manner on a subset $S(X)$ of all the other variables; here 'deterministic' means that the value of X is completely determined by the values of the variables in $S(X)$. Now if an intervention takes place that sets the value of a variable X to some value, then some of the variables in $S(X)$ become independent of X while the rest of the variables in $S(X)$ keep their dependence. If we equate an *action* with a *set of interventions*, this induces a semantics for the concept of 'action'. This resulting semantics, which will be discussed in more detail below, has sometimes been called *the sufficient cause principle* [34, 33]. An action that sets the value of a variable X is then called a 'sufficient cause' for X.

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

Causal Theories The basic entities of first-order causal theories are *fluents* (properties of the world that persist), *events* and (nonnegative integer) *time-points*. The main predicates used are *Ho*, *Do* and *Ab*.

- $Ho(f,t)$ denotes that fluent f holds at time t; $Ho(e,t)$ denotes that event e takes place at time t.
- $Do(x,b,t)$ denotes that 'an intervention takes place that sets fluent or event x to value b' at time t.
- $Ab_1(x,t), Ab_2(x,t)$ both denote that something abnormal is the case concerning fluent or event x at time t. In our use of Ab_1 and Ab_2, an 'abnormality' can *always* be seen as something *unexpected* or *surprising*.

Causal Models We determine the 'causal' models of causal theories in several steps. We start with a theory consisting of two sets of axioms: CONS and EQ. CONS contains all axioms specific to the domain under consideration; EQ contains general axioms involving persistence and the effects of interventions. Then we:

1. circumscribe (minimize) the *Do*-predicate in CONS, keeping *Ho*, Ab_1 and Ab_2 fixed.
2. add the axioms in EQ to the theory resulting from (1).
3. minimize the *Ab*-predicates in the theory resulting from (2).

The models resulting from step (3) are the *preferred causal models* of the causal theory defined by CONS and EQ. Note that we use nonmonotonic inference in steps (1) and (3).

Essential Features of our Theory Our theory is characterized by two features:

1. It follows the *alternative research goal* stated in Section 8.4 of the previous chapter, page 197. The distinction introduced there is reflected in our use of nonmonotonicity: in step (1) above, nonmonotonicity serves to enable a concise representation of causal theories. In step (3), nonmonotonicity serves to determine the 'most expected', or 'least surprising' models of a fully specified causal theory.
2. It follows *Pearl's theory of causation*. This is embodied in the semantics and the use of *Do*. The semantics of *Do* is derived from Pearl's 'sufficient cause principle'. The use of *Do* involves one of the central ideas in Pearl's theory: *observations and interventions should be modeled differently*. Observations are modeled using *Ho*, interventions are modeled using *Do*.

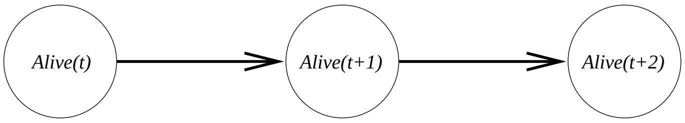

Figure 9.1: A Very Simple Causal Graph.

$Alive(t+1)$ and $Alive(t+2)$ has not disappeared: we can still infer that the turkey will *not* be alive at time $t+2$.

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

Causal Graphs and Structural Equations

We can depict the dependencies between the different variables of a given causal theory as a (possibly cyclic) *causal graph*[1] [34, 33]. Figure 9.1 shows the causal graph for the example above. The semantics of actions can be characterized in terms of these causal graphs: if an action takes place that sets the variable X to some value, then that variable becomes independent of all its non-descendants in the graph. Hence the intervention changes the theory into a new, updated one that is characterized by the causal graph in which all *incoming* arcs into the node corresponding to X have been removed while all other arcs remain. The graph does not give any information about the exact functional relationships between the variables in the domain, but only about which variables become independent in the presence of an intervention; we can now either regard causal theories as causal graphs together with a description of the functional relations between parents and children in the graph, or, as has been done by Pearl in his more recent papers [119, 120] we can avoid using graphs by writing the functional relations in the form of *structural equations*: these are equations with a directionality attached to them. A set of structural equations for our example would look as follows: $\{Alive(1) = Alive(0)\ ;\ Alive(2) = Alive(1)\ ; \ldots\}$. Our action semantics, i.e. the sufficient cause principle, can now be rephrased like this: if an intervention takes place that sets variable X to value x, then the structural equation $X = \ldots$ is replaced by the new equation $X = x$. In our example, if just one intervention takes place, and this intervention sets $Alive(2)$ to FALSE, then the equation $Alive(2) = Alive(1)$ gets replaced by $Alive(2) = \text{FALSE}$, while all other structural equations remain unchanged.

In Pearl's terminology, each structural equation can be seen as a *micro theory* [34]. It describes a single 'mechanism' whose input and output are given by, respectively, the right-hand and the left-hand side of the equation. All these mechanisms are coupled to form a complete theory. But if an intervention takes place, some of the mechanisms get *decoupled* from the rest and the output they normally provide is replaced by the result of the intervention. In other words, in Pearl's theory interventions are defined as *surgeries* on the set of mechanisms: an intervention keeps some of the mechanisms

[1] Historically, causal graphs have arisen as an extension of the Bayesian networks we saw in Chapter 6.

intact, while 'turning off' others, thereby changing the behaviour of the system of mechanisms as a whole.

Why we need to extend Pearl's theories

In common-sense reasoning about action, we are typically interested in domains where the fact whether or not an intervention takes place may depend on the values of some of the variables in the domain; this is impossible to model in Pearl's [119, 120] basic theory. Pearl's theory also lacks the representational power needed to express *global* domain constraints. For example, we may have $Alive(t) \equiv \neg Dead(t)$ irrespective of any action that influences the value of $Alive$; this formula should always hold in our domains and should not be replaced when $Alive$ is set to FALSE. It is therefore not a structural equation. All constraints between variables in Pearl's basic theory must be given in the form of structural equations, so it is impossible to model such a global constraint. In our work we extend Pearlian causal theories in a way that allows us to represent both kinds of domain knowledge mentioned here. We do this by simply adding to the set of structural equations a set of *global constraints*, consisting of formulas that should hold in all models of our theories irrespective of what interventions take place. These global constraints are also allowed to mention interventions. To be able to express those, we introduce for each propositional variable X in our domain two additional propositional variables $Do(X, \text{TRUE})$ and $Do(X, \text{FALSE})$, representing interventions setting X either to TRUE or to FALSE.

9.3 Propositional Causal Theories

Our propositional causal theories are a straightforward extension of Pearl's causal theories. In an appendix to this chapter (page 235) we show exactly how they are related to Pearl's causal theories as defined in [119, 120].

Preliminaries

Let $\mathbf{B} = \{\text{TRUE}, \text{FALSE}\}$. A *truth value* is an element of $\mathbf{B}$. Let $\mathbf{X}$ be a set of propositional variables, i.e. variables taking on truth values. We assume a standard propositional language for $\mathbf{X}$ containing the additional symbols TRUE and FALSE which are abbreviations of $X \vee \neg X$ and $X \wedge \neg X$ respectively; here X is any variable in $\mathbf{X}$. A *valuation* or *interpretation* of $\mathbf{X}$ is an assignment of truth values to the variables in $\mathbf{X}$. A valuation $\mathcal{M}$ is a *model* for a set of propositional formulas Γ (written as $\mathcal{M} \models \Gamma$) if Γ is true in $\mathcal{M}$; here truth of propositional formulas with respect to interpretations is defined as usual. Hence, from now on,'$\models$' will stand for standard propositional entailment, later it may also stand for standard first-order entailment.

Causal Theories and their Models

Here is our initial definition of causal theories:

Definition 9.1 *A propositional causal theory T is a 4-tuple* $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS} \rangle$ *where*

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

$$X_i \equiv \Phi_i(X_1, \ldots, X_n, U_1, \ldots, U_m) \tag{9.1}$$

where Φ_i is a formula involving zero or more of the variables $X_1, \ldots, X_n$, $U_1, \ldots, U_m$. For each $X_i \in \mathbf{V}$, EQ *contains at most one structural equation with X_i on the left-hand side.*

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

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

Notice that expressions of the form '$Do(X_i, b)$' simply stand for propositional variables here. The set $\mathbf{U}$ can be used to model external influences we are ignorant about or that we consider unlikely; we will only need it in Section 9.5.4.

Example 9.2 Let $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS}\rangle$ be a causal theory with $\mathbf{V} = \{Alive(0), Alive(1)\}$, $\mathbf{U} = \emptyset$, EQ $= \{Alive(1) \equiv Alive(0)\}$, CONS $= \{Alive(0), \neg Do(Alive(1), \text{FALSE}), \neg Do(Alive(1), \text{TRUE})\}$. The equation in EQ expresses that the value of *Alive* persists (from time 0 to 1) if there are no interventions; the formulas in CONS ensure that $Alive(0)$ should hold in all models of T and that there is no intervention at time 1.

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

Definition 9.3 *A causal model for a propositional causal theory $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS}\rangle$ is a valuation $\mathcal{M}$ for the variables in $\mathbf{V} \cup \mathbf{U} \cup \mathcal{A}(\mathbf{V})$ such that*

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

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

If $\mathcal{M}$ is a causal model for T and it is clear from the context that T is a causal theory rather than a set of propositional formulas, then we simply write '$\mathcal{M}$ is a model for T'.

Example 9.2, continued Let T be as in Example 9.2 above. Condition (1) of Definition 9.3 ensures that all causal models $\mathcal{M}$ for T have $\mathcal{M} \models Alive(0)$ and that for all $b \in \mathbf{B}$, $\mathcal{M} \models \neg Do(Alive(1), b)$. By condition (2), $\text{EQ}' = \text{EQ}$ for all models and so we also have $Alive(1)$ in all models for T: *Alive* persists. Now consider a theory T' identical to T except that CONS is now equal to $\text{CONS} = \{Alive(0), Do(Alive(1), \text{FALSE})\}$. By condition 1 of Definition 9.3 we now have $\mathcal{M} \models Alive(0) \wedge Do(Alive(1), \text{FALSE})$ for all causal models $\mathcal{M}$. By condition 2 of that definition, we have $\text{EQ}' = \{Alive(1) \equiv \text{FALSE}\}$ for all these models and thus all models for T' must have $\neg Alive(1)$.

9.4 Causal Theories Involving Persistence

Throughout this and the next chapter we consider domains in which two basic kinds of entities exist: *fluents*, which are properties of the world that, if no interventions take place, do not change over time, and *events*, which will stand for the 'triggers' of interventions. We use the term 'event' rather than 'action' here since our 'events' do not necessarily have to be performed by some agent – the difference however is not crucial in what follows. Causal theories for our domains of interest will thus be defined with respect to a set of fluents $\mathbf{F}$ and a set of events $\mathbf{E}$; we assume these sets to be finite. Specifically, from now on we assume that for any causal theory for the sets $\mathbf{E}$ and $\mathbf{F}$, the set of observables $\mathbf{V}$ can be partitioned into two subsets: $\mathbf{V_E}$, the set of *event – time pairs*, and $\mathbf{V_F}$, the set of *fluent – time pairs*. In this section and the next, we are only interested in the behaviour of our domains directly before and directly after some actions happen. We therefore restrict ourselves to *2-state causal theories* which represent domains using only two points in time: the initial point in time $t = 0$ and the final point in time $t = 1$. Given the set $\mathbf{F} = \{F_1, \ldots, F_n\}$, $\mathbf{V_F}$ can now be written as $\mathbf{V_F} = \{F_1(0), \ldots, F_n(0), F_1(1), \ldots, F_n(1)\}$. Here $F_i(t)$ denotes fluent i at time t. We assume that all actions we are interested in happen at the same time, namely directly after the initial point in time. Hence we do not have to index actions by time points and can simply let $\mathbf{V_E} = \mathbf{E}$.

We define a fluent literal to be an expression of the form F or $\neg F$ where F is some fluent in $\mathbf{F}$. The *initial state* of a causal model $\mathcal{M}$ for a causal theory for sets $\mathbf{E}$ and $\mathbf{F}$ is the set $\{F \mid \mathcal{M} \models F(0)\} \cup \{\neg F \mid \mathcal{M} \models \neg F(0)\}$. The *final state* of a causal model $\mathcal{M}$ is the set $\{F \mid \mathcal{M} \models F(1)\} \cup \{\neg F \mid \mathcal{M} \models \neg F(1)\}$.

Example 9.4 We now extend our turkey domain with a new fluent *Dark*. $Dark(t)$ will denote that it is dark at time t; shooting at the turkey will have its intended effect only if it is not dark (so that one can aim properly). For this, let T be a causal theory such that $\mathbf{V_E} = \{Shoot\}$, $\mathbf{V_F} = \{Alive(0), Dark(0), Alive(1), Dark(1)\}$; $\mathbf{U} = \emptyset$. EQ contains two equations: $\{Alive(1) \equiv Alive(0)\ ; Dark(1) \equiv Dark(0)\}$. CONS contains the single axiom

$$[\ \neg Dark(0) \wedge Shoot\ ] \supset Do(Alive(1), \text{FALSE}) \tag{9.2}$$

There are potentially four different initial and final states for T, corresponding to the four possible interpretations of *Alive* and *Dark*. Let us look at the set of valuations $\mathbf{M}$ in which no *Shoot*-event and no interventions take place; i.e. we have $\neg Shoot$

and $\neg Do(X, b)$ for any $X \in \mathbf{V}, b \in \mathbf{B}$. Clearly, for any member $\mathcal{M}$ of $\mathbf{M}$ we have $\mathcal{M} \models$ CONS; hence condition (1) of Definition 9.3 is satisfied. Since, according to condition (2) of that definition, the updated set of equations EQ$'$ is equal to EQ, the valuations in $\mathbf{M}$ which also satisfy this condition are clearly exactly those in which the values of both *Alive* and *Dark* persist. Thus $\mathbf{M}$ contains four models, each of which has the same initial and final state. Hence in models in which no interventions take place, we have persistence, which is in accord with intuition.

9.4.1 We need more...

Let us now look at the set of all causal models $\mathbf{M}'$ for T in which the *Shoot*-event *does* take place and in which it is not dark in the initial situation. Hence for all $\mathcal{M} \in \mathbf{M}'$ it holds $\mathcal{M} \models \neg Dark(0) \wedge Shoot$. By axiom (9.2) all $\mathcal{M} \in \mathbf{M}'$ have $\mathcal{M} \models Do(Alive(1), \text{FALSE})$. The set EQ$'$ for these models must therefore contain $Alive(1) \equiv$ FALSE so each $\mathcal{M} \in \mathbf{M}'$ also has $\mathcal{M} \models \neg Alive(1)$. But notice that there is also a model $\mathcal{M}' \in \mathbf{M}'$ with

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

since both conditions (1) and (2) of Definition 9.3 are satisfied for $\mathcal{M}'$.

This is not what we would intuitively expect! Intuitively, we would reason as follows: (1) there are no interventions except those triggered by the *Shoot*-event; (2) *Shoot* does not affect *Dark*, so the value of *Dark* should remain unchanged after the *Shoot*-action.

What we have yet forgotten to model are the implicit assumptions behind (1) and (2), namely that (a) *there are no external interventions* and that (b) *actions affect no more things in the world than those we explicitly state they affect.* Notice that, unlike assumption (a) and the assumption of persistence, assumption (b) is *not* an assumption about the *physics* of our reasoning domains, i.e. it is not an assumption about how the world works. Rather it is an assumption about *what we really mean when we specify our domain knowledge in a certain way.* It allows us to specify domains in a compact way and as such corresponds to the second part of the distinction indicated in Chapter 8, Section 8.4 (page 198). In contrast, the persistence assumption (expressed by the axioms in EQ), the effect laws (expressed by the axioms in CONS) and the sufficient cause principle (expressed by the semantics of causal theories, i.e. Definition 9.3) are directly about the 'physics' of our domains: they describe how the domains under consideration really work, and as such belong to the first part of the distinction made on page 198.

In order to find a way to deal with assumption (b), we need to ask ourselves when exactly an 'intervention' takes place, or equivalently: under what conditions is a variable *set* to a value? Variables are always set to a value in a specific *context*: for example, $Alive(1)$ is set to FALSE only in a context in which *Shoot* and $\neg Dark(0)$ is true. Now suppose we are in one specific context; let us call it C. If there is a model $\mathcal{M}_1$ with this context and with $\mathcal{M}_1 \models Do(X_i, b)$ while there is also a model $\mathcal{M}_2$ with the same context C and with $\mathcal{M}_2 \models \neg Do(X_i, b)$, then we can be sure that nothing in our axioms states that in context C the fluent X_i is set to value b. By assumptions

(a) and (b) above, we should now prefer model $\mathcal{M}_2$. Apparently, we should partition our models into classes sharing the same context, and then within each such class pick the model which has $\neg Do(X_i, b)$ for as many X_i and b as possible.

Now what does it mean that two models 'have the same context'? Clearly, the context should contain at least *every fact in the world that can possibly influence whether or not an intervention takes place.* Since we assume (assumption (a)) that there are no external events, we may safely say that two models in which exactly the same events take place and the same fluents hold at the same time share the same context. Any two such models interpret all the variables in $\mathbf{V}$ and $\mathbf{U}$ the same; hence we should partition our models into equivalence classes where each class corresponds to one particular interpretation of the variables in $\mathbf{V} \cup \mathbf{U}$ and contains all models with that particular interpretation. For each equivalence class we should then pick the models which have a minimal interpretation (see below) of Do. But notice that we must do all this *before* we replace the structural equation set EQ by EQ$'$ (step 2 of Definition 9.3), since this replacement will only work if the Do-propositions already have the right interpretations in each model.

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

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

Definition 9.5 *We call $\mathcal{M}$ a minimal model for Γ of the variables $\mathbf{Y}$ within context $\mathcal{C}\rfloor_{\mathbf{Z}}$ iff:*

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

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

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

We call causal theories in which Do still has to be minimized *partially specified* (since it is only specified which fluents are affected by events and it is not specified which fluents are not). In both this chapter and the next, we will assume that all our causal theories are actually partially specified ones. This brings us to the final definition of 2-point causal theories and their models. Definition 9.6 defines 2-point causal theories as a special case of the general propositional causal theories as defined in Definition 9.1: each 2-point causal theory is also a propositional causal theory.

Definition 9.6 *A 2-point causal theory for the set of fluents $\mathbf{F}$ and the set of events $\mathbf{E}$ is a tuple $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS} \rangle$ such that*

- $\mathbf{V} = \mathbf{V_F} \cup \mathbf{V_E}$ *with* $\mathbf{V_F} = \{F_i(0), F_i(1) \mid F_i \in \mathbf{F}\}$ *and* $\mathbf{V_E} = \mathbf{E}$.
- $\mathbf{U} = \{U_1, \dots, U_m\}$ *is a set of unobserved propositional variables*
- $\text{EQ} = \{F_i(1) \equiv F_i(0) \mid F_i \in \mathbf{F}\}$
- CONS, *the set of* constraints, *is a finite set of propositional formulas over variables* $\mathbf{V} \cup \mathbf{U} \cup \mathcal{A}(\mathbf{V})$. *Here* $\mathcal{A}(\mathbf{V})$ *is defined as the set of propositional variables*

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

Notice that the following definition only differs from Definition 9.3 in its first condition.

Definition 9.7 *A causal model for a 2-point causal theory* $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS} \rangle$ *is a valuation* $\mathcal{M}$ *for the variables in* $\mathbf{V} \cup \mathbf{U} \cup \mathcal{A}(\mathbf{V})$ *such that*

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

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

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

We have now reached our first 'milestone'; that is, a theory that can be successfully applied.

9.5 The Power of Two-Point Causal Theories

2-point causal theories allow us to handle many interesting problem domains, specifically involving *ramifications* and actions with *disjunctive* or *non-deterministic* effects. In this section we will give several examples to illustrate this fact. We will briefly indicate how other approaches handle these examples and how our approach avoids some of their difficulties. Problem domains which involve more than two points in time have to wait until the next section, where our theory is extended to deal with them.

The key to all the examples presented in this section is that we are allowed the use of *Do* in domain constraints and effect axioms. Using *Do*, we are able to express the difference between an observation (e.g. $\neg Alive(1)$) and an intervention (e.g. $Do(Alive(1), \text{FALSE})$).

9.5.1 Ramifications

The *ramification problem* concerns the problem of concisely and correctly representing indirect effects of actions. It was introduced in Chapter 8, Section 8.2.4. The key to our handling of ramifications are axioms of the form

$$Do(F_1, b) \supset Do(F_2, b') \tag{9.3}$$

Such axioms can be interpreted as saying that *any event in the domain that sets the value of F_1 to b also sets the value of F_2 to b'*. Forms of (9.3) will be used in all the examples we are about to present.

Before turning to more complicated issues, we start with two standard examples involving ramifications: the extended 'walking turkey' problem and the 'two switches problem'.

Example 9.8 [The Walking Turkey] We introduced the 'walking turkey' in Chapter 8, Example 8.5, page 193. Recall that it was about a turkey that may be *Alive* or not and that may be *Walking* or not. In the basic version of the domain, we want to express that if a turkey stops being alive (for example, because it is shot at), then it should also stop walking: $\neg$ *Walking* is a *ramification* of $\neg$*Alive*. Later, a more complicated version of the domain was proposed by McCain and Turner [103]. They noted that if the turkey is not alive and somebody tries to make it walking (for example, by performing the action *Entice*), it will not suddenly become alive! *Alive* is *not* a ramification of the effect of *Entice*; rather, *Alive* is a *qualification* of *Entice* (by 'A is a qualification of B' we mean that the action B will have its usual effect only in situations in which A is the case). Clearly, a sentence of the form $\neg \mathit{Alive}(t) \supset \neg \mathit{Walking}(t)$ is not enough to express this knowledge, since it treats $\mathit{Alive}(t)$ and $\neg \mathit{Walking}(t)$ in an equivalent manner. Therefore it cannot represent the different 'dynamics' of *Alive* and $\neg$ *Walking*.

Let us formalize our domain. Consider the 2-point causal theory T_{WT} for the event and fluent sets $\mathbf{E} = \{\mathit{Shoot}, \mathit{Entice}\}$ and $\mathbf{F} = \{\mathit{Alive}, \mathit{Walking}\}$. $\mathbf{U} = \emptyset$. $\textsc{eq} = \{\mathit{Alive}(1) \equiv \mathit{Alive}(0)\ ;\ \mathit{Walking}(1) \equiv \mathit{Walking}(0)\}$. $\textsc{cons}$ consists of the following four axioms:

$$\mathit{Shoot} \supset \mathit{Do}(\mathit{Alive}(1), \textsc{false}) \tag{9.4}$$

$$\mathit{Entice} \supset \mathit{Do}(\mathit{Walking}(1), \textsc{true}) \tag{9.5}$$

$$\neg \mathit{Alive}(t) \supset \neg \mathit{Walking}(t) \tag{9.6}$$

$$\mathit{Do}(\mathit{Alive}(t), \textsc{false}) \supset \mathit{Do}(\mathit{Walking}(t), \textsc{false}) \tag{9.7}$$

Here axioms of the form $\phi(t)$ should be read as $\phi(0) \wedge \phi(1)$. At first sight, axiom (9.7) seems a consequence of (9.6). However, axiom (9.6) denotes a *static* domain constraint ('there can be no state in which a turkey is both walking and not alive') while axiom (9.7) denotes its corresponding 'dynamics'. Axiom (9.7) will allow us to infer that interventions that set *Alive* to false also set *Walking* to false. Note that it is an example of an axiom of form (9.3). By the minimization of *Do* in Definition 9.7, we will *not* be able to conclude that an intervention which sets *Walking* to true also sets *Alive* to true. We now show this in detail.

Let us denote by $\mathbf{M}$ the set of causal models for T_{WT} in which *Shoot* takes place while *Entice* does not; by axioms (9.4) and (9.7) we have $\mathit{Do}(\mathit{Alive}(1), \textsc{false})$ and $\mathit{Do}(\mathit{Walking}(1), \textsc{false})$ in all such models. By condition (1) of Definition 9.7, we also have $\neg \mathit{Do}(X_i, b)$ for all other (X_i, b). Hence, the updated set of structural equations $\textsc{eq}'$ for the models in $\mathbf{M}$ (Definition 9.6) becomes:

$$\mathit{Alive}(1) \equiv \textsc{false} \quad ; \quad \mathit{Walking}(1) \equiv \textsc{false}$$

Hence all models in $\mathbf{M}$ have as their final state $\{\neg Alive, \neg Walking\}$. We also see (by checking whether all the axioms in CONS and EQ$'$ hold) that $\mathbf{M}$ contains models for three different initial states: $\{Alive, Walking\}$, $\{Alive, \neg Walking\}$, $\{\neg Alive, \neg Walking\}$. So the case in which the domain constraints should entail a ramification of the *Shoot*-event works fine. Now for the models in which no *Shoot* but an *Entice* event takes place. All such models have $Do(Walking(1), \text{TRUE})$. By the minimization of *Do* in condition (1) in Definition 9.7 they also have $\neg Do(Alive(1), \text{TRUE})$. Hence the set EQ$'$ in condition (2) of that definition becomes

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

By axiom (9.6) all models $\mathcal{M}$ must then also have $\mathcal{M} \models Alive(1)$, and, since $\mathcal{M} \models$ EQ$'$ also $\mathcal{M} \models Alive(0)$. This means that there are no causal models for T_{WT} with both $\neg Alive(0)$ and *Entice*: *Entice* has *Alive* as an *implicit qualification*.

Our next example is due to Lifschitz' [97]. It is closely related to Lin's [99] 'suitcase problem' and to Ginsberg's 'stuffy room problem' [135].

Example 9.9 [The Suitcase & Switches Problem] Imagine a light that is connected to two switches; the light is only on if both of the switches are in the *on*-position: any event that puts a switch into the *on*-position in a context in which the other switch is on, will have as a ramification that the light goes on. However, if the event takes place in a context in which the other switch is not on, the light will not be affected. The main importance of this example is that some previous approaches cannot express the domain properly: they cannot rule out models in which, as a result of turning on one switch in a context in which the other one is on already, the other switch may jump into the *off*-position while the light remains out; see Lin [99] for details. We formalize the domain using event and fluent sets $\mathbf{E} = \emptyset$ and $\mathbf{F} = \{Swi_1, Swi_2, Light\}$. Swi_i denotes that switch i is in the *on*-position. Let T_{SW} be a causal theory for $\mathbf{E}$ and $\mathbf{F}$ with $\mathbf{U} = \emptyset$, EQ as usual and CONS as follows:

$$[\ Swi_1(t) \wedge Do(Swi_2(t), \text{TRUE})\ ] \supset Do(Light(t), \text{TRUE}) \tag{9.8}$$
$$[\ Swi_2(t) \wedge Do(Swi_1(t), \text{TRUE})\ ] \supset Do(Light(t), \text{TRUE}) \tag{9.9}$$
$$Do(Swi_1(t), \text{FALSE}) \supset Do(Light(t), \text{FALSE}) \tag{9.10}$$
$$Do(Swi_2(t), \text{FALSE}) \supset Do(Light(t), \text{FALSE}) \tag{9.11}$$
$$(Swi_1(t) \wedge Swi_2(t)) \equiv Light(t) \tag{9.12}$$

Here axiom (9.12) describes a domain constraint that must hold in all models of the domain; the other axioms describe the 'dynamics' of the domain. Let us see whether we get the expected models for T_{SW}: suppose first we turn on the first switch while the second one is off; i.e. we add the axiom $\neg Swi_2(0) \wedge Do(Swi_1(1), \text{TRUE})$ to CONS. By the minimization of *Do*, we get $\neg Do(Swi_2(1), b)$ for all b in all models. Hence EQ$'$ contains $Swi_2(1) \equiv Swi_2(0)$, so all models for T_{SW} must have $\neg Swi_2(1)$. But in all models with $\neg Swi_2(1)$, we already have (again by the minimization of *Do*) $\neg Do(Light(1), \text{TRUE})$ and $\neg Light(1)$. It follows that all models will have $\neg Light$ in their final state; it is easy to see that such models indeed exist. One can show in a similar manner that one obtains the intended models in all other possible scenarios.

9.5.2 A First Glimpse at Other Approaches

We have seen in the introduction that, in Pearl's terminology, the proposition $Do(X_i, b)$ can be read as 'there is a sufficient cause for X_i to take on the value b'. Indeed, usage of 'Do' in axioms often corresponds to the colloquial use of the word 'causes'. It turns out that if the word Do is replaced by the word `causes`, then our axioms start to look very similar to those used in various other approaches. We leave detailed comparisons for Chapter 10 and give only a suggestive example for the time being. This will serve to make clear that existing approaches solve the ramification problem in a quite similar, yet subtly different manner. The difference causes problems for domains involving 'causal cycles', which we are about to present.

In Thielscher's work, the equivalent of axiom (9.7) looks as follows ([151], page 330):

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

In Lin's work [99, 100], we would get:

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

In McCain & Turner's approach [103, 104, 105], it would be formalized as:

$$\neg Alive \Rightarrow \neg Walking \tag{9.13}$$

where '$\Rightarrow$' is to be read as 'if $\neg Alive$, then the fact that $\neg Walking$ is caused' [104]. McCain & Turner's semantics interprets (9.13) as something like

$$\neg Alive(t) \supset Do(Walking(t), \text{FALSE}) \tag{9.14}$$

In the turkey example, replacing (9.7) by (9.14) does not make any difference: it is easy to show that one obtains exactly the same models in both cases. However, as we will see below, in general, $(X \equiv b_1) \supset Do(Y, b_2)$ is *not* equivalent to $Do(X, b_1) \supset Do(Y, b_2)$.

9.5.3 Causal Cycles

We have just seen that $Do(X, \text{TRUE}) \supset Do(Y, \text{TRUE})$ may often be read as 'X causes Y'. As pointed out by Sandewall and Gustafsson & Doherty [66, 134], we do not always want to reason in the 'causal direction' that is implied by the above; for example, consider the switches domain (Example 9.9) again: if we know that the light has been put off, we may want to conclude that (at least) one of the two switches has been put off too. We model this as follows:

Example 9.10 Let $T_{\text{SW},2}$ be as T_{SW} but with the following axioms added to CONS:

$$\begin{aligned}
Do(Light(t), \text{FALSE}) \quad &\supset \quad [\ Swi_1(t) \supset Do(Swi_2(t), \text{FALSE})\ ] && (9.15)\\
Do(Light(t), \text{FALSE}) \quad &\supset \quad [\ Swi_2(t) \supset Do(Swi_1(t), \text{FALSE})\ ] && (9.16)\\
Do(Light(t), \text{TRUE}) \quad &\supset \quad Do(Swi_1(t), \text{TRUE}) && (9.17)\\
Do(Light(t), \text{TRUE}) \quad &\supset \quad Do(Swi_2(t), \text{TRUE}) && (9.18)
\end{aligned}$$

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

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

$$Do(Light(1), \text{TRUE}) \wedge \neg Swi_1(0) \wedge \neg Swi_2(0) \tag{9.19}$$

Axioms (9.17) and (9.18) ensure that we have $Do(F(1), \text{TRUE})$ for $F \in \{Swi_1, Swi_2, Light\}$ in all models; so EQ' becomes $F(1) \equiv \text{TRUE}$ for all these F in all models, and all models get final state $\{Swi_1, Swi_2, Light\}$; it is easy to see that models with such a final state indeed exist, so we get exactly the result we wanted. What happens if we turn off the light will be discussed in Example 9.12.

The example above leads to a problem for some of the earlier causal approaches (i.e. those of Gustafsson & Doherty and McCain & Turner; see [66] for the details). The reason is that in these approaches, constructions similar to $X \supset Do(Y, \text{TRUE})$ are used at places where one should use $Do(X, \text{TRUE}) \supset Do(Y, \text{TRUE})$; see the remark at the end of Section 9.5.2. Had we chosen the first possibility, axioms (9.8) and (9.9) would have been replaced by the single axiom

$$[\ Swi_1(t) \wedge Swi_2(t)\ ] \supset Do(Light(t), \text{TRUE}) \tag{9.20}$$

while axioms (9.17) and (9.18) would have become:

$$Light(t) \supset [\ Do(Swi_1(t), \text{TRUE}) \wedge Do(Swi_2(t), \text{TRUE})] \tag{9.21}$$

and axioms (9.15) and (9.16) would have become:

$$\neg Light(t) \supset \\ [\ [Swi_1(t) \supset Do(Swi_2(t), \text{FALSE})] \wedge [Swi_2(t) \supset Do(Swi_1(t), \text{FALSE})]\ ] \tag{9.22}$$

By axioms (9.20) and (9.21) we would then get that each model with final state $S = \{Light\ ;\ Swi_1\ ;\ Swi_2\}$ also has $Do(F(1), \text{TRUE})$ for F equal to any of the three fluents. Since (a) there is nothing in our axioms contradicting models with final state S and (b) the minimization of *Do* cannot rule out particular interpretations of $\mathbf{V} \cup \mathbf{U}$, such models exist and are not ruled out by condition (1) of Definition 9.7. Therefore, for such models the set EQ' in condition (2) of Definition 9.7 would become $F \equiv \text{TRUE}$ for all three F and all three persistence relations would be broken: there is an 'automatic' intervention that sets their value to TRUE. It follows that no matter what the initial state is, there is always a model with the final state above: there is a cycle involved in that the fact that the light is on implies an intervention that puts both switches on, while the fact that the switches are on implies an intervention that puts the light on. We have already seen that in our own formalization (the one using axioms (9.8)-(9.12) and (9.15)-(9.18)) this cannot happen.

Thielscher ([151], page 329) solves the problem in a way similar to ours, but since his equivalent of *Do* is called `causes`, the corresponding rules look somewhat strange: he obtains rules like '*Light* `causes` Swi_1', while, intuitively, turning on the light does not 'cause' turning on the switch. Related observations have led some authors to claim that there is more to ramification than mere causal relationships – as stated in Gustafsson & Doherty, 'physical causality is simply one of several reasons one might set up dependencies between fluents' [66].

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

9.5.4 Disjunctive Effects and Nondeterminism

Example 9.10 raises the interesting question of how to formalize actions whose effect is a *disjunction* of fluents. Recently, some authors have, either explicitly or implicitly, begun to take up this question [105, 151]. A special case of this question concerns the more well-known issue of representing actions with non-deterministic effects [11, 105, 135]. If an event occurs that sets the value of a proposition X to TRUE, where $X \equiv Y \vee Z$, it is not immediately clear what should happen: should we exempt both Y and Z from persistence, or should we keep as much persistence as is logically consistent? In Example 9.10 we implicitly chose the latter option: setting *Light* to FALSE in an initial situation with both switches on will always keep one switch in the *on*-position: as will be shown below, there will be no final states with both $\neg Swi_1$ and $\neg Swi_2$. But is this correct, i.e. is it what we intuitively expect? A small poll conducted by the present author reveals that people have differing intuitions about this: some think that the position of both switches should be exempted from persistence, while others prefer the 'inert models' where there are as few changes as possible. We think that this clash of intuitions is not so surprising in light of Pearl's theory: if we intervene to set the value of a variable X that can be regarded as a disjunction of variables Y and Z, then we simply have an *incomplete specification* of our problem: it is not clear what should happen to the structural equations for Y and Z – should they both be removed from the set EQ, or should we remove as few equations as possible? In our view, this may change from domain to domain and should therefore be reflected in the domain axioms. We will therefore *not* attempt to extend the semantics of *Do* to constructions of the form $Do(Y \vee Z, b)$, but rather always make explicit what should happen to the structural equations of Y and Z if an intervention takes place. We can do so in our causal theories by using the *assumption symbols* in $\mathbf{U}$ (see Definition 9.1).

We illustrate their use by considering a domain[2] which clearly asks for eliminating all structural equations involved in the disjunction.

Example 9.11 We drop a pencil on a table with a piece of paper on it. As a result, the pencil may either lie fully on the paper, or touch both part of the paper and part of the table's surface, or touch only the table's surface. We write $\mathit{Touches\_table}(t)$ $(\mathit{Touches\_paper}(t))$ iff the pencil touches the surface of the table (the paper) at time t. We can model this using a causal theory with $\mathbf{E} = \{\mathit{Drop}\}$, $\mathbf{F} = \{\mathit{Touches\_table}, \mathit{Touches\_paper}\}$, $\mathbf{U} = \{U_1, U_2\}$, $\mathbf{V} = \mathbf{V_E} \cup \mathbf{V_F}$, EQ as in Definition 9.6 and CONS as follows:

$$\neg\mathit{Touches\_table}(0) \wedge \neg\mathit{Touches\_paper}(0) \tag{9.23}$$

$$\begin{aligned} \mathit{Drop} \supset [\; & (\; \mathit{Do}(\mathit{Touches\_table}(1), \text{TRUE}) \wedge U_1 \;) \vee \\ & (\; \mathit{Do}(\mathit{Touches\_paper}(1), \text{TRUE}) \wedge U_2 \;) \;] \end{aligned} \tag{9.24}$$

In this case, among the models with *Drop*, there will be three subclasses, corresponding to the three extensions of (U_1, U_2) that are consistent with (9.24). By the minimization of *Do* in Definition 9.7 and by axiom (9.24), we get the following interpretation for each class of models:

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

(9.25)

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

Example 9.12 Let us now consider a case in which we want 'as much persistence as possible'. Let us say there is an intervention which sets the value of variable X where $X \equiv Y \vee Z$. If we let $X \equiv \neg\mathit{Light}$, $Y \equiv \neg\mathit{Swi}_1$ and $Z \equiv \neg\mathit{Swi}_2$, then we can see that we have already implicitly treated this case in the extended switches domain, Example[3] 9.10. Consider a scenario in the context of that example with an intervention setting *Light* to FALSE, i.e. we add the following axiom to the set CONS of $T_{\text{SW},2}$:

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

We see that for the subclass of models with $\mathit{Swi}_1(1)$ and $\neg\mathit{Swi}_2(1)$ we have $\mathit{Do}(\mathit{Swi}_2(1),$ FALSE$)$ (by axiom (9.15)) and $\neg\mathit{Do}(\mathit{Swi}_1(1), \text{FALSE})$ (by the minimization of *Do*). For these models EQ$'$ contains $\mathit{Swi}_2(1) \equiv \text{FALSE}$ and $\mathit{Swi}_1(1) \equiv \mathit{Swi}_1(0)$. It is now easy

[2]Example 9.11 is an adaptation of a scenario that, according to M. Shanahan [139], is due to R. Reiter.

[3]We would like to stress that we do not say that the switches domain *should* be formalized so as to keep 'as much persistence as possible' – we formalized it this way just for illustrative purposes, and we feel that one may equally well decide to formalize it along the lines of Example 9.11.

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

Example 9.13 [Exclusive Non-Deterministic Effects] The classic example of an action with a non-deterministic effect is the tossing of a coin: if we toss a coin, it will come up either heads (*Heads*) or tails, but we do not know which. However, we do know that whether or not we have *Heads* after tossing is independent of the fact whether or not we had *Heads* before tossing, thus there is no persistence. In light of the previous examples, this can clearly be modeled by a causal theory containing the following CONS:

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

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

9.6 Handling many Time-points, Events & and Surprises

In this section we extend our 2-point causal theories to handle arbitrarily many points in time. This means we will have to move to a first-order language. In the subsections to come, we first (Section 9.6.1) show how first-order theories arise as a natural extension of 2-point theories. In Section 9.6.2 we introduce 'basic' first-order causal theories. We then instantiate these theories in more and more complicated ways: in Section 9.6.3, we start with simple instantiations for handling domains in which exactly one event happens at a time. Section 9.6.4 extends our theories to handle 'dependent fluents' which form yet another instance of the ramification problem. Then (Section 9.6.5) we move over to domains in which more than one event may happen at the same time and (Section 9.6.6) we show how this allows us to formalize causal chains of events. Finally, we extend our domains to handle complete surprises (Section 9.6.7).

9.6.1 From Propositional to First-Order Causal Theories

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

For any propositional causal theory $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS} \rangle$ defined according to Definition 9.1, let T' be the propositional theory over variables $\mathbf{U} \cup \mathbf{V} \cup \mathcal{A}(V)$ such that $T' = \text{CONS} \cup \text{EQ}_{\text{new}}$. Here EQ_{new} results from EQ by replacing any structural equation $X_i \equiv \Phi$ in EQ by the following three axioms:

$$[\ \neg Do(X_i, \text{TRUE}) \wedge \neg Do(X_i, \text{FALSE})\ ] \ \supset (X_i \equiv \Phi) \tag{9.26}$$

$$Do(X_i, \text{TRUE}) \supset X_i \tag{9.27}$$

$$Do(X_i, \text{FALSE}) \supset \neg X_i \tag{9.28}$$

Here the first axiom represents the original structural equation: if no interventions take place, then X_i should still be equivalent to Φ (and hence the structural equation applies). The second and third axioms represent the effect of actions. To give an example, if $\mathcal{M} \models Do(X_i, \text{TRUE})$ (and hence 'X_i is set to true in $\mathcal{M}$'), then the left-hand side of (9.26) does not hold in $\mathcal{M}$, and $\mathcal{M}$ automatically satisfies (9.26). On the other hand, the left hand side of (9.27) does hold in $\mathcal{M}$, hence $\mathcal{M}$ must satisfy $\mathcal{M} \models X_i$.

The following theorem shows that the notion of 'causal model' for theories $T = \langle \mathbf{V}, \mathbf{U}, \text{EQ}, \text{CONS} \rangle$ is simply equivalent to the notion of classical model for theories $T' = \text{CONS} \cup \text{EQ}_{\text{new}}$.

Theorem 9.14 *For any causal theory T and propositional theory T' obtained from T as described above, we have*

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

The proof (which is almost trivial) can be found in [64].

The theorem shows that propositional causal theories can be equivalently modeled as 'plain' propositional theories. But these can be extended in a straightforward way to first-order theories:

Quantifying over Structural Equations

Let us suppose for the moment that we want to use causal theories for persistence involving $n+1$ points in time. Theorem 9.14 shows that in this case, all axioms in EQ can equivalently be represented by a set EQ_{new} which contains:

$$\begin{array}{rcl}
\neg Do(F_i(1), \text{TRUE}) \wedge \neg Do(F_i(1), \text{FALSE}) & \supset & (F_i(1) \equiv F_i(0)) \\
\neg Do(F_i(2), \text{TRUE}) \wedge \neg Do(F_i(2), \text{FALSE}) & \supset & (F_i(2) \equiv F_i(1)) \\
 & \vdots & \\
\neg Do(F_i(n), \text{TRUE}) \wedge \neg Do(F_i(n), \text{FALSE}) & \supset & (F_i(n) \equiv F_i(n-1))
\end{array} \tag{9.29}$$

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

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

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

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

$$\forall t > 0 \ . \ [\neg Do(F_i(t), \text{TRUE}) \wedge \neg Do(F_i(t), \text{FALSE})] \supset (F_i(t) = F_i(t-1)) \tag{9.30}$$

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

$$\forall t \ . \ Do(F_i(t), \text{TRUE}) \supset (F_i(t) = \text{TRUE}) \tag{9.31}$$

$$\forall t \ . \ Do(F_i(t), \text{FALSE}) \supset (F_i(t) = \text{FALSE}) \tag{9.32}$$

Axioms of form (9.30) will be called *structural equation axioms*. Notice that they really contain an infinitude of structural equations. Axioms of form (9.32) will be called *intervention axioms*.

We will want to quantify not only over time points but also over fluents. This can easily be accomplished by introducing a new predicate Ho and writing $Ho(F_i, t)$ instead of $F_i(t)$, a common practice in common-sense temporal reasoning. This final extension immediately leads to the formal definitions of first-order theories that we will present in the next subsection.

9.6.2 First-Order Causal Theories

Preliminaries We use a many-sorted first-order language $\mathcal{L}$. Structures, interpretations, truth in a model and entailment are defined as usual; see Chapter 8, page 187. Specifically, from now on '$\models$' stands for first-order rather than propositional entailment. For a sort X indicated by the letter X, we write $|\mathcal{M}|_x$ to denote the universe of the sort X. The language $\mathcal{L}$ in turn depends on the sets $\mathbf{E}$ and $\mathbf{F}$. We therefore sometimes write $\mathcal{L}(\mathbf{E}, \mathbf{F})$. $\mathcal{L}(\mathbf{E}, \mathbf{F})$ contains three sorts: *Booleans* (variables of the sort will be denoted by b); *time points* (t) and *observables* (x). There are two 'subsorts' to observables: *fluents* (variables of the sort denoted by f) and *events* (e). We explain what we mean by 'subsort' below. The set $\mathbf{B}$ of Boolean constants contains two elements: $\mathbf{B} = \{\text{TRUE}, \text{FALSE}\}$. The set of time-point constants is $\mathbf{N}_0 = \{0, 1, 2, \ldots\}$, i.e. the set of nonnegative integers. The set of fluent constants coincides with $\mathbf{F}$; the set of event constants is identified with $\mathbf{E}$. $\mathcal{L}(\mathbf{E}, \mathbf{F})$ contains the following functions and predicates:

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

- Ab_1, Ab_2: $Ab_2(f,t)$ and $Ab_1(e,t)$ will stand for *abnormalities.* They will be used and explained only in Section 9.6.5.

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

General First-Order Causal Theories

We are now ready to define first-order causal theories. In the appendix to this chapter (page 235), we show formally that the definitions below are indeed a straightforward extension of the propositional causal theories defined earlier.

Definition 9.15 *A first-order causal theory for a language $\mathcal{L}(\mathbf{E},\mathbf{F})$ is a tuple $\langle \text{EQ}, \text{CONS} \rangle$ where*

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

$$\forall x,t. \qquad Do(x,\text{TRUE},t) \supset Ho(x,t) \tag{9.33}$$

$$\forall x,t. \qquad Do(x,\text{FALSE},t) \supset \neg Ho(x,t) \tag{9.34}$$

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

$$\forall\ x,t\ .\ [\ \neg Do(x,\text{TRUE},t) \quad \wedge \quad \neg Do(x,\text{FALSE},t)\ ]\ \supset \\ [\ Ho(x,t) \equiv\ \Phi(x,t)\ ] \tag{9.35}$$

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

2. CONS *is a set of sentences for $\mathcal{L}(\mathbf{E},\mathbf{F})$ containing at least uniqueness-of-names (UNA) and domain closure (DC) axioms[4] for the sets $\mathbf{B},\mathbf{E}$ and $\mathbf{F}$.*

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

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

This 'minimization within a context' can be conveniently implemented using circumscription [94, 106]. For details about circumscription, we refer to [94]. In an

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

appendix to this chapter (page 234) we give Lifschitz' characterization of circumscription in model-theoretic terms. From that definition, it can be seen immediately that circumscribing *Do* in CONS with all other functions and predicates *fixed* will give us exactly the models we want: if a predicate is kept fixed while circumscribing *Do*, each interpretation of the predicate will serve as a context within which *Do* is minimized. Also, just as in the propositional case, we have to minimize *Do* in CONS *before* we add the set of structural equations EQ to it – cf. the remark in Section 9.4.1. This is reflected in the following definition. The expression $Circum(\textsc{cons}; Do)$ stands for the *circumscription of Do in* CONS *with all other functions and predicates kept fixed.*

Definition 9.16 *A structure $\mathcal{M}$ for the language $\mathcal{L}$ is a* model *for the first-order causal theory T (written as $\mathcal{M} \models_{\mathbf{c}} T$) iff*

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

Note that for any causal theory $T = \langle \textsc{eq}, \textsc{cons} \rangle$ and any $\mathcal{M}$ we have that if $\mathcal{M} \models_{\mathbf{c}} T$, then also $\mathcal{M} \models \textsc{cons}$. Since by Definition 9.15 above CONS contains both UNA- and DC-axioms for **B**, **E** and **F**, we easily see that the following proposition holds:

Proposition 9.17 *We may assume without loss of generality that for any model $\mathcal{M}$ for a causal theory T, we have*

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

We now turn to the exact kind of rules of form (9.35) that we will need in order to model persistence. Fluents are supposed to behave as before: if no intervention takes place, they persist. This is modeled by the following axiom in EQ which we call our *persistence axiom*:

$$\begin{array}{l}\forall f, t\ .(t > 0) \supset \\ \quad [\ \neg Do(f, \textsc{true}, t) \quad \wedge \quad \neg Do(f, \textsc{false}, t)\ ] \ \supset \ [\ Ho(f, t) \equiv \ Ho(f, t-1)\ ]\end{array} \tag{9.36}$$

which can easily be rewritten into the form prescribed by the definition of first-order causal theories above. Concerning events, we can opt for either one of two possibilities, The first is reminiscent of the way actions are treated in the standard situation calculus [107]: between each two 'time points', exactly one action happens. We can formalize this by adding an extra axiom to CONS ('$\exists!$' stands for 'there exists exactly one'):

$$\forall t \exists! e.\ Ho(e, t) \tag{9.37}$$

The formula $Ho(e, t)$ is to be interpreted as saying that event e takes place between time t and time $t+1$. The second, more complicated possibility is to allow for multiple actions happening at the same time; this is similar to what happens in the works of Morgenstern & Stein and Baral, Gelfond & Provetti [12, 147]. We first discuss the former possibility, delaying treatment of the latter until Section 9.6.5.

9.6.3 Handling Events like in the Situation Calculus

We first extend our definition of causal theories to incorporate axioms (9.36) and (9.37).

Definition 9.18 *A* first-order causal theory with persistence *for the tuple* $\langle \mathbf{E}, \mathbf{F} \rangle$ *is a tuple* $\langle \text{EQ}, \text{CONS} \rangle$ *where* EQ *and* CONS *are sets of sentences for the language* $\mathcal{L}(\mathbf{E}, \mathbf{F})$*,* EQ *consists of intervention axioms (9.33),(9.34) and the persistence axiom (9.36) while* CONS *contains at least UNA- and DC-axioms for* $\mathbf{B}, \mathbf{E}$ *and* $\mathbf{F}$ *and axiom (9.37).*

Example 9.19 [Yale Shooting Problem] Armed with this definition, we are able to handle most standard reasoning domains involving more than two time points. As an example, consider the Yale Shooting domain as described on page 190, where it was modeled in a 'naive' way. To model it instead as a causal theory, let T_{YSP} be a causal theory with persistence for sets $\mathbf{E} = \{Load, Wait, Shoot\}$, $\mathbf{F} = \{Alive, Loaded\}$. Apart from the axioms mentioned above, the set CONS further contains the following axioms:

$$Ho(Load, t) \supset Do(Loaded, \text{TRUE}, t+1) \tag{9.38}$$

$$Ho(Loaded, t) \wedge Ho(Shoot, t) \supset Do(Alive, \text{FALSE}, t+1) \tag{9.39}$$

$$Ho(Alive, 0) \wedge \neg Ho(Loaded, 0) \tag{9.40}$$

$$Ho(Load, 0) \wedge Ho(Wait, 1) \wedge Ho(Shoot, 2) \tag{9.41}$$

It is instructive to compare these axioms with the original axioms on page 190. Recall that with the original formulation, we obtained an unintended model in which Fred remained alive after the shooting. We will show that we now only get intended models. To see this, notice first that by axiom (9.37) we have that the only events taking place in *any* model at times 0, 1 and 2 are those introduced in axiom (9.41). The circumscription of Do in the definition of causal models then makes sure that in all models of T_{YSP} we have $\neg Do(Loaded, b, t)$ for all b and $t \in \{0, 2\}$. On the other hand, by axioms (9.41) and (9.38) we have $Do(Loaded, \text{TRUE}, 1)$ in all models, too. It follows by axiom (9.33) that we have $Ho(Loaded, 1)$ in all models and by the persistence axiom (9.36) that we have $Ho(Loaded, 2)$ in all models. Since by axiom (9.41) we have $Ho(Shoot, 2)$ in all models, the antecedent of axiom (9.39) holds in all models for t instantiated to 2 and we have $Do(Alive, \text{FALSE}, 3)$ in all models. By axiom (9.36) we then have $\neg Ho(Alive, 3)$ in all models for T_{YSP}. It is easy to show that such models indeed exist.

We state without proof that we also handle the related 'Stanford Murder Mystery' [8]. The current version still has a problem if the schedule of actions that take place in the domain is incompletely specified. This problem will be discussed and solved in Section 9.6.5, where we show how to extend our theories for domains where arbitrarily many events may happen at the same time. First, we will discuss another well-known reasoning domain that we can already express with our current restrictions on the occurrence of events.

9.6.4 Ramifications Again: Dependent Fluents

Giunchiglia and Lifschitz [57] argue that we sometimes want to express dependencies between fluents where these dependencies may be partially unknown. Let us consider

Giunchiglia and Lifschitz' [57] motivating example: suppose we are in a room with a baby, an object and a table. The object may be dangerous to the baby (for example, it might be a hammer) but we are not sure about that. We do know however that if the object is placed on the table, it is out of reach of the baby, and hence it is safe. In other words, the safeness of an object depends on whether or not it is on the table, but we do not know exactly in what way: we know it is safe if it is on the table; if it is not on the table, we just know that this fact *determines* whether it is safe or not (see [57] for details). This implies that if we put an object on the table and then remove it again, we want there to be only two possibilities: either the object was safe both before and after it lay on the table or it was unsafe both before and after it lay on the table. Giunchiglia and Lifschitz introduce what they call the 'high-level action language $\mathcal{ARD}$' in order to deal with this kind of dependencies. It turns out that causal theories can be instantiated for 'dependent fluents' in a straightforward manner, the reason being that we are allowed the use of *assumption symbols* (unobserved variables) in causal theories.

We first extend our formalism by introducing the new subsort of *dependent fluents*; we will assume there are a finite number of them, listed in the set $\mathbf{D}$. Variables of the subsort will be indicated by d. Like regular fluents, dependent fluents may or may not persist. But unlike the situation for regular fluents, the behaviour of dependent fluents is fully defined in terms of other (regular) fluents and assumption symbols. We first extend our definitions of causal theories to deal with the subsort of dependent fluents. The language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F})$ is as the language $\mathcal{L}(\mathbf{E}, \mathbf{F})$ but now with dependent fluents as a new subsort of observables and with $\mathbf{D}$ indicating the constants of this new subsort.

Definition 9.20 *A* first-order causal theory with persistence and dependent fluents *for the tuple* $\langle \mathbf{D}, \mathbf{E}, \mathbf{F} \rangle$ *is a tuple* $\langle \text{EQ}, \text{CONS} \rangle$ *where* EQ *and* CONS *are sets of sentences for the language* $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F})$*,* EQ *consists of intervention axioms (9.33),(9.34) and the persistence axiom (9.36) while* CONS *contains at least UNA- and DC-axioms for* $\mathbf{B}, \mathbf{D}, \mathbf{E}$ *and* $\mathbf{F}$ *and axiom (9.37).*

Note that the only difference between causal theories as defined here and those defined as in the previous section (Definition 9.18) is the addition of UNA- and DC-axioms for dependent fluents. The new definition allows us to formalize the example described above:

Example 9.21 Let T_{DF} be a first-order causal theory with persistence for the language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F})$ containing the additional 0-ary predicate U_{Safe}. Here $\mathbf{D} = \{Safe\}$, $\mathbf{E} = \{Put\_On\_Table, Remove\_From\_Table\}$ and $\mathbf{F} = \{On\_Table\}$ and CONS contains the following additional axioms:

$$Ho(Put\_On\_Table, t) \supset Do(On\_Table, \text{TRUE}, t + 1) \tag{9.42}$$

$$Ho(Remove\_From\_Table, t) \supset Do(On\_Table, \text{FALSE}, t + 1) \tag{9.43}$$

$$Ho(Safe, t) \equiv [\ Ho(On\_Table, t) \lor (\neg Ho(On\_Table, t) \land U_{Safe})\ ] \tag{9.44}$$

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

$$\neg Ho(On\_Table, 0) \wedge Ho(Put\_On\_Table, 0) \wedge Ho(Remove\_From\_Table, 1) \qquad (9.45)$$

Notice that (9.45) together with (9.42) and (9.43) make sure that we have $Do(On\_Table, \text{TRUE}, 1)$ and $Do(On\_Table, \text{FALSE}, 2)$ in all models. The circumscription of Do in the definition of models for causal theories makes sure that we have $\neg Do(x, b, t)$ for all other x, b and $t \in \{0, 1\}$. It follows that all models for T_{DF} have

$$\neg Ho(On\_Table, 0) \wedge Ho(On\_Table, 1) \wedge \neg Ho(On\_Table, 2) \qquad (9.46)$$

Now in any model either U_{Safe} holds or it does not. For the class of models $\mathbf{M}$ with $\neg U_{Safe}$, we have by (9.44) and (9.46) that $\mathbf{M} \models \neg Ho(Safe, 0) \wedge \neg Ho(Safe, 2)$. For the class $\mathbf{M}'$ with U_{Safe}, we have $\mathbf{M}' \models Ho(Safe, 0) \wedge Ho(Safe, 2)$. It immediately follows that there can be no models with $Ho(Safe, 0) \equiv \neg Ho(Safe, 2)$. It remains to be shown that there do exist models for T_{DF} with U_{Safe} and models with $\neg U_{Safe}$; such models can indeed easily be constructed; we omit the details. Summarizing:

Proposition 9.22 *There exists a model $\mathcal{M}$ for T_{DF} with $\mathcal{M} \models Ho(Safe, 0) \wedge Ho(Safe, 1) \wedge Ho(Safe, 2)$. There is a model $\mathcal{M}'$ for T_{DF} with $\mathcal{M} \models \neg Ho(Safe, 0) \wedge Ho(Safe, 1) \wedge \neg Ho(Safe, 2)$. There are no models for T_{DF} with any other interpretation of $Ho(Safe, t)$ for $t \in \{0, 1, 2\}$.*

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

9.6.5 Minimizing Occurrence of Events

We have seen how to deal with domains in which we allow one event at a time to happen. Following Morgenstern & Stein [112, 147], Baral, Gelfond and Provetti [11, 12] and several other authors, we would like to extend this to the case where more than one event may happen at a time. We want our domains to be subject to the assumption that normally, events don't happen unless there is a specific reason for them to happen.

This is easy to model with our Pearlian theories: it turns out that we can properly model events, just like fluents, using structural equations. The advantage of using structural equations instead of axioms in CONS will become clear in Section 9.6.6. The new structural equations will say that 'if no intervention takes place that makes an event happen, and if nothing abnormal is the case, then the event will not happen'. The corresponding axiom in EQ will look as follows (the notation has been chosen to conform to (9.35)) :

$$\forall e, t \,.\, [\, \neg Do(e, \text{TRUE}, t) \;\;\wedge\;\; \neg Do(e, \text{FALSE}, t) \,] \;\supset\; [\, Ho(e, t) \equiv \; Ab_1(e, t) \,] \qquad (9.47)$$

This uses an 'abnormality predicate' $Ab_1(e, t)$ defined for all event-time pairs. An instantiated abnormality predicate plays a role similar to the 'assumption symbols' we have seen before. But unlike these, which represented things we were completely ignorant about, the abnormalities stand for things we consider abnormal, or, in other words, *unlikely*. For this, we extend the notion of causal model to 'preferred causal model'. We always prefer those models of our theory that are the least 'abnormal'. Since in the next subsection we will encounter domains involving both *rather* and *highly* abnormal eventualities, we need to introduce *two* abnormality predicates Ab_1 and Ab_2 in the definition below. Ab_1, standing for the 'weak' abnormality, is defined over event-time pairs (e, t). Ab_2, standing for 'strong' abnormality, over fluent-time pairs (f, t) (we will see how to use Ab_2 in the next section). The definition below says that the models with the smallest interpretations of the strong abnormality predicate Ab_2 should be preferred, and, among the remaining models, the models with the smallest interpretations of the weak abnormality predicate Ab_1.

Definition 9.23 *A model $\mathcal{M}$ is a preferred causal model for causal theory T if*

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

In the Epilogue to part III of this thesis (page 269) we show how the notion of 'preferred models' is connected to probability theory, and as such is already implicitly present in Pearl's original (probabilistic) theories. Having defined preferred models, we are in a position to give the definition of causal theories with concurrent events. The only difference to the previous definition (Definition 9.20) is that EQ must now also contain the 'no events'-axiom (9.47) while CONS does not contain the 'one-event-at-a-time' axiom (9.37) any more.

Definition 9.24 *A* first-order causal theory with persistence, dependent fluents and concurrent events *for the tuple* $\langle \mathbf{D}, \mathbf{E}, \mathbf{F} \rangle$ *is a tuple* $\langle \text{EQ}, \text{CONS} \rangle$ *where* EQ *and* CONS *are sets of sentences for the language* $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F})$, EQ *consists of intervention axioms (9.33),(9.34), persistence axiom (9.36) and no-events axiom (9.47) while* CONS *contains at least UNA- and DC-axioms for* $\mathbf{B}, \mathbf{D}, \mathbf{E}$ *and* $\mathbf{F}$.

Note first that with this definition, we *still* handle standard reasoning domains like the Yale Shooting Problem:

Example 9.25 [YSP, continued] Let us consider the theory $T_{\text{YSP},2}$ which is as T_{YSP} but adapted to Definition 9.24: EQ now contains axioms (9.33), (9.34), (9.36) and (9.47), while CONS contains UNA- and DC-axioms and additionally axioms (9.38)–(9.41). One sees that the circumscription of CONS rules out all models with $Do(e, b, t)$ for any e, b and t. It follows by axioms (9.47) and (9.41) that all models for $T_{\text{YSP},2}$ have the abnormalities

$$Ab_1(Load, 0), Ab_1(Wait, 1), Ab_1(Shoot, 2) \tag{9.48}$$

Clearly, there will be no models with even more abnormalities: no more events will happen than those we specified to happen. It is easy to check that this means that in all preferred models for $T_{\text{YSP},2}$, Fred is not alive any more at $t = 3$. We can now also handle the case where at some points in time, no actions at all need to take place. If we replace (9.41) by the following axiom,

$$Ho(Load, 0) \wedge \; \exists t_1, t_2. \; [\; Ho(Wait, t_1) \wedge Ho(Shoot, t_2) \wedge t_1 > 0 \wedge t_2 > t_1 \;] \tag{9.49}$$

then we will still only prefer models $\mathcal{M}$ with, for some t_1 and $t_2 > t_1 > 0$:

$$\mathcal{M} \models Ho(Wait, t_1) \wedge Ho(Alive, t_2) \wedge Ho(Shoot, t_2) \wedge \neg Ho(Alive, t_2 + 1).$$

9.6.6 Causal Chains of Events

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

Example 9.26 Imagine a domain where, if you push somebody, he or she falls down a moment later. We can formalize this using a causal theory according to Definition 9.24 such that CONS contains the axiom

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

Now suppose CONS further contains axiom $Ho(Push, 0)$. From inspection of axiom (9.47) and the intervention axioms (9.33) and (9.34), we see that all preferred models for this theory will have $Ho(Push, 0) \wedge Ho(Fall, 1)$. But if CONS had not contained $Ho(Push, 0)$, then the preferred models would be those in which no events at all take place.

9.6.7 Surprise, Surprise

What if a fluent changes value while we have no event in our domain which can account for that? This is the kind of situation that Sandewall and Shoham [135] call a 'surprise'; Lifschitz and Rabinov [98] call it a 'miracle'. In order to deal with it, we have to weaken our structural equations concerning regular fluents: persistence may now be broken not only by some intervention, but also by some 'abnormal' (unlikely) external influence. For such abnormalities, we will use our second abnormality predicate Ab_2. We change axiom (9.36) into the following *weakened persistence axiom*:

$$\begin{aligned} &\forall f, t \;.\; (t > 0) \;\supset\; [\; (\neg Do(f, \text{TRUE}, t) \quad \wedge \quad \neg Do(f, \text{FALSE}, t)) \;\supset \\ &\quad [\; Ho(f, t) \equiv \;[\; (Ho(f, t-1) \wedge \neg Ab_2(f, t)) \vee (\neg Ho(f, t-1) \wedge Ab_2(f, t)) \;]\;]\;] \end{aligned} \tag{9.50}$$

The notation (9.50) has been chosen so as to conform to the syntactic form we used in (9.36). An equivalent, perhaps more intuitive way of stating it is by the following two axioms, whose conjunction is equivalent to (9.50):

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

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

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

Definition 9.27 *A* first-order causal theory with persistence, dependent fluents, concurrent events and surprises *for the tuple* $\langle \mathbf{D}, \mathbf{E}, \mathbf{F} \rangle$ *is a tuple* $\langle \text{EQ}, \text{CONS} \rangle$ *where* EQ *and* CONS *are sets of sentences for the language* $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F})$*,* EQ *consists of intervention axioms (9.33),(9.34),* weakened *persistence axiom (9.50) and no-events axiom (9.47) while* CONS *contains at least UNA- and DC-axioms for* $\mathbf{B}, \mathbf{D}, \mathbf{E}$ *and* $\mathbf{F}$.

We see that the only difference to the previous Definition 9.24 is that persistence axiom (9.36) has been replaced by its weakened version (9.50). We illustrate the use of 'surprises' by Kautz' Stolen Car Domain which we introduced on page 191.

Example 9.28 [stolen car problems] Here we consider two variations of this domain, modeled by theories $T_{\text{SC},1} = \langle \text{EQ}, \text{CONS}_1 \rangle$ and $T_{\text{SC},2} = \langle \text{EQ}, \text{CONS}_2 \rangle$. Both are causal theories according to Definition 9.27 for the tuple $\langle \mathbf{D}, \mathbf{E}, \mathbf{F} \rangle$ with $\mathbf{D} = \emptyset$, $\mathbf{E} = \{\mathit{Steal\_Car}\}$, $\mathbf{F} = \{\mathit{Car\_In\_Lot}\}$. On top of the UNA- and DC-axioms, CONS_1 and CONS_2 both contain observations:

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

CONS_2 contains no further axioms, while CONS_1 additionally contains:

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

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

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

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

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

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

9.7 Conclusion

We have shown that by extending Pearl's causal theories we arrive at a powerful approach to common sense reasoning about action and change. We had to extend Pearl's theory at several places; however, the basic idea behind Pearl's theory, i.e. the sufficient cause principle, remained unchanged. The main ingredient of our causal theories is the *Do*-operator, which allows us to express any propositional combinations of observations (in the propositional case, these are elements of **V**; in the first-order case, instances of *Ho*) and interventions (instances of *Do*).

In the introduction to this chapter we set ourselves three goals. We said that some of them would be dealt with in the present chapter and some in the next. Let us see what we have achieved this far:

Our first goal was to gain a better understanding of the rôle of causality in NMTR. Our contribution towards this goal was given mainly in Section 9.5. There we showed in detail that, while

$$Do(F_1, \textsc{true}) \supset Do(F_2, \textsc{true}) \tag{9.51}$$

can often be read as 'F_1 causes F_2', it has a semantics that can be understood fully in non-causal terms. We showed that this makes it possible to use axioms like (9.51) even in contexts where they do not correspond to the colloquial usage of 'causes'. We also showed how other approaches to NMTR, which use a predicate called *Caused* (or similarly) with a semantics very similar to *Do*, express relationships like (9.51) in a different way, sometimes leading to unintended results (Section 9.5.3). Finally, in Section 9.6.5 we showed that the *Do*-predicate with its 'sufficient cause principle' semantics can be used to model causal chains of events in a natural way.

The second goal was to provide a theory that could deal with a wide class of reasoning domains. Towards this goal, we have given many specific examples of reasoning domains that are problematic for many other approaches, and we have shown that our theory deals with them in an intuitive way. However, as we stressed in Chapter 8, Section 8.3.1, this does not give much confidence that our causal theories will work well on any full *class* of reasoning domains. Providing evidence that they do work well for full classes too is one of the two main goals of the next chapter.

The third goal, which is to show how existing approaches are interrelated and to explain some of their successes and failures, will be the second of the two main goals of the next chapter.

The Overall Research Goal To end this chapter, we indicate how our causal theories fit in the general 'alternative research goal' we set ourselves in Chapter 8, Section 8.4. Let us concentrate on the most sophisticated version of our theories, i.e. Definition 9.27 on page 232. Recall that the alternative research goal concerned a fundamental distinction between finding a good Model of a domain and finding a concise representation of that Model ('Model', when written with capital M, is used in the general sense; not in the logician's – see page 198). In the present causal theories, this distinction is adhered to as follows. We specify our causal theories by two sets of axioms: EQ and CONS. EQ contains axioms embodying persistence and the sufficient cause principle, CONS contains axioms standing for state constraints and for effects of events (e.g. $Ho(Shoot, t) \supset Do(Alive, \text{FALSE}, t+1)$) and of interventions (e.g. $Do(Alive, \text{FALSE}, t) \supset Do(Walking, \text{FALSE}, t)$). Nowhere do we mention any sentences about the non-effects of events; nevertheless there are many of these which we want to hold in all our models. For example, in the YSP domain (page 227) as modeled by theory T_{YSP}, we surely want the following to hold in *all* models:

$$\neg Ho(Shoot, t) \supset \neg Do(Alive, \text{FALSE}, t+1)$$

Yet this is neither enforced by CONS nor by EQ. Hence, T_{YSP} is an incomplete specification (in the sense of page 199) of the Model we want to specify by the theory T_{YSP}. T_{YSP} is turned into a complete specification in the first step of the definition of causal models (Definition 9.16), by circumscribing Do in CONS and then adding EQ. This first step performs the 'completion mapping' we defined on page 200. The set $\mathbf{M}$ of (logician's) models of $Circum(\text{CONS}; Do) \wedge \text{EQ}$ corresponds to our intended Model of the Yale Shooting domain. The minimization of abnormalities was done with respect to this set of models $\mathbf{M}$. It only served to select the set of *least surprising* models within $\mathbf{M}$; *all* models in the set $\mathbf{M}$ stand for *possible* realizations of the domain we want to model. We see that both uses of nonmonotonicity are strictly separated, as required by our alternative research goal: the circumscription of Do corresponds to the completion mapping, the preference order of abnormalities to the internal preference structure (page 200) of $\mathbf{M}$.

9.8 Appendix: The Model-Theoretic Characterization of Circumscription

The following is all taken from [94].

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

(i) $|\mathcal{M}_1| = |\mathcal{M}_2|$
(ii) $\mathcal{M}_1[K] = \mathcal{M}_2[K]$ for every constant K not in P, Z
(iii) $\mathcal{M}_1[P_i] \subseteq \mathcal{M}_2[P_i]$ for every P_i in P.

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

The 'circumscription of P in T with Z varied', written as $Circum(T;P;Z)$ is defined as a formula in second-order logic (we will not repeat this formula here, see [106, 94]). However, we may also characterize circumscription as follows:

Proposition 9.29 (Lifschitz 1985) *A structure $\mathcal{M}$ is a model of $Circum(T;P;Z)$ iff $\mathcal{M}$ is minimal in the class of models of T with respect to $<^{P;Z}$.*

9.9 Appendix: Pearl's Causal Theories

We claimed at various places that our theories are simple extensions of Pearl's. However, none of our definitions appear anywhere in Pearl's articles. Hence, in order to convince the sceptical reader we will establish a formal relationship between Pearl's original theories and ours. Both our propositional and first-order causal theories for dealing with persistence have their roots in definitions 9.1 and 9.3, so it is enough if we can show how these definitions are formally related to Pearl's original definitions. That is what we will do in this appendix: we show in detail how exactly to extend Pearl's causal theories in order to arrive at definitions 9.1 and 9.3. We use the version of Pearl's theories introduced in [119, 120]. We first give Pearl's original definitions. We then show how a set of global constraints CONS is introduced; this gives us a new kind of causal theories that are more powerful than Pearl's. We then show that these are equivalent to the propositional causal theories introduced in Section 9.3. We end the appendix by saying something about the probabilistic ingredient in Pearl's theories, which is ignored in our analysis.

Propositional Causal Theories

Here is the definition of causal theories as given by Pearl [119, 120]:

Definition 9.30 *A causal theory T is a 4-tuple $T = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \{f_i\}\rangle$ where*

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

$$X_i = f_i(X_1, \ldots, X_n, U_1, \ldots, U_m) \tag{9.52}$$

Equations of the form (9.52) are called structural equations.

Usually, causal theories come together with one or more *actions* of the form $Do(X_i, x)$. In the following, we assume that $\{X_{i_1}, \ldots, X_{i_l}\}$ is an arbitrary subset of the variables $\mathbf{V}$ and that for all $1 \leq k \leq l$, x_k is a value in the domain of X_{i_k}. Here is how Pearl defines the effect of actions [119, 120]:

Definition 9.31 *(Effect of actions) The effect of the set of actions*

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

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

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

Definition 9.32 *A valuation $\mathcal{M}$ for the variables in $\mathbf{V} \cup \mathbf{U}$ belonging to a causal theory T is called a* model *of T iff all of the equations (9.52) associated with T hold in $\mathcal{M}$. A valuation $\mathcal{M}$ is called* a model for the causal theory T and the set of actions $\mathbf{A}$ *iff $\mathcal{M}$ is a model for $T(\mathbf{A})$, where $T(\mathbf{A})$ is defined as in Definition 9.31.*

It is often assumed [120] that the set of equations (9.52) has a unique solution for $X_i, \ldots, X_n$, given any value of the disturbances $U_1, \ldots, U_m$; in other words, each set of values for the disturbances determines a unique model for T. Therefore the distribution $P(\mathbf{u})$ induces a unique distribution on the set of variables $\mathbf{U} \cup \mathbf{V}$, or, equivalently, on the set of *models*: a model $\mathcal{M}$ such that $U_1 = u_1, \ldots, U_m = u_m$ will receive probability $P\{\mathcal{M}\} = P\{U_1 = u_1, \ldots, U_m = u_m\}$. We can define the notion of *preferred* model in terms of its probability:

Definition 9.33 *For any causal theory T, any valuation $\mathcal{M}$ of the variables in $\mathbf{V} \cup \mathbf{U}$ with maximum probability $P\{\mathcal{M}\}$ will be called a* preferred model *for the causal theory.*

For the time being we just look at plain models for causal theories, i.e. we do not care about the distribution $P(\mathbf{u})$. In that case, the assumption that the set of equations (9.52) has a unique solution for $X_i, \ldots, X_n$, given any value of the disturbances $U_1, \ldots, U_m$ is not needed. We return to the use of $P(\mathbf{u})$ and to 'preferred models' at the end of this appendix (page 238).

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

Definition 9.34 *An extended propositional causal theory T_{EP} is a 5-tuple*

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

such that

1. *$\langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \{f_i\} \rangle$ is a propositional basic causal theory as defined in Definition 9.30.*

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

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

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

Definition 9.35 *A model for an extended causal theory* $T_{EP} = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \{f_i\}$, CONS$\rangle$ *is a valuation* $\mathcal{M}$ *for the variables in* $\mathbf{V} \cup \mathbf{U} \cup \mathcal{A}(\mathbf{V})$ *such that*

1. $\mathcal{M} \models \text{CONS}$
2. *The restriction of* $\mathcal{M}$ *to the variables in* $\mathbf{V} \cup \mathbf{U}$ *is a model of the basic causal theory* $T^*_{EP}(\mathbf{A})$. *Here* $T^*_{EP}(\mathbf{A})$ *is the effect of the set of actions* $\mathbf{A}$ *on the basic causal theory* T^*_{EP}*, where*

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

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

Note that the set $\mathbf{A}$ in item 2 of this definition depends on the model $\mathcal{M}$, i.e. it can be different for different $\mathcal{M}$.

We would like to show that the definitions of T_{EP} and its models are equivalent to the definitions of causal theories and their models used in the main text (definitions 9.1 and 9.3). However, there are still a few small differences, the most important one being that, in contrast to the theories of the main text, theories T_{EP} must have one structural equation with X_i on the left hand side for *every* X_i in $\mathbf{V}$. But this difference is not crucial, as we proceed to show:

For any extended causal theory $T_{EP} = \langle \mathbf{V}, \mathbf{U}, P(\mathbf{u}), \{f_i\}, \text{CONS}\rangle$, let T'_{EP} be the theory that results from deleting from T_{EP} all functions f_i in $\{f_i\}$ of the form $X_i = U$ where U is some element of $\mathbf{U}$. It is easy to see the following:

Proposition 9.36 $\mathcal{M}$ *is a model for* T_{EP} *iff* $\mathcal{M}$ *is a model for* T'_{EP}.

Hence when working with extended propositional theories T_{EP}, we do not have to worry about specifying functions f_i for *all* the $X_i \in X$ – if an f_i is left out, it just means that we are indifferent about the corresponding X_i. But now notice that, since we are working with propositional variables, we can replace the equality sign '=' in the structural equations (9.52) by logical equivalence '≡' without changing their meaning. From this, together with the Proposition 9.36 above and the definition of effects (Definition 9.31) it immediately follows that the definitions of extended propositional causal theories T_{EP} and their models (definitions 9.34 and 9.35, resp.) are completely equivalent to the definitions of causal theories and models given in Section 9.3 (definitions 9.1 and 9.3, resp.). We have thus shown how these definitions arise as natural extensions of Pearl's definitions, which was our goal in this appendix.

9.9.1 Conclusion; the use of $P(\mathbf{u})$

We have seen how our propositional causal theories can be viewed as extensions of Pearl's theories. In Section 9.6, we lifted these theories to the first-order case. In the end this led to Definition 9.27 (page 232), the most sophisticated version of our causal theories. These were given a non-monotonic semantics in Definition 9.23: we prefer (roughly speaking) the models of such theories with the least number of abnormalities. This minimization of abnormalities is also already implicit in Pearl's original theories, where it can be handled by the probability distribution $P(\mathbf{u})$ that is implicit in causal theories but that we neglected thus far (see the remark below Definition 9.33). One can show that, if one assumes the probability distribution $P(\mathbf{u})$ to satisfy certain natural requirements, then the 'preferred models' of a causal theory (Definition 9.33) coincide with the models with the least abnormalities. We shall return to this issue in the Epilogue to part III of this thesis, where we show the relation between abnormality and probability.

Chapter 10

Causal Theories and Other Approaches

In the previous chapter we introduced 'causal theories'. In this chapter we give a detailed comparison of causal theories to three existing 'state-of-the-art' approaches to reasoning about action: those of McCain & Turner (Section 10.2), Lin (Section 10.3) and Baral, Gelfond & Provetti (Section 10.4). We briefly mention similarities to some other approaches (Section 10.5) and we end with some conclusions. The chapter is followed by two appendices containing proofs of the theorems and propositions presented throughout the chapter.

10.1 Introduction

In this chapter we give detailed comparisons of causal theories to other, existing approaches. This will serve two goals:

1. Provide evidence that our approach can truly handle a large class of reasoning domains.
2. Understand how existing approaches are inter-related; understand their successes and failures.

Both goals will be dealt with using a 'pragmatic' variation of the 'systematic methodology' introduced in Chapter 8, Section 8.3.1: in order to compare our approach to some alternative approach A we will show *formally* that it yields the same inferences as A on a large *class* of reasoning domains. In some cases we will also provide a domain that falls outside this class, and for which our approach gives more intuitive results than approach A. Such a result provides evidence (but not proof of course) that our approach gives intuitive results for a superset of the domains for which approach A gives intuitive results. In this way, goal (1) above will be achieved. The comparisons and formal correspondences will connect alternative approaches not only to our approach, but also to Pearl's 'sufficient cause principle'. This will shed a new light on their treatment of causality, and as such, goal (2) will be achieved.

10.1.1 How The Comparisons Will Be Done

We will consider three approaches in detail: those of McCain & Turner (Section 10.2), Lin (Section 10.3) and Baral, Gelfond & Provetti (Section 10.4). Each of these will be compared to the instantiation of our theory that is conceptually closest to it. Hence McCain & Turner's approach, which has been specifically designed to handle the ramification problem, will be compared to our 2-point causal theories (page 214) which are complex enough to deal with ramifications but which contain no additional features that would only serve to complicate the comparison. In the same vein, Lin's approach will be compared to first-order theories for persistence in which only one event may happen at a time (page 225), and Baral, Gelfond & Provetti's method will be compared to the version of our theory that minimizes occurrences of events but that does not handle surprises (page 229).

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

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

The importance of the equivalence theorems is that they show that some approaches which at first glance look rather different from ours are actually quite similar; therefore they are also connected to Pearl's sufficient cause principle, albeit implicitly. In our comparison to Lin's [99, 100] approach, we have not bothered to try and prove an equivalence theorem, since it is easy to see, just by looking at the definitions, that his approach is almost equivalent to ours.

10.2 McCain & Turner's Theory of Ramifications and Qualifications

Recently, McCain & Turner (MT) have introduced a 'causal theory' that focuses on handling ramification constraints [103, 105, 104]. We will now compare their approach to our two-point causal theories as defined by Definition 9.6 and 9.7 on page 214.

MT consider theories that are triplets (S, E, C), defined for a propositional language where each atom stands for a fluent. The 'state of the world' S is an interpretation for all fluents. An interpretation is denoted by the set of literals true in it. E is a set of 'explicit effects', i.e. propositional combinations of fluents. Intuitively, they are the formulas that are explicitly caused to hold by some (unspecified) action. C is a set of 'causal laws' that determine the ramifications of effects. MT define the function[1] $\Pi_1(S, E, C)$ such that it gives the set of possible states of the world after an action with effects E has taken place in state S. The exact definition of $\Pi_1(S, E, C)$ can be found in Appendix 10.7.1; here we just give an example of its use:

$$\begin{aligned} S &= \{\mathit{Alive}, \mathit{Walking}\} \\ E &= \{\neg \mathit{Alive}\} \\ C &= \{\neg \mathit{Alive} \Rightarrow \neg \mathit{Walking}\} \end{aligned} \tag{10.1}$$

In this case, $\Pi_1(S, E, C) = \{\{\neg \mathit{Alive}, \neg \mathit{Walking}\}\}$ i.e. the change of *Alive* brought about a change of *Walking*. However, if we had had

$$S' = \{\neg \mathit{Alive}, \neg \mathit{Walking}\}, E' = \{\mathit{Walking}\}, C' = C,$$

then $\Pi_1(S', E', C')$ would have been empty: '$\Rightarrow$' has a function similar to our '*Do*', enabling changes of right-hand side fluent values given changes of left-hand side values, but *not* the other way around (compare this to Example 9.8 of Chapter 9). There is one sort of domain constraint that can be expressed in MT's approach but not in ours: MT allow effects to be *any* propositional combination of fluents, and thus an effect may be a disjunction of two fluents. This cannot be expressed by our 2-point causal theories (we have no construct of the form $Do(X \vee Y, b)$). But it is exactly here that MT can give counterintuitive results; to see this, consider the general case where there is an effect X that further causes $Y \vee Z$, and an initial state with $\neg X, \neg Y$ and $\neg Z$:

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

Proposition 10.1 $\Pi_1(S, E, C) = \{\{X, Y, \neg Z\}, \{X, \neg Y, Z\}\}$

The proof of this proposition can be found in Appendix 10.7.2. We have seen in Example 9.11 (Chapter 9, Section 9.5.4) that this seems too strong in general.

Another difference between MT's approach and ours has already been pointed out in Section 9.5.2 of the previous chapter: there are domain constraints that would be modeled as a single axiom $A \Rightarrow B$ in MT's approach, while we prefer to model them using two separate axioms:

$$\begin{aligned} Do(A(t), \text{TRUE}) \supset Do(B(t), \text{TRUE}) \\ A(t) \supset B(t) \end{aligned} \tag{10.2}$$

[1] $\Pi_1(S, E, C)$ is the function McCain & Turner use in their recent paper [105]; it stands at the basis of their 'Causal Theory of Action and Change' [104]. It is a slight modification of their earlier next-state function $Res^4_C(E, S)$ [103]. For the precise relation between $\Pi_1(S, E, C)$ and $Res^4_C(E, S)$, see [105].

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

$$A(t) \supset Do(B(t), \text{TRUE}) \tag{10.3}$$

We have seen in Chapter 9, Section 9.5.3 that this is in general *not* the same as (10.2).

But *if* we translate constraints like $A \Rightarrow B$ indeed as (10.3), and we restrict ourselves to domains without disjunctive effects, then it turns out that MT and our approach agree on all problem domains that can be represented in the languages of both approaches. In Definitions 10.2 and 10.3 (page 243) syntactic ($\sim$) and semantical ($\cong$) correspondence for 2-point causal theories and MT's theories are defined. $\sim$ has been defined such that disjunctive effects cannot occur in corresponding theories and such that rules of the form '$A \Rightarrow B$' are translated into sentences of the form (10.3). $\cong$ is defined such that $\mathcal{M} \cong (S, S')$ iff the fluents that hold in S hold in $\mathcal{M}$ at time 0 and the fluents that hold in S' hold in $\mathcal{M}$ at time 1. As an example of how Definition 10.2 works, we consider the causal theory (S, E, C) defined by (10.1) at the beginning of this section. The theory described there corresponds to a 2-point causal theory T_{C} (the subscript C stands for 'Causal') such that CONS contains the axioms:

$$\begin{gathered} Alive(0) \wedge Walking(0) \\ Do(Alive(1), \text{FALSE}) \\ \neg Alive(1) \supset Do(Walking(1), \text{FALSE}) \end{gathered}$$

which can be seen to stem from items 3,4 and 5 of Definition 10.9, respectively. EQ contains

$$Alive(1) \equiv Alive(0) \;;\; Walking(1) \equiv Walking(0)$$

which can be seen from item 1 of the definition.

We are now ready to state our equivalence theorem, the proof of which can be found in Appendix 10.7.3. In the theorem, $\models_{\mathbf{c}}$ is defined as in Definition 9.7.

Theorem 10.4 *For any 2-point causal theory T_{C} and any domain description $T_{\text{MT}} = (S, E, C)$ such that $T_{\text{C}} \sim T_{\text{MT}}$, we have:*

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

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

10.3 Lin's Embrace of Causality

Lin [99, 100] has recently introduced a new method for reasoning about action that is based on a version of the situation calculus [107]. In the situation calculus, time is modeled by *situations s* rather than time points. The sort *actions* corresponds to our *events*; however, in Lin's approach symbols of the sort are not necessarily constants (they can be n-ary function symbols rather than only 0-ary). Similarly, fluents in situation calculus correspond to our fluents, and again, Lin allows fluent symbols to

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

Definition 10.2 *For any 2-point causal theory $T_{\mathrm{C}} = \langle \mathbf{V}, \mathbf{U}, \mathrm{EQ}, \mathrm{CONS}\rangle$ for sets $\mathbf{E}$ and $\mathbf{F}$ and any $T_{\mathrm{MT}} = (S, E, C)$ 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. *$\mathbf{E} = \emptyset$; $\mathbf{F} = \{F_1, \ldots, F_n\}$; $F_i \in \mathbf{F}$ iff F_i is an atom in the propositional language for which T_{MT} is defined; $\mathbf{U} = \emptyset$; EQ is defined as required by Definition 9.6, page 214.*
2. *Each sentence in CONS corresponds either to S or to a sentence in E or to a sentence in C. The set S and the sentences in E and C all correspond to a sentence in CONS. Here 'correspondence' is defined as in items 3-5 below:*
3. *the sentence Γ_{C} in CONS corresponds to S (and vice versa) iff Γ_{C} is of the form $F_1^{b_1}(0) \wedge \ldots F_n^{b_n}(0)$ and $S = (F_1^{b_1}, \ldots, F_n^{b_n})$.*
   Here $b_i \in \mathbf{B}$ for all $1 \leq i \leq n$. .
4. *A sentence Γ_{C} in CONS corresponds to a sentence Γ_{MT} in E iff Γ_{C} is of the form $Do(F_i(1), b)$ and Γ_{MT} is of the form F_i^b. Here $b \in \mathbf{B}$.*
5. *A sentence Γ_{C} in CONS corresponds to a sentence Γ_{MT} in C iff Γ_{C} is of the form $(m \geq 1)$*

$$\Phi \supset Do(F_{i_1}(1), b_1) \wedge \ldots \wedge Do(F_{i_m}(1), b_m) \tag{10.4}$$

   and Γ_{MT} is of the form

$$\Phi' \Rightarrow F_{i_1}^{b_1} \wedge \ldots \wedge F_{i_m}^{b_m} \tag{10.5}$$

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

Definition 10.3 *Given a model $\mathcal{M}_C$ for a causal theory T_{C} and a pair of states (S, S') for a T_{MT} with $T_{\mathrm{C}} \sim T_{\mathrm{MT}}$ we say that $\mathcal{M}_C$ corresponds to (S, S') iff for all $F_i \in \mathbf{F}$ we have*

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

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

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

$$Caused(f, \text{TRUE}, s) \supset Ho(f, s) \tag{10.6}$$
$$Caused(f, \text{FALSE}, s) \supset \neg Ho(f, s) \tag{10.7}$$

2. Circumscribe *Caused* 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.

$$\begin{aligned} &Poss(a, s) \supset \\ &\quad \{(\neg Caused(f, \text{TRUE}, Result(a, s)) \wedge \neg Caused(f, \text{FALSE}, Result(a, s))) \\ &\qquad \supset [Ho(f, Result(a, s)) \equiv Ho(f, s)]\} \end{aligned} \tag{10.8}$$

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

be n-ary. For any action e and situation s, the function[2] $Result(e, s)$ stands for the situation that results when performing e in s. Ordinary situation calculus contains only the functions described above and the predicate *Ho*, defined on fluent-situation pairs. In addition to this, Lin uses two additional predicates *Caused* and *Poss*.

It turns out that Lin's method is *very* similar to ours. We will compare it to the instantiation of causal theories in which only one event is allowed to happen at a time (Chapter 9, Definition 9.18), since this instantiation is conceptually closest to the situation calculus. Like us, Lin uses an additional sort *truth values*; a variable of the sort can be either TRUE or FALSE. Lin's *Caused*-predicate is ternary: $Caused(f, v, s)$ is true if the fluent f is caused to have the truth value v in situation s. On page 244 we give a (somewhat extended) quote of [99] 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 our work, so it will be of no concern to us. We first need the following:

Proposition 10.5 *If we exchange step 1 and step 2 in Lin's procedure we will arrive at an equivalent final theory* T'''_{LIN} .

Proof: If we rewrite (10.6) and (10.7) as disjunctions, then *Caused* appears only in negated form in them. It then easily follows from the model-theoretic characterization of circumscription (see Appendix 9.8) that the theory obtained by adding (10.6) and (10.7) before the circumscription has the same models as the theory obtained by adding them after the circumscription. □

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

Let us write $\text{EQ}_{\text{LIN}} = \{(10.6) \cup (10.7) \cup (10.8)\}$. With this notation, step 1 and step 2 interchanged and step 4 left out, Lin's approach can be rephrased as follows:

$$T''_{\text{LIN}} = \text{EQ}_{\text{LIN}} \cup \mathit{Circum}(T_{\text{LIN}}\ ; \mathit{Caused}) \tag{10.9}$$

Now if we compare EQ_{LIN} to the set of axioms EQ in our definition of causal theories (Definition 9.18, page 227), we see that they are strikingly similar: Once we rename *Caused* to *Do*, (10.6) and (10.7) actually become equivalent to our intervention axioms (9.33) and (9.34)! If we furthermore realize that $\mathit{Result}(a, s)$ refers to the first time point considered after the time point corresponding to situation s, we see that also (10.8) is almost the same as our persistence axiom (9.36). And most importantly, if we compare the characterization of Lin's approach (10.9) to the definition of models for causal theories (Definition 9.16, page 226), we see that, if EQ is as in Definition 9.18, then (10.9) above is nearly equivalent to Definition 9.16, item 1. The differences are that our approach uses integer time while Lin's uses situations and that our approach does not handle the qualification problem and infinite numbers of fluents and events (CONS contains a domain closure axiom while T_{LIN} does not). On the other hand, Lin's approach does not handle the extensions to our approach introduced in sections 9.6.4-9.6.7 of the previous chapter. Otherwise, as is hopefully clear from just looking at the respective definitions, Lin's approach is almost identical to ours. Though we have not proven it formally, we conjecture the two approaches to be equivalent on the set of reasoning domains for which both are defined.

Interestingly, in Lin's papers neither the choice to keep *Ho* fixed during circumscription of *Caused*, nor the choice to add the persistence axiom only after this first circumscription is motivated in terms much other than 'if you do not do it, you get counterintuitive results'. Our work thus provides an external motivation for these choices, cf. the remarks in Chapter 9, Section 9.4: the axioms (10.6)-(10.8) can be interpreted as mimicking the replacement of structural equations. This replacement can only be done properly if the right interpretations of *Do* have been obtained already for each particular interpretation of *Ho*. Note that, for the 'replacement' to happen correctly, it is *not* crucial that the intervention axioms are added after the circumscription of CONS (which is why Lin can add his version of them to T_{LIN} before circumscribing). It *is* however crucial that the persistence axiom is only added after the circumscription. Otherwise, as can be easily seen from the persistence axiom (9.36), we would infer that in each model with $\neg[\mathit{Ho}(F, t-1) \equiv \mathit{Ho}(F, t)]$ for some fluent F, we would have $\mathit{Do}(F, b, t)$ for some b. Thus any change would automatically be accompanied by an intervention, and we would select models with spurious changes.

10.4 Baral and Gelfond

We now undertake the most difficult and extensive comparison: we compare our approach to Baral and Gelfond's (BG) approach based on the *action description language* $\mathcal{L}_3$ [11]. $\mathcal{L}_3$ is an extension of Baral, Gelfond and Provetti's language $\mathcal{L}_1$ [12] which in turn is an extension of Lifschitz' language $\mathcal{A}$ [54, 135]. $\mathcal{L}_3$ extends $\mathcal{A}$ to deal with concurrent actions, actions with non-deterministic effects and observations of the actions that take place and the fluents that hold in arbitrary situations. On the other hand, it cannot at all deal with ramifications. The way it treats concurrent actions is similar to the way we treat events: BG always prefer the models of a domain in which as few actions as possible take place. Otherwise, at least on the surface, BG's approach looks quite different from ours. It is definitely not based on Pearl's ideas. Still, it turns out to be equivalent to our approach on most reasoning domains for which both approaches are defined.

We will compare BG's approach to our instantiation of causal theories in which concurrent events are allowed to happen but surprises are not, that is, we use Definition 9.24 (Chapter 9, page 230). The comparison will follow the same pattern as the comparison to MT's approach did: we first give an example of a reasoning domain where our approach gives better results than BG's does. We will then define syntactical and semantical correspondence relations and provide a theorem stating that, for the subset of possible reasoning domains for which the syntactical correspondence relation is defined, corresponding theories have corresponding models. We first have to explain the basics of BG's approach though. For more details we refer to [11] and to the much more extended [12]. Strictly speaking, the latter reference is about $\mathcal{L}_1$ and not $\mathcal{L}_3$, but $\mathcal{L}_3$ is only a minor extension of $\mathcal{L}_1$.

Review of $\mathcal{L}_3$

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

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

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

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

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

Third, there are *occurrence facts* which are expressions of the form

$$\alpha \textbf{ occurs\_at } s$$

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[3].

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

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

If all the rules of a domain description D are true in an interpretation M, we call M a *model* of D. The definition of a rule 'being true in interpretation M' is relatively straightforward. The precise definition of 'truth in a model' and of modelhood can be found in Appendix 10.8.1.

A Distinguishing Example We will see that for most of the domains expressible in both formalisms, BG and our causal theories give the same results. However, in domains involving *specificity*, BG and our approach give different results, and we claim that for these domains, BG gives the less intuitive ones. The default assumption of specificity says that 'more specific information about actions overrides less specific information'. We illustrate this using a standard example [150]: suppose that if you lift a bowl of soup with either your left or your right hand, but not both, then you will spill the soup and the table will get wet. If you lift the bowl with both hands however, then you will not spill the soup. BG formalize this using the following domain description D, containing only two situations s_0 and s_N:

$$\{\mathit{Lift\_left}\} \quad \textbf{causes} \quad \mathit{Wet} \tag{10.10}$$

$$\{\mathit{Lift\_right}\} \quad \textbf{causes} \quad \mathit{Wet} \tag{10.11}$$

$$\{\mathit{Lift\_left}, \mathit{Lift\_right}\} \quad \textbf{causes} \quad \neg\,\mathit{Wet} \ \textbf{if} \ \neg\,\mathit{Wet} \tag{10.12}$$

BG now formalize the assumption of specificity as follows: if the preconditions of (10.12) hold, then rules (10.10) and (10.11) are ignored when determining the effects of *Lift_left* and *Lift_right* – see the paragraph on 'causal models' in Appendix 10.8.1 for details on how this is achieved.

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

$$\mathit{Turn\_on\_sprinkler} \ \textbf{causes} \quad \mathit{Wet} \tag{10.13}$$

[3]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.

Now consider a situation in which your table is dry. Suppose in this situation you lift the soup bowl with both hands while somebody else turns on the sprinkler. While you would intuitively expect the table to get wet, according to BG's approach, this is an inconsistent domain description! More formally, suppose we have the additional facts

$$\begin{aligned} & s_0 \textbf{ precedes } s_N \ \wedge \neg \, \mathit{Wet} \textbf{ at } s_0 \wedge \\ & [\{\mathit{Turn\_on\_sprinkler}, \mathit{Lift\_left}, \mathit{Lift\_right}\}] \textbf{ occurs\_at } s_0 \end{aligned} \tag{10.14}$$

Proposition 10.6 *There exists no model M for domain description* $D = \{(10.10) - (10.14)\}$.

The proof of this proposition can be found in Appendix 10.8.2.

One should stress that there is nothing wrong with 'specificity' in itself! From our point of view, specificity just says that, just as you assume that 'no events happen in general' when determining your set of models, you may already assume it in your specification of effect axioms; in the present example, 'specificity' amounts to the default assumption that if only *Lift_left* is known to take place, then *Lift_right* is assumed not to take place (and vice versa). We did not build in such a feature in our causal theories. While this somewhat restricts the applicability of our approach, it is unrelated to the problem with BG's approach mentioned above. That problem is caused by the inappropriate use of **causes** in (10.12): there is definitely no *sufficient cause* for not getting wet if you lift the soup bowl with two hands: no intervention that sets the value of *Wet* is performed. However, **causes** does receive a semantics in BG's approach as if it would always represent an intervention. From a Pearlian point of view, it comes as no surprise that this may lead to counterintuitive results.

Because we do not feature specificity, we would have to formalize (10.10) as

$$\mathit{Ho}(\mathit{Lift\_left}, t) \wedge \neg \mathit{Ho}(\mathit{Lift\_right}, t) \supset \mathit{Do}(\mathit{Wet}, \textsc{true}, t+1) \tag{10.15}$$

and (10.11) accordingly. (10.12) would then simply disappear. If the sprinkler were turned on, then according to the definition of models for causal theories (Definition 9.16, page 226), the table would definitely get *Wet*.

Correspondence and Equivalence

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

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

  Definition 10.7 *If a domain description D in the language $\mathcal{L}_3$ contains two effect laws*

$$a \textbf{ causes } f \textbf{ if } p_1, \ldots, p_n \quad \text{and} \quad a' \textbf{ causes } \neg f \textbf{ if } p'_1, \ldots, p'_m \tag{10.16}$$

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

The exclusion of ambiguous domain descriptions will be echoed in condition (b) of Theorem 10.11 below.

- The second problem is that the definition of modelhood comes in two versions in BG's papers [12]. In one of the two versions, logically consistent initial states are *never* ruled out, even if this would lead to models in which more actions occur than is strictly necessary. In the other version only the models with a minimal number of actions are selected. Since in our approach we always prefer the models with the least number of events (i.e. the smallest, in the subset sense, interpretations of Ab_1), our approach should clearly be compared to the second version. We therefore assume, in the following, that whenever we speak of a model of BG's approach, we mean a model according to the second version of the definition of modelhood – this second version is the definition given in Appendix 10.8.1.

- The third problem is that BG use *names* for situations. In order to create corresponding theories, we need names for our time points, too. The way we define of our language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F})$ does not allow for this. We therefore have to extend Definition 9.24 of page 230 in the following straightforward way:

  Definition 10.8 *A first-order causal theory* $T_{\mathrm{C}} = \langle \mathrm{EQ}, \mathrm{CONS} \rangle$ *for a language* $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F}, \mathbf{T})$ *(where* $\mathbf{T}$ *is a finite set of constants) is a first-order causal theory with persistence, dependent fluents and concurrent events (i.e. a theory according to Definition 9.24, page 230) for the tuple* $\langle \mathbf{D}, \mathbf{E}, \mathbf{F} \rangle$ *extended with the constants in* $\mathbf{T}$*). These constants are all of sort time points.*

  The definition of models for causal theories (Definition 9.16, page 226) ensures that the constants in $\mathbf{T}$, which we will call *time names*, will always be interpreted as nonnegative integers. Whenever in the following we speak of T_{C}, we mean a theory T_{C} defined according to Definition 10.8 above.

Having taken care of these problems, we are ready to define syntactic correspondence. This is done in Definition 10.9 on page 250. As an example of how that definition works, we will give a simple domain description and a causal theory that corresponds to it. For this, let D be the domain description consisting of (10.10), (10.11),(10.13) and (10.14). This is just the soup-bowl domain we considered before but without the malfunctioning specificity axiom (10.12). According to Definition 10.9, this domain is equivalent to a causal theory T_{C} with an EQ containing the standard two intervention axioms and our persistence and no-events axiom (9.47), and a CONS containing UNA-

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

Definition 10.9 *For any theory T_{C} for a language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F}, \mathbf{T})$ and domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$, we define $T_{\mathrm{C}} \sim D$ ('T_{C} corresponds to D') to hold iff all of the following hold:*

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

2. **general axioms** EQ *is as required by Definition 10.8.* CONS *contains UNA- and DC- axioms for $\mathbf{E}, \mathbf{F}$ and $\mathbf{B}$ and the additional axiom $s_0 = 0$.*

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

4. **effect laws** *A sentence Γ_{C} corresponds to an effect law L in D iff Γ_{C} is of the form*
$$\forall t.[\, Ho(F_{j_1}^{b_1}, t) \wedge \ldots \wedge Ho(F_{j_m}^{b_m}, t) \;\wedge Ho(a^1, t) \wedge \ldots \wedge Ho(a^n, t)\,] \supset Do(F_i, b, t+1) \quad (10.17)$$
while L is of the form: $$\{a^1, \ldots, a^n\} \textbf{ causes } F_i^b \textbf{ if } F_{j_1}^{b_1}, \ldots, F_{j_m}^{b_m} \quad (10.18)$$

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

6. **occurrence facts** *A sentence Γ_{C} corresponds to an occurrence fact O in D iff O is of the form '$[a_1, \ldots, a_n]$* **occurs_at** *s', where $n > 0$, $a_i = a_{i1} \mid \ldots \mid a_{im_i}$, $a_{ij} = \{a_{ij}^1, \ldots, a_{ij}^{k(i,j)}\}$, $a_{ij}^{k(i,j)} \in \mathcal{A}$, while Γ_{C} is of the form:*
$$\begin{array}{r}
[\, \mathbf{H}(a_{11}, s) \vee \;\ldots\; \vee \mathbf{H}(a_{1m_1}, s)\,] \;\wedge \\
[\, \mathbf{H}(a_{21}, s+1) \vee \;\ldots\; \vee \mathbf{H}(a_{2m_2}, s+1)\,] \;\wedge \\
\vdots \\
\wedge\, [\, \mathbf{H}(a_{n1}, s+n-1) \vee \;\ldots\; \vee \mathbf{H}(a_{nm_n}, s+n-1)\,]
\end{array} \quad (10.19)$$
where $\mathbf{H}(a_{ij}, s)$ is short for $Ho(a_{ij}^1, s) \wedge \ldots \wedge Ho(a_{ij}^{k(i,j)}, s)$.

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

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

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

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

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

Definition 10.10 *For any T_{C} for a language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F}, \mathbf{T})$ and any non-ambiguous domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$ such that $T_{\mathrm{C}} \sim D$, let $M = (\Psi, \Sigma)$ be any interpretation of $(\mathcal{F}, \mathcal{A}, \mathcal{S})$ and $\mathcal{M}$ be any interpretation of our language. We say that M corresponds to $\mathcal{M}$ (written as '$M \cong \mathcal{M}$') iff (a) $\mathcal{M}$ interprets all time points as integers and '$+$' and '$<$' accordingly; and (b) for all $f \in \mathcal{F}, a \in \mathcal{A}$ and $s \in \mathcal{S}$ and all $t \in \mathbf{N}_0$:*

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

and DC-axioms and on top of that the axioms:

$$\begin{gathered} s_0 = 0 \\ Ho(Lift\_left, t) \supset Do(Wet, \text{TRUE}, t+1) \\ Ho(Lift\_right, t) \supset Do(Wet, \text{TRUE}, t+1) \\ Ho(Turn\_on\_sprinkler, t) \supset Do(Wet, \text{TRUE}, t+1) \\ (s_0 < s_N) \wedge \neg Ho(Wet, s_0) \wedge \\ [\ Ho(Turn\_on\_sprinkler, s_0) \wedge Ho(Lift\_left, s_0) \wedge Ho(Lift\_right, s_0)\ ] \end{gathered} \tag{10.20}$$

Here the first axiom is introduced by item 2 of the definition; the second, third and fourth axioms are introduced by items 3 and 4 of the definition and the fifth axiom is introduced by item 8 of the definition. Item 6 of definition 10.9 deserves special attention: it translates any rule in D of the form '$[a_1, \ldots, a_n]$ **occurs_at** s' into an axiom in CONS which expresses that a_1 should hold at time point s, a_2 at time point $s+1$ etc. In other words, if, in BG's formalism, several actions $[a_1, \ldots, a_n]$ take place sequentially in a situation s, we interpret this as saying, in our formalism, that the first of them happens at the point in time t corresponding to s, the second to the point in time directly thereafter etc. Any situation s' such that s precedes s' in BG's formalism will therefore be mapped to a point in time t' that is sufficiently larger than t so as to allow for all the actions $[a_1, \ldots, a_n]$ to happen in between t and t'.

The semantical correspondence relation '$\cong$' is introduced in Definition 10.10 on page 251. Just above that definition, we introduce the notation a_t to stand for the $(t+1)$-th compound action taking place in an interpretation M. Similarly, σ_t stands for the state of the world (i.e. the set of all fluents that hold) just before the $(t+1)$-th compound action takes place. In the definition itself, it is checked whether for any set of actions denoted by a_t, the same actions take place in $\mathcal{M}$ *at time* t; σ_t is treated similarly.

Having discussed both syntactic and semantical correspondence, we are now ready to state our theorem. The proof can be found in Appendix 10.8.3.

Theorem 10.11 *For any theory* T_{C} *for a language* $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F}, \mathbf{T})$ *and any domain description* D *for a language* $\mathcal{L}_3$ *of Baral and Gelfond's such that a)* $T_{\mathrm{C}} \sim D$ *and b)* D *is not* ambiguous*, we have:*

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

and we further have

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

10.5 A Brief Note on Other Approaches based on Pearl's ideas

There have been several earlier papers on common-sense reasoning about action that were (partially) based on Pearl's ideas; we mention the papers of Darwiche and Pearl [33, 34], Boutilier and Goldszmidt [19] and Geffner [52]. However, in the works of Darwiche, Pearl, Boutilier and Goldszmidt actions are not directly treated as interventions, but rather compiled into nodes in a causal graph (in Darwiche and Pearl's case) or a Bayesian Network (in Boutilier and Goldszmidt's case). The resulting theories are quite different from ours in that they lack the *Do*-operator, which is fundamental to our solution of the ramification problem. Only the approach of Geffner is able to handle comparable instances of the ramification problem. We plan a more in-depth comparison to Geffner's approach in future work.

Earlier, the present author has introduced the two model selection criteria $\mathbf{S}_0$ and $\mathbf{I}_0$ [62, 63] which were claimed to be based on Pearl's causal graphs. However, the connection was not investigated in detail. The work reported here can be seen as an extension of this original work, but now with the connection worked out in full detail (see the appendix to Chapter 9). Putting the theory in a form that makes as explicit as possible the connection to Pearl's work has caused the causal theories discussed in this thesis to look – superficially – quite different from $\mathbf{S}_0$ and $\mathbf{I}_0$.

10.6 Conclusion

In this chapter we have compared our causal theories to existing (causal) approaches to NMTR. We have seen that our theories permit the same inferences as these existing approaches for large classes of reasoning domains, while showing (in the case of McCain and Turner and Baral, Gelfond and Provetti) that our approach sometimes gives better results on domains falling outside these classes. In this way, the two goals we set ourselves at the beginning of the chapter have been achieved: we provided evidence that our approach 'truely works well', and we gained insight in how the treatment of causality in existing approaches is related to Pearl's 'sufficient cause principle': both the counterexamples we referred to above (given by proposition 10.1 and 10.6, respectively) can be interpreted as stemming from not following this principle. In contrast, Lin's persistence and causation axioms (axioms (10.6)-(10.8)) are almost equivalent to our persistence and intervention axioms, which (as explained in the previous chapter, Section 9.6.1) are directly based on the sufficient cause principle.

Future Work In future work, we would like to make a more extensive comparison between our approach and that of Thielscher [151]. Apart from Pearl, Thielscher is one of the few authors in the field who tries to make completely clear what he means by a causal law, i.e. he attempts to give a semantics to statements of the form 'A causes B' in non-causal terms. Interestingly, Thielscher's is also the only approach we are aware of that can correctly handle reasoning domains for which our approach fails ([151], Example 18). On the other hand, our use of Do is much less restrictive than Thielscher's use of `causes`. We plan to study the exact differences between Thielscher's and our approach in the near future.

Another approach that deserves further attention is the one based on Sandewall's and Doherty's *occlusion* concept [135, 134, 66]. The underlying idea behind occlusion is similar to Pearl's ideas about interventions: in Sandewall, Gustafsson and Doherty's work, if a fluent is *affected* by some action, then it becomes *occluded* for the duration of the action, which means that its value becomes independent of the value it had directly before the action started. Compare this to the sufficient cause principle: if an action takes place that *sets* the value of some fluent, then the fluent becomes independent of the values of any variables in the domain which normally influence it. We see that the two concepts are closer than their names suggest and we think that they should be compared in detail.

This brings us once again to what may be the most interesting aspect of our approach: it partially bridges the conceptual gaps that exist between different paradigms in the field of reasoning about action. First, within the logical (nonmonotonic reasoning) paradigm, it relates the causal approaches to the non-causal approaches. Second and more generally, it establishes a connection between formalizations of causal knowledge that have been developed within this logical paradigm (such as Lin's and McCain and Turner's) and those that have their roots in the probabilistic/ Bayesian network tradition (i.e. Pearl's causal networks).

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

10.7.1 Formal Definition of MT's Next-State Function

The definitions in this section have all been copied from [105, 104]. MT start with a propositional language which includes the 0-ary logical connectives TRUE and FALSE. TRUE and ¬FALSE are tautologies in which no atoms occur. Each interpretation is identified with the set of literals true in it. A *causal law* is defined to be an expression of the from $\phi \Rightarrow \psi$ where ϕ and ψ are formulas of the propositional language. A set of causal laws is called a *causal theory.* Now for every causal theory D and interpretation I, we let

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

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

Definition 10.12 *Let D be a causal theory and I be an interpretation. We say that I is* causally explained *according to D if I is the unique model of D^I.*

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

Definition 10.13 *For any interpretation S, set of propositional formulas E and set of causal laws C, $\Pi_1(S,E,C)$ is defined to be the set consisting of all states that are causally explained according to the causal theory*

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

We will repeatedly use the following lemma which gives a precise characterization of $\Pi_1(S,E,C)$. It was first presented by McCain & Turner [105] but without proof. The proof however is not difficult; we provide one elsewhere.

Lemma 10.14 (McCain & Turner) *For any interpretation S, set of propositional formulas E and set of causal laws C, a state S' belongs to $\Pi_1(S,E,C)$ if and only if*

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

The proof can be found in [64].

10.7.2 Proof of Proposition 10.1

Take S, E and C as in the statement of the proposition. Let S' be any member of $\Pi_1(S,E,C)$. By Lemma 10.14 in Appendix 10.7.1 it follows that $S' \models E$ and hence $X \in S'$. This means $C^{S'}$ must contain $Y \vee Z$. Applying Lemma 10.14 again, we find $S' \models Y \vee Z$. So there are three possibilities left for S': $S' \in \{S_1, S_2, S_3\}$ where $S_1 = \{X, Y, Z\}$, $S_2 = \{X, Y, \neg Z\}$, $S_1 = \{X, \neg Y, Z\}$. Now the second part of Lemma 10.14 does not hold for $S' = S_1$ while it does hold for $S' = S_2$ and $S' = S_3$. The first part of the lemma also holds for $S' = S_2$ and $S' = S_3$, so $\Pi_1(S,E,C) = \{S_2, S_3\}$.

10.7.3 Proof of Theorem 10.4

Our proof makes repeated use of Lemma 10.14 of Appendix 10.7.1. We first prove the first part of Theorem 10.4, which is restated below. We use the notation introduced in Definition 10.2, i.e. we will sometimes write F_i^{TRUE} to denote F_i and F_i^{FALSE} to denote $\neg F_i$.

Theorem 10.4 (part I) *For any theory T_{C} and any domain description $T_{\mathrm{MT}} = (S, E, C)$ such that $T_{\mathrm{C}} \sim T_{\mathrm{MT}}$, we have:*

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

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

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

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

It is clear that a model $\mathcal{M}$ satisfying all the conditions above can indeed be constructed. It follows immediately from the definition of '$\cong$' (Definition 10.3) and item 1 in the construction of $\mathcal{M}$ that $\mathcal{M} \cong (S, S')$. We now prove that $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$ in three stages. In stage 1, we show $\mathcal{M} \models$ CONS. We use this to show in stage 2 that $\mathcal{M}$ is a minimal model for CONS of $\mathcal{A}(\mathbf{V})$ within context $\mathcal{M}\rfloor_{\mathbf{V} \cup \mathbf{U}}$. This shows that $\mathcal{M}$ satisfies condition (1) of Definition 9.7. In stage 3, we show that $\mathcal{M}$ also satisfies condition (2) of that definition.

Stage 1 We have to show that $\mathcal{M} \models \phi$ for all axioms ϕ contained in CONS. From Definition 10.2, item 2, it follows that it is sufficient to show that $\mathcal{M} \models \phi$ for all sentences in CONS that correspond to either S or to the sentences in E and C. By items 3 and 4 of the same definition and the construction of $\mathcal{M}$ above, it follows immediately that indeed $\mathcal{M} \models \phi$ for the sentences ϕ that correspond to S and E.

Concerning C, let Γ be any sentence in CONS of the form (10.4). We will show $\mathcal{M} \models \Gamma$. Let us write $\Gamma \equiv \Phi \supset \Psi$ with Ψ of the form $Do(F_{i_1}(1), b_1) \wedge \ldots \wedge Do(F_{i_m}(1), b_m)$. If $\mathcal{M} \not\models \Phi$ we are done, so let us suppose $\mathcal{M} \models \Phi$. By construction of $\mathcal{M}$, we have $S' \models \Phi'$ where Φ' is obtained from Φ by replacing all occurrences of $F_i(1)$ in Φ by F_i. By item 5 of Definition 10.2, C must contain a rule of form $\Phi' \Rightarrow F_{i_1}^{b_1} \wedge \ldots \wedge F_{i_m}^{b_m}$ and, since $S' \models \Phi'$, we conclude that $F_{i_j}^{b_j} \in C^{S'}$ for all $1 \leq j \leq m$. Hence by item 2 of the construction of $\mathcal{M}$, $\mathcal{M} \models \Psi$ and hence $\mathcal{M} \models \Gamma$.

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

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

For this, let $\mathcal{M}' \in \mathbf{M}(\mathcal{M})$ be a model for CONS. Suppose, by way of contradiction, that $\{\phi \in \mathcal{A}(\mathbf{V}) \mid \mathcal{M}' \models \phi\}$ is a proper subset of $\{\phi \in \mathcal{A}(\mathbf{V}) \mid \mathcal{M} \models \phi\}$. Then there is a pair (F_i, b) such that $\mathcal{M}' \not\models Do(F_i, b)$ while $\mathcal{M} \models Do(F_i, b)$. $\mathcal{M}$ is constructed such that $F_i^b \in E \cup C^{S'}$. Now if $F_i^b \in E$, then, by Definition 10.2, item 4, $\mathcal{M}' \not\models$ CONS and we arrive at a contradiction. So suppose $F_i^b \in C^{S'}$. It follows from the definition of $C^{S'}$ that $S' \models \Phi'$ for a Φ' such that C contains some sentence of form $\Phi' \Rightarrow \Psi'$ and $F_i^b \in \Psi'$. Hence by Definition 10.2, item 5, CONS contains the sentence $\Phi \supset \Psi$ where Φ, Ψ stand to Φ', Ψ' as required in that item of the definition. Since $S' \models \Phi'$, by the construction of $\mathcal{M}$, we have $\mathcal{M} \models \Phi$. Since $\mathcal{M}'$ and $\mathcal{M}$ share the same interpretations of the variables in $\mathbf{V}$, we also have $\mathcal{M}' \models \Phi$. We assumed $\mathcal{M}' \models$ CONS so also $\mathcal{M}' \models \Psi$ and in particular $\mathcal{M}' \models Do(F_i, b)$. But we assumed the contrary so we have arrived at a contradiction.

Stage 3 We have to show that $\mathcal{M} \models$ EQ$'$ where EQ$'$ is the updated set of structural equations referred to in condition (2) of Definition 9.7. By construction of $\mathcal{M}$, there are two cases for each F_i: either EQ$'$ contains $F_i(1) \equiv b$ for some b or EQ$'$ contains $F_i(1) \equiv F_i(0)$.

The first case happens if $\mathcal{M} \models Do(F_i(1), b)$. By construction of $\mathcal{M}$, it follows that in this case $F_i^b \in E \cup C^{S'}$. By Lemma 10.14 and the fact that $S' \in \Pi_1(S, E, C)$ we have $S' \models E \cup C^{S'}$ and thus $S' \models F_i^b$. Then by construction of $\mathcal{M}$, indeed $\mathcal{M} \models F_i(1) \equiv b$ and we are done.

The second case happens if $\mathcal{M} \not\models Do(F_i, b)$ for any $b \in \mathbf{B}$. It follows from the construction of $\mathcal{M}$ that in this case there is no b such that $F_i^b \in E \cup C^{S'}$. Suppose, by way of contradiction, that $\mathcal{M} \not\models F_i(1) \equiv F_i(0)$. Then (by construction of $\mathcal{M}$), $F_i^b \in S' \setminus S$ for some b. By Lemma 10.14, $(S \cap S') \cup E \cup C^{S'} \models F_i^b$. Clearly, $F_i^b \notin S \cap S'$. We have already shown that F_i^b cannot be in $E \cup C^{S'}$ either. Since both $S \cap S'$ and $E \cup C^{S'}$ only contain conjunctions of literals, it follows that $(S \cap S') \cup E \cup C^{S'} \not\models F_i^b$ and we have reached a contradiction. □

We now restate and prove the second part of Theorem 10.4.

Theorem 10.4 (part II) *If we have a theory* T_{C} *and a* $T_{\mathrm{MT}} = (S, E, C)$ *such that* $T_{\mathrm{C}} \sim T_{\mathrm{MT}}$*, then*

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

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

Since $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$ and thus $\mathcal{M} \models$ CONS we know by items 2 and 3 of Definition 10.2 that for all F_i: $F_i \in S \Leftrightarrow \mathcal{M} \models F_i(0)$. From this and the definition of S' above it follows $\mathcal{M} \cong (S, S')$. It remains to be shown that $S' \in \Pi_1(S, E, C)$. By Lemma 10.14 it is sufficient to prove claims 1 and 2 below.

Claim 1 $S' \models E \cup C^{S'}$.
To prove this, we show that for all F_i^b with $F_i^b \in E$ and for all F_i^b with $F_i^b \in C^{S'}$ we have $S' \models F_i^b$. In the first case ($F_i^b \in E$), as $\mathcal{M} \models$ CONS and by Definition 10.2, item 4, we have $\mathcal{M} \models Do(F_i(1), b)$. Since $\mathcal{M} \models$ EQ$'$ this means $\mathcal{M} \models F_i^b(1)$. From the definition of S', we now have $S' \models F_i^b$ and we are done. In the second case, i.e. $F_i^b \in C^{S'}$, C must contain a law of form $\Phi' \Rightarrow \Psi'$ such that $S' \models \Phi'$ and one of the conjuncts in Ψ' is F_i^b. By definition of S' and Definition 10.2, item 5, we also have $\mathcal{M} \models \Phi$ where $\Phi \supset \Psi$ is the sentence in CONS that corresponds to $\Phi' \supset \Psi'$. As $\mathcal{M} \models$ CONS, we have $\mathcal{M} \models \Psi$ and in particular $\mathcal{M} \models Do(F_i(1), b)$. As $\mathcal{M} \models$ EQ$'$, it follows $\mathcal{M} \models F_i^b(1)$ and hence, by definition of S', $S' \models F_i^b$.

Claim 2 $S' \setminus S \subseteq \{\phi : (S \cap S') \cup E \cup C^{S'} \models \phi\}$.
To prove this, we show that for any F_i^b, if $F_i^b \in S' \setminus S$, then we must also have $F_i^b \in (S \cap S') \cup E \cup C^{S'}$. So suppose $F_i^b \in S' \setminus S$. By definition of S', we have for such an F_i^b that $\mathcal{M} \models F_i^b(1) \wedge \neg F_i^b(0)$. Since $\mathcal{M} \models$ EQ$'$, the structural equation $F_i(1) \equiv F_i(0)$ cannot be in EQ$'$ and thus there must have been an intervention: we must have $\mathcal{M} \models Do(F_i(1), b)$. Now as $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$, there can be no other $\mathcal{M}' \models$ CONS with the same interpretations of $\mathbf{V}$ and $\mathbf{U}$ but with a smaller (in the subset sense) interpretation of the 'Do' variables. This means that CONS must contain an axiom somehow mentioning $Do(F_i(1), b)$. From Definition 10.2 we see there are two possibilities: The first possibility is that CONS contains the axiom $Do(F_i(1), b)$, i.e. an axiom of the form given in item 4 of Definition 10.2. In this case, it follows $F_i^b \in E$ so $F_i^b \in (S \cap S') \cup E \cup C^{S'}$ and we are done.

The second possibility is that CONS contains an axiom of form (10.4) while $\mathcal{M} \models \Phi$ where Φ is as in (10.4) and one of the conjuncts on the right-hand side of the axiom is $Do(F_i(1), b)$. In that case, C contains the corresponding rule $\Phi' \Rightarrow \Psi'$ with $F_i^b \in \Psi'$ and, by definition of S', $\Phi' \in S'$. Therefore, $F_i^b \in C^{S'}$ and hence $F_i^b \in (S \cap S') \cup E \cup C^{S'}$ and we are done. □

10.8 Appendix: $\mathcal{L}_3$ and First-Order Causal Theories

10.8.1 Formal Definition of Modelhood

The definitions in this section have all been copied from [11]. In order to formally define the notion of 'model', we first need to introduce a menagerie of other concepts. For detailed explanation of these concepts and clarifying examples, we refer to [11]. Roughly, the definition of 'model' for BG's domain descriptions D (Definition 10.19) makes use of two other, more basic notions: the notion of a *causal model* and the notion of 'a fact being true in an interpretation M'. Both concepts will be introduced below.

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

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

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

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

A fluent f is an *inherited effect* of (executing) a in σ if there is $b \subset a$ such that (1) f is an immediate effect of b in σ; (2) there is no action c such that $b \subset c \subseteq a$ and $\neg f$ is an immediate effect of c in σ.

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

Now let

$$E_a^+(\sigma) = \{f : f \text{ is an effect of } a \text{ in } \sigma\} \ ; \ E_a^-(\sigma) = \{f : \neg f \text{ is an effect of } a \text{ in } \sigma\}$$

and $Res(a, \sigma) = \sigma \cup E_a^+(\sigma) \setminus E_a^-(\sigma)$.

Definition 10.15 *A causal interpretation Ψ* satisfies *effect laws of D if for any sequence $\alpha \circ a$ from the language of D:*

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

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

Truth of Facts We already defined a *situation assignment* on page 246. To introduce the notion of 'truth of facts in an interpretation', we first need to make our definition of 'interpretation' (also page 246) more precise:

Definition 10.16 *For any domain description D and any causal interpretation Ψ that is a causal model of D, an* interpretation *M of $\mathcal{L}_3$ is a pair (Ψ, Σ) where Ψ is a causal model of D, Σ is a situation assignment of $\mathcal{S}$ and $\Sigma(s_N)$ belongs to the domain of Ψ.*

Definition 10.17 *For any interpretation $M = (\Psi, \Sigma)$,*

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

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

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

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

Definition 10.18 *Let $\alpha = [a_1, \ldots, a_n]$ and $\beta = [b_1, \ldots, b_m]$ be sequences of actions. We say that $\alpha \leq \beta$ if there exists a subsequence*[5] *$[b_{i_1}, \ldots, b_{i_n}]$ of β such that for every $a_k \in \alpha, a_k \subseteq b_{i_k}$.*

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

Definition 10.19 [11] *An interpretation $M = (\Psi, \Sigma)$ is called a* model *of a domain description D in $\mathcal{L}_3$ if the following conditions are satisfied:*

C1. *Ψ is a causal model of D*

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

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

The definition given above is slightly different from the one given in [11, 12]. This is because we opt for the version of BG's approach in which the models with the least number of actions are always preferred (see page 249). At page 10 of [12], it is claimed that in this case, the definition should be with **C1** and **C2** as above, but **C3** replaced by

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

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

10.8.2 Proof of Proposition 10.6

We abbreviate all fluent and action names in the obvious way.

We have $\mathcal{F} = \{W\}, \mathcal{A} = \{LL, LR, T\}, \mathcal{S} = \{s_0, s_N\}$. Suppose, by means of contradiction, that a model $M = (\Psi, \Sigma)$ of D exists. We then have by Definition 10.19, **C2** and (10.14) that $\neg W$ **at** s_0 must hold in M. By Definition 10.17, item 2 it follows that $\Psi([]) = \emptyset$ (i.e. $\neg W$ is true in $\Psi([])$). Also, Ψ must be a causal model of D. It is straightforward to verify that for $a = \{T, LL, LR\}$ we have $E_a^+(\Psi(\alpha)) \cap E_a^-(\Psi(\alpha)) = \{W\}$. Definition 10.15 now tells us that $\Psi([\{T, LL, LR\}])$ is undefined. Since by Definition 10.17, item 2, we have for any model $M = (\Psi, \Sigma)$ of D that $[\{T, LL, LR\}]$ is a prefix of $\Sigma(s_N)$, it follows that $\Psi(\Sigma(s_N))$ is undefined too. But then, by Definition 10.16, M is not an interpretation of D. Therefore it is not a model of D either and we have reached the desired contradiction. □

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

10.8.3 Proof of Theorem 10.11

The proof of Theorem 10.11 is necessarily quite involved, the reason being that both our definition of 'preferred models' and BG's definition of models contain a self-reference: a model $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$ is preferred if there is *no other* model $\mathcal{M}' \models_{\mathbf{c}} T_{\mathrm{C}}$ with a smaller interpretation of Ab_1. Similarly, M is a model of D if there is *no other* model satisfying conditions **C1** and **C2** of Definition 10.19 in which fewer actions happen. For this reason, we will first show that for corresponding M and $\mathcal{M}$, we have that the same fluents and events hold at the same time (Lemma 10.20). We use this result to show that for corresponding M and $\mathcal{M}$, we have $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$ iff conditions **C1** and **C2** hold for M (this is essentially what happens in lemmas 10.22, 10.23 and 10.24). Only then will we be in a position to prove the theorem itself. We start by proving the four lemmas mentioned.

Correspondence of Facts

In the following, whenever we mention T_{C}, we mean a causal theory T_{C} defined for a language $\mathcal{L}(\mathbf{D}, \mathbf{E}, \mathbf{F}, \mathbf{T})$ according to Definition 10.8.

Lemma 10.20 *For any T_{C} for an instance of our language and any non-ambiguous domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$ such that $T_{\mathrm{C}} \sim D$, let $M = (\Psi, \Sigma)$ be any interpretation of $(\mathcal{F}, \mathcal{A}, \mathcal{S})$ and $\mathcal{M}$ be any interpretation of our language such that $M \cong \mathcal{M}$. Let F be any fact in D and ϕ be any axiom in* CONS *such that F and ϕ are corresponding facts (where correspondence is defined as in Definition 10.9, items 5-8). We have:*

$$F \textit{ is true in } M \quad \Leftrightarrow \mathcal{M} \models \phi \tag{10.21}$$

Proof of Lemma 10.20: The fact F mentioned in the lemma can either be atomic or non-atomic. We first consider atomic facts. They can be of three kinds: fluent facts, occurrence facts and precedence facts. We only give the proof for fluent facts. The proofs for the other two cases follow the same scheme as the one for fluent facts so we omit them.

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

So suppose F is true in M. It follows that f is true in $\Psi(\Sigma(s))$, so $f \in \sigma_t$ for the t that is equal to the length of the sequence $\Sigma(s)$. Now since $M \cong \mathcal{M}$, it follows from Definition 10.10 that $\mathcal{M} \models Ho(f, t)$ and $\mathcal{M} \models s = t$, so $\mathcal{M} \models Ho(f, s)$, so $\mathcal{M} \models \phi$.

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

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

Causal Consistency and its Lemma's

Definition 10.21 (Causally Consistent) *For any T_{C} for an instance of our language and any non-ambiguous domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$ such that $T_{\mathrm{C}} \sim D$, let $M = (\Psi, \Sigma)$ be any interpretation of $(\mathcal{F}, \mathcal{A}, \mathcal{S})$ and $\mathcal{M}$ be any interpretation of our language. We say that M and $\mathcal{M}$ are* causally consistent *iff*

1. $M \cong \mathcal{M}$.
2. $\mathcal{M} \models \forall e, b, t.\ \neg Do(e, b, t)$
3. *For all F_i^b: a) $\mathcal{M} \models \neg Do(F_i, b, 0)$. b) for all $t \in \{1, 2, \ldots\}$: $\mathcal{M} \models Do(F_i, b, t)$ iff* CONS *contains an axiom ϕ of form (10.17) such that (1) the F_i and b mentioned in the right-hand side of ϕ are equal to the F_i and b mentioned here and (2) the left-hand side of ϕ holds in $\mathcal{M}$.*
4. $\mathcal{M} \models \forall e, t\ .\ Ab_1(e, t) \equiv Ho(e, t)$.
5. $|\mathcal{M}|_b = \mathbf{B}; |\mathcal{M}|_e = \mathbf{E}; |\mathcal{M}|_f = \mathbf{F}$*; $\mathcal{M}$ interprets all constants in $\mathbf{B}$, $\mathbf{E}$ and $\mathbf{F}$ as themselves.*
6. *Ψ is a causal model of D.*

Lemma 10.22 *For any T_{C} for an instance of our language and any non-ambiguous domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$ such that $T_{\mathrm{C}} \sim D$, let $M = (\Psi, \Sigma)$ be any interpretation of $(\mathcal{F}, \mathcal{A}, \mathcal{S})$ and $\mathcal{M}$ be any interpretation of our language such that $M \cong \mathcal{M}$. We have:*

$$\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}} \Rightarrow \textit{All facts of } D \textit{ are true in } M \tag{10.22}$$

and furthermore

$$\begin{gathered} \textit{All facts of } D \textit{ are true in } M \ \&\ M \textit{ and } \mathcal{M} \textit{ are causally consistent} \\ \Rightarrow \\ \mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}} \end{gathered} \tag{10.23}$$

Lemma 10.23 *For any T_{C} for an instance of our language and any domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$ such that $T_{\mathrm{C}} \sim D$, let $M = (\Psi, \Sigma)$ be any interpretation of D such that M satisfies conditions* **C1** *and* **C2** *of Definition 10.19. Then there exists an interpretation $\mathcal{M}$ of our language such that M and $\mathcal{M}$ are causally consistent.*

Lemma 10.24 *For any T_{C} for an instance of our language and any non-ambiguous domain description D for a language $\mathcal{L}_3 = (\mathcal{F}, \mathcal{A}, \mathcal{S})$ such that $T_{\mathrm{C}} \sim D$, let $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$. Then there exists an interpretation $M = (\Psi, \Sigma)$ of $\mathcal{L}_3$ such that M and $\mathcal{M}$ are causally consistent.*

Proofs of lemmas 10.22-10.24

Proof of Lemma 10.22 (first part) We prove the contrapositive of (10.22). Suppose not all facts of D hold in M. By Lemma 10.20, the fact that $M \cong \mathcal{M}$ and the definition of corresponding theories (Definition 10.9) it follows that $\mathcal{M} \not\models$ CONS and therefore $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$. □

Proof of Lemma 10.22 (second part) Suppose we have an M and $\mathcal{M}$ such that $M \cong \mathcal{M}$ and (10.23) holds. We show in three stages that $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$: in stage 1, we show $\mathcal{M} \models$ CONS. We use this to show in stage 2 that $\mathcal{M} \models$ *Circum*(CONS; *Do*). In stage 3, we show that $\mathcal{M} \models$ EQ. Stage 1–3 together show that $\mathcal{M}$ satisfies condition (1) of the definition of models for causal theories (Definition 9.16, page 226). That $\mathcal{M}$ satisfies condition (2) of that definition follows already from the definition of '$\cong$' (Definition 10.10, item (a)).

Stage 1 We show that $\mathcal{M} \models \phi$ for all axioms ϕ contained in CONS. From Definition 10.9, item 2, it follows that CONS contains UNA- and DC-axioms for $\mathbf{B}$, $\mathbf{E}$ and $\mathbf{F}$. It follows directly from the fact that M and $\mathcal{M}$ are causally consistent that these axioms hold for $\mathcal{M}$ too. We now consider the axioms of form (10.17) in CONS (Definition 10.9 item 4). Let us denote by ϕ any of these axioms. ϕ can be written as $\forall t.\mathbf{H}(t) \supset Do(F_i^b, t+1)$ where $\mathbf{H}$ is some conjunction of instances of *Ho*. We now show that for all $t \geq 0$, $\mathcal{M} \models \mathbf{H}(t) \supset Do(F_i, b, t+1)$. Since we assume that *time points* are interpreted as the nonnegative integers, this is sufficient to show $\mathcal{M} \models \phi$. So take any $t \geq 0$. For this t, either $\mathcal{M} \not\models \mathbf{H}(t)$ and we are done. If $\mathcal{M} \models \mathbf{H}(t)$ then by the definition of causal consistency, we have $\mathcal{M} \models Do(F_i, b, t+1)$ and we are done again.

It follows from Lemma 10.20 and the fact that all facts of D are true in M that $\mathcal{M} \models \phi$ for all axioms ϕ in T_{C} mentioned in items 5–8 of Definition 10.9 (all these axioms correspond to some fact in D). From item 3 of Definition 10.9, we conclude that CONS contains no more axioms than those for which we already checked that they hold in $\mathcal{M}$. This completes stage 1.

Stage 2 We show that $\mathcal{M} \models$ *Circum*(CONS; *Do*). Notice first that by Proposition 9.17 all models share the same universes for CONS. From this and the characterization of circumscription given in Proposition 9.29 it follows that it is sufficient to show that among all the models for CONS with the same interpretation of *Ho*, there is no $\mathcal{M}'$ whose interpretation of *Do* is a proper subset of that of $\mathcal{M}$. Suppose, by means of contradiction, that such an $\mathcal{M}'$ exists, i.e. $\mathcal{M}' \models$ CONS, $\mathcal{M}[Ho] = \mathcal{M}'[Ho]$ but $\mathcal{M}'[Do] \subsetneq \mathcal{M}[Do]$. By the definition of causal consistency, item 2, $\mathcal{M}$ and therefore also $\mathcal{M}'$ have $\neg Do(e, b, t)$ for any e, b, t. It follows there is a pair $(F_{i^*}^{b^*}, t^*)$ with $\mathcal{M}' \models \neg Do(F_{i^*}, b^*, t^* + 1)$ while $\mathcal{M} \models Do(F_{i^*}, b^*, t^* + 1)$. It follows from the definition of causal consistency, item 3 that T_{C} contains an axiom ϕ of form (10.17) where (1) the F_i,b and t mentioned in the right-hand side of ϕ are equal to F_{i^*},b^* and t^* and (2) the left-hand side of ϕ holds in $\mathcal{M}$, and therefore also in $\mathcal{M}'$. But also $\mathcal{M}' \models$ CONS so $\mathcal{M}' \models Do(F_{i^*}, b^*, t^*)$ and we arrive at a contradiction.

Stage 3 We show that $\mathcal{M} \models$ EQ by showing that axioms (9.33) and (9.34), (9.36) and (9.47) hold in $\mathcal{M}$. By items 2 and 4 of the definition of causal consistency and the fact that M and $\mathcal{M}$ are causally consistent it follows that (9.47) holds in $\mathcal{M}$. Concerning (9.33) and (9.34), we must show that for all $x \in \mathbf{E} \cup \mathbf{F}, t \in \mathbf{N}_0$, if $\mathcal{M} \models Do(x, \text{TRUE}, t)$ then we also have $\mathcal{M} \models Ho(x, t)$ and similarly, if $\mathcal{M} \models Do(x, \text{FALSE}, t)$

then we also have $\mathcal{M} \models \neg Ho(x,t)$. Let us thus assume $\mathcal{M} \models Do(x,b,t)$ for some $x \in \mathbf{E} \cup \mathbf{F}$, $t \in \mathbf{N}_0$ and $b \in \mathbf{B}$. Since $\mathcal{M} \models \neg Do(e,b,t)$ for any e, b and t (see the definition of causal consistency), we may assume $x = F_i^b$ for some $F_i \in \mathbf{F}$. From the same definition, it follows $t \geq 1$ and that CONS must contain an axiom ϕ of form (10.17) such that the right-hand side of ϕ is equal to $Do(F_i,b,t)$ and the left-hand side holds in $\mathcal{M}$. It follows D contains an effect law of form (10.18), i.e. of form $\{a^1,\ldots,a^n\}$ **causes** F_i^b **if** $F_{j_1}^{b_1},\ldots,F_{j_m}^{b_m}$. Since $M = (\Psi,\Sigma)$ corresponds to $\mathcal{M}$ we have that $\{a^1,\ldots,a^n\} \subseteq a_{t-1}$ and $F_{j_1}^{b_1},\ldots,F_{j_m}^{b_m}$ all hold in σ_{t-1}. By the following claim we have that F_i^b is an effect of a_{t-1} in σ_{t-1}.

Sublemma 10.25 *Let σ be a state for a non-ambiguous domain description D. Let f be a fluent literal and a be a compound action. Now suppose there is $b \subseteq a$ such that f is an* immediate *effect of b in σ. Then f is an (ordinary) effect of a in σ.*

Proof: follows almost directly from the definitions of immediate effect, inherited effect (Section 10.8.1 and ambiguity (Definition 10.7) . We omit the details. □

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

Claim Let Ψ be a causal model and Σ be a situation assignment for some domain description D. If $\Psi(\Sigma(s_N))$ is defined, then for all $t \geq 0$, σ_t is defined.
Proof: is immediate from the definition of σ_t in Definition 10.10. □

Since σ_t is defined, it follows by Definition 10.15 that $\sigma_t = Res(a_{t-1},\sigma_{t-1})$. Since F_i^b is an effect of a_{t-1} in σ_{t-1}, this means F_i^b is contained in σ_t. Since $M \cong \mathcal{M}$, we have $\mathcal{M} \models Ho(F_i^b,t)$, which is what we had to prove.
We will now show that $\mathcal{M} \models$ (9.36). For this, suppose $\mathcal{M} \models \neg Ho(F_i^b,t-1) \wedge Ho(F_i^b,t)$ for some F_i^b and t. One can show $\mathcal{M} \models Do(F_i,b,t)$ in a manner analogous to (but simpler than) the reasoning in the first part of stage 3, above. We omit the details of this part of the proof. Since we can show this for *any* $t > 0$, *any* b and *any* F_i, it follows by Proposition 9.17 that the following axiom holds for $\mathcal{M}$:

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

which is clearly equivalent to (9.36). This ends the proof of Lemma 10.22. □

Proof of Lemma 10.23 M is an interpretation of (Ψ,Σ) so, by Definition 10.16, $\Psi(\Sigma(s_N))$ is defined; this means Definition 10.10 applies. Items 1 and 2 of that definition determine the interpretation of $Ho(x,t)$ for all $x \in \mathbf{E} \cup \mathbf{F}$ and all $t \geq 0$ in a way that clearly cannot lead to any contradiction. Also, it is clear that item 3 in the definition cannot lead to any contradictory assignment in $\mathcal{M}$ of time name *symbols* to time *points*. Hence an interpretation $\mathcal{M} \cong M$ exists. Now items 2 to 5 of the definition of causal consistency determine the interpretations of all atoms in $\mathcal{M}$ other than those of the form $Ho(x,t)$, again in a way that cannot give rise to any contradiction. □

Proof of Lemma 10.24 It follows immediately from the fact that $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$ and the form that CONS must have according to Definition 10.9, that conditions 2 and 3 of the definition of causal consistency hold for $\mathcal{M}$. Since condition 2 holds and $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$ and so, in particular, $\mathcal{M} \models \mathit{Circum}(\textsc{cons}; \mathit{Do})$, condition 4 holds too. By Proposition 9.17 condition 5 holds too. Hence it only remains to prove that an interpretation $M = (\Psi, \Sigma)$ exists with $M \cong \mathcal{M}$ such that Ψ is a causal model for D. Using Definition 10.10 we can construct such an M as follows:

1. We set $\Sigma(s_N) = [a_0, \ldots, a_{last}]$ where a_i is defined as in Definition 10.10, item 1. We define, for any $s \in \mathcal{S}$, $\Sigma(s)$ to be the prefix of $\Sigma(s_N)$ of length t for the t for which $\mathcal{M} \models s = t$. This completes the definition of Σ.
2. We now define Ψ: we set for all $t \geq 0$, $\Psi([a_0, \ldots, a_{t-1}]) = \sigma_t$ (we use the convention that $[a_{-1}] = []$, so $\Psi([]) = \sigma_0$), where the a_i's are defined as above and σ_t is defined as in Definition 10.10, item 2. Let us denote by the Σ-*sequences* those sequences that start with a prefix of $\Sigma(s_N)$ (note this includes the empty sequence). Now for *any* sequence of compound actions α that is not a Σ-sequence we define $\Psi(\alpha)$ to be such that it satisfies Definition 10.15. Note that this is always possible, since according to that definition $\Psi(\alpha)$ is allowed to be undefined for some α. It can be seen from the definition that, since $\Psi([])$ has already been defined above, this uniquely defines the function Ψ for any sequence that does not start with a prefix of $\Sigma(s_N)$. This completes the definition of Ψ.

For any sequence of compound actions except possibly the Σ-sequences, we know (from the way we defined Ψ) that Ψ satisfies all the effect laws of D. If we can prove that the Σ-sequences, too, satisfy all the effect laws of D, we have by Definition 10.15 that Ψ is a causal model of D and we are done.

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

Claim For any Σ-sequence $\alpha \circ a_t = [a_0, \ldots a_{t-1}, a_t]$, we have $E^+_{a_t}(\Psi(\alpha)) \cap E^-_{a_t}(\Psi(\alpha)) \neq \emptyset$.
Proof: Easy and omitted; see [64]. □

From the claim, it follows that we must prove that for any Σ-sequence $\alpha \circ a_t = [a_0, \ldots a_{t-1}, a_t]$, for any $t \geq 0$, we have $\Psi(\alpha \circ a_t) = \mathit{Res}(a_t, \Psi(\alpha))$. From the definition of σ_t and Definition 10.15, we know that this is equivalent to showing that for all $F_i \in \mathcal{F}$, for all $t \geq 0$,

$$F_i \in \sigma_{t+1} \Leftrightarrow F_i \in [\sigma_t \cup E^+_{a_t}(\sigma_t)] \setminus E^-_{a_t}(\sigma_t)] \qquad (10.24)$$

We will prove (10.24) using the following

Claim Let either $b = \textsc{true}$ and $\overline{b} = \textsc{false}$ or vice versa. For any $\mathcal{M}$ with $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$ and any t and F_i we have

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

Proof: trivial & omitted. □

We show the if-direction of (10.24). Suppose $F_i \in \sigma_{t+1}$. Then $\mathcal{M} \models Ho(F_i, t+1)$ and by the claim above, $\mathcal{M} \models \neg Do(F_i, \text{FALSE}, t+1)$. This means CONS contains no axiom of form (10.17) with right-hand side $Do(F_i, \text{FALSE}, t+1)$ and left-hand side such that it holds at time t. As $T_C \sim D$ and $M \cong \mathcal{M}$, there is no rule in D of form (10.18) such that $\{a^1, \ldots, a^n\} \subseteq a_t$ and $F_{j_1}^{b_1}, \ldots, F_{j_m}^{b_m}$ all hold in σ_t. This means $\neg F_i$ is not an effect of a_t in σ_t so $F_i \notin E^-_{a_t}(\sigma_t)$.

From our assumption $F_i \in \sigma_{t+1}$ it also follows that either $\mathcal{M} \models Ho(F_i, t)$ or $\mathcal{M} \models Do(F_i, \text{TRUE}, t+1)$. In the first case, $F_i \in \sigma_t$ and we are done; in the second case, since $\mathcal{M} \models_{\mathbf{c}} T_C$ and thus $\mathcal{M} \models \mathit{Circum}(\text{CONS}; Do)$, using Propositions 9.29 and 9.17, there are no models $\mathcal{M}'$ for CONS with $\mathcal{M}'[Ho] = \mathcal{M}[Ho]$ and $\mathcal{M}' \not\models Do(F_i, \text{TRUE}, t+1)$. This means that for some t, CONS must mention $Do(F_i, \text{TRUE}, t+1)$ in one of its axioms. This can only be an axiom ϕ of form (10.17) and it follows that the left-hand side of this axiom, instantiated to time t, holds in $\mathcal{M}$ and hence that D contains an effect law of form (10.18) corresponding to ϕ such that $\{a^1, \ldots, a^m\} \subseteq a_t$ and $F_{j_1}^{b_1}, \ldots, F_{j_m}^{b_m}$ all hold in σ_t. From Sublemma 10.25 on page 263 it follows that F_i is an effect of a_t in σ_t, so $F_i \in E^+_{a_t}(\sigma_t)$. □

The only-if direction of (10.24) can be proven in a manner completely analogously to the if-direction, again using the claim above. We omit the details.

This finishes the proof of Lemma 10.24. □

The Actual Proof

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

Proposition 10.26 *For any T_C for an instance of our language and any domain description D for a language $\mathcal{L}_3$ such that $T_C \sim D$, let $\mathcal{M}$ be any model with $\mathcal{M} \models_{\mathbf{c}} T_C$. For such an $\mathcal{M}$ we have:*

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

Proof: Easy and omitted. □

We now first prove the first part of Theorem 10.11, which is restated below.

Theorem 10.11 (part I) *For any theory T_C for our language and any domain description D for a language $\mathcal{L}_3$ of Baral and Gelfond's such that (a) $T_C \sim D$ and (b) D is not* ambiguous, *we have:*

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

Proof: By Lemma 10.22 and Lemma 10.23 together we have that there exists an $\mathcal{M}$ with $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$ that is causally consistent to M. Clearly $\mathcal{M} \cong M$. Suppose that $\mathcal{M}$ is *not* preferred, i.e. that there is an $\mathcal{M}^* \models_{\mathbf{c}} T_{\mathrm{C}}$ with $\mathcal{M}^*[Ab_1] \subsetneqq \mathcal{M}[Ab_1]$. By Lemma 10.24, this means that there is an $M^* = (\Psi^*, \Sigma^*)$ such that M^* and $\mathcal{M}^*$ are causally consistent and hence condition **C1** of the definition of modelhood (Definition 10.19) is satisfied for M^*. Applying Lemma 10.22 (first part) we find that for such an M^*, condition **C2** of that definition is satisfied too. Since $\mathcal{M}^*[Ab_1] \neq \mathcal{M}[Ab_1]$, we have by Proposition 10.26 and the fact that $\mathcal{M} \cong M$ and $\mathcal{M}^* \cong M^*$ that $\Sigma^*(s_N) \neq \Sigma(s_N)$. We are assuming M is a model of D, so we have by condition **C3** in Definition 10.19 that it is *not* the case that $\Sigma^*(s_N) \leq \Sigma(s_N)$. Then it is not the case either that $a_t^* \subseteq a_t$ for all t (where a_t is defined as in Definition 10.10, and a_t^* stands for the a_t belonging to Σ^*). Since $M \cong \mathcal{M}$ and $M^* \cong \mathcal{M}^*$, it follows by Definition 10.10 and Proposition 10.26 that it is not the case that $\mathcal{M}^*[Ab_1] \subsetneqq \mathcal{M}[Ab_1]$ and we have arrived at a contradiction. □

We now restate and prove the second part of Theorem 10.11.

Theorem 10.11 (part II) *For any theory T_{C} for our language and any domain description D for a language $\mathcal{L}_3$ of Baral and Gelfond's such that (a) $T_{\mathrm{C}} \sim D$ and (b) D is not* ambiguous, *we have, for all $\mathcal{M}$:*

$$\mathcal{M} \textit{ is a preferred model of } T_{\mathrm{C}} \Rightarrow$$
$$\textit{there exists a model } M \textit{ for } D \textit{ such that } M \cong \mathcal{M}$$

Proof: Clearly, if $\mathcal{M}$ is a preferred model of T_{C} then also $\mathcal{M} \models_{\mathbf{c}} T_{\mathrm{C}}$. By Lemma 10.24 we have that there exists an M that is causally consistent to $\mathcal{M}$ and that therefore satisfies condition **C1** of the definition of modelhood (Definition 10.19). Clearly, $M \cong \mathcal{M}$. By Lemma 10.22 (first part) we further have that M also satisfies condition **C2** of Definition 10.19. We now prove that M is a model of D by showing that assuming that condition **C3** of Definition 10.19 does *not* hold for M leads to a contradiction.

So assume this condition does not hold for M. Then there exists $M' = (\Psi', \Sigma')$ that satisfies **C1** and **C2** of Definition 10.19 with $\Sigma'(s_N) \neq \Sigma(s_N)$ and $\Sigma'(s_N) \leq \Sigma(s_N)$ and $\Psi'(\Sigma'(s_N))$ is defined. By Lemma 10.23 and Lemma 10.22 (second part) together we have that there is an $\mathcal{M}' \models_{\mathbf{c}} T_{\mathrm{C}}$ with $\mathcal{M}' \cong M'$. We will now 'stretch' $\mathcal{M}'$ into another $\mathcal{M}'' \models_{\mathbf{c}} T_{\mathrm{C}}$ which will have the property that $\mathcal{M}''[Ab_1] \subsetneqq \mathcal{M}[Ab_1]$. In $\mathcal{M}''$ exactly the same events and changes take place as in $\mathcal{M}'$, but the time in between two events/changes may be larger. We now prepare the construction of $\mathcal{M}''$:

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

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

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

Clearly, all predicates in the language for $\mathcal{M}''$ have now been defined for all t, f, e, b (notice that Ab_1 only exists for event-time pairs, not for fluent-time pairs, see Chapter 9, Section 9.6.5).

One easily verifies that an $\mathcal{M}''$ as defined above indeed exists, and that from the fact that $\mathcal{M}' \models_{\mathbf{c}} T_{\mathrm{C}}$, it follows that also $\mathcal{M}'' \models_{\mathbf{c}} T_{\mathrm{C}}$ (simply check all the axioms in CONS and EQ; we omit the details). We see that $\{(e,t) | \mathcal{M}'' \models \mathit{Ho}(e,t)\} \subseteq \{(e,t) | \mathcal{M} \models \mathit{Ho}(e,t)\}$ and hence by Proposition 10.26 $\mathcal{M}''[Ab_1] \subseteq \mathcal{M}[Ab_1]$. Since $\Sigma'(s_N) \neq \Sigma(s_N)$, by construction of $\mathcal{M}''$ and Proposition 10.26 we also have $\mathcal{M}''[Ab_1] \neq \mathcal{M}[Ab_1]$. This means $\mathcal{M}''$ is preferred over $\mathcal{M}$ and hence $\mathcal{M}$ is not a preferred model of T_{C}, in contradiction with our assumption. $\square$

Epilogue: Nonmonotonicity and Probability

In Chapter 9, Section 9.6.5, we extended our causal theories with a nonmonotonic component so as to be able to handle causal chains of events and 'surprises'. We used the simplest form of preferential semantics - simply picking the models of a theory with a minimal interpretation of the Ab-predicate. In this Epilogue, we show how this form of nonmonotonicity can be given a probabilistic interpretation, and we shall argue that there is advantage in taking a probabilistic point of view. However, we stress at the outset that we do *not* claim that '*all* uses of nonmonotonicity can (or should) be dealt with probabilistically'; more about this below.

To establish the connection between minimizing abnormality and probability theory, we need to invoke two controversial ideas: the 'Principle of Insufficient Reason' (page xi) (or equivalently, the Maximum Entropy Principle, Chapter 3) and the idea of viewing 'abnormalities' as 'events with small probability'. Both ideas have been used before to endow non-monotonic logics with probabilistic semantics [116, 58, 59, 114, 5]; however, their use remains controversial [114].

Some logicians feel that a probabilistic treatment of uncertainty, which necessarily involves such controversial steps, may lead to the inference of unwarranted conclusions. We will see that, the way they are used here, both the principle of insufficient reason and the probabilistic view on abnormalities can be understood from a point of view inspired by MDL. We shall argue that, if this MDL interpretation is taken seriously, then no unwarranted conclusions can be drawn. On the contrary, we shall show that the probabilistic/MDL interpretation forces one to explicitly state certain assumptions that remain hidden in the non-probabilistic approach. By establishing a connection between minimizing abnormality and the MDL Principle, we relate the present (third) part of the thesis to the first two parts.

Before We Start: Non-Probabilistic Nonmonotonicity Exists!

Before we start we emphasize that we do not claim *all* forms of nonmonotonic reasoning to be related to probability theory. In Chapter 8, Section 8.4 we presented an 'alternative research goal' in which we distinguished between the use of nonmonotonicity for theory completion (implemented by a nonmonotonic entailment relation $\approx$ which we called *completion mapping*) and the use of nonmonotonicity to impose a preference ordering over the set of possible realizations of a domain. In our 'causal theories' of chapters 9 and 10 the distinction returned: the minimization of Do corresponded to a completion mapping, the minimization of Ab_1 and Ab_2 corresponded to finding the most preferred (least surprising) models of our (already completely specified) domains. Here, we shall only be concerned about the latter use of nonmonotonicity, the minimization of abnormality in order to find the most preferred models of a 'completely specified' domain. We do not claim that the use of nonmonotonicity for theory completion is in any way related to probability theory.

1 Minimizing Abnormality and Maximizing Probability

Preliminaries

For simplicity, we will only consider the propositional case (we say something about the first-order case later). We start by assuming the notation and terminology introduced in Chapter 9, page 210. We introduce some additional notation concerning probability distributions over sets of models. Let $\mathbf{X} = \{X_1, \ldots, X_m\}$ be a finite set of propositional variables. By a probability distribution over $\mathbf{X}$ we mean a probability distribution over the sample space $\mathbf{E}$ consisting of all interpretations of $\mathbf{X}$. We write $P(\mathcal{M})$ to denote the probability of interpretation $\mathcal{M}$. Let $\mathbf{M}$ be some set of interpretations over $\mathbf{X}$. By $P(\mathbf{M})$ we denote the probability of the 'event' that the actual interpretation is contained in $\mathbf{M}$. Hence $P(\mathbf{M}) = \sum_{\mathcal{M} \in \mathbf{M}} P(\mathcal{M})$.

For a propositional formula Γ, we define $P(\Gamma) = P(\mathbf{M})$ where $\mathbf{M}$ is the set of $\mathcal{M}$ with $\mathcal{M} \models \Gamma$. Hence $P(\Gamma) = \sum_{\mathcal{M} \models \Gamma} P(\mathcal{M})$. Conditional probability is defined as usual: $P(\Gamma_1|\Gamma_2) = P(\Gamma_1 \wedge \Gamma_2)/P(\Gamma_2)$. We freely write $P(X_i = \text{TRUE})$ to denote $P(X_i)$ and $P(X_i = \text{FALSE})$ to denote $P(\neg X_i)$. Note that P is defined both over sets of models (as in $P(\mathbf{M})$) and over propositions (as in $P(\Gamma)$).

For now we shall assume that there are only two kinds of elements in $\mathbf{X}$: those whose value we are ignorant about (or, more properly, about which we want to make no assumptions) and those whose truth is considered 'abnormal' or 'unusual'. We can therefore partition $\mathbf{X}$ as follows: $\mathbf{X} = \mathbf{Ind} \cup \mathbf{Ab}$. (**Ind** being short for 'indifferent'). **Ind** contains the atoms we want to make no assumptions about while **Ab** contains the atoms that represent abnormalities.

Preferred Models According to the simple form of nonmonotonic reasoning introduced on page 184 and used in Chapter 9, Section 9.6.5, we should prefer the models of a theory T with the smallest interpretation of the elements in **Ab**; here we call the models that are selected in this way the L-preferred models:

Definition E.1 *We call a model $\mathcal{M}$ for a propositional theory T an L-preferred model of T iff there is no model $\mathcal{M}'$ for T with*

$$\{Ab \mid Ab \in \mathbf{Ab}, \mathcal{M}' \models Ab\} \subsetneq \{Ab \mid Ab \in \mathbf{Ab}, \mathcal{M} \models Ab\} \tag{E.1}$$

(note that Ab is used as a meta-variable here whose value can be any of the variables X_i that are contained in $\mathbf{Ab}$).

From the probabilistic point of view, our 'preference ordering' over models should be expressed by a probability distribution P over the space of models and we should prefer the models with maximum probability[6]. We will now show that the minimization of abnormality corresponds to picking the most probable model when a certain class of probability distributions is used. We call these distributions *simple*:

Definition E.2 *Let P be a distribution over $\mathbf{X} = \mathbf{Ind} \cup \mathbf{Ab}$. We call P* simple *iff*

1. *All variables in $\mathbf{X}$ are independent: for all $(b_1, \ldots, b_m) \in \mathbf{B}^m$ and $1 \leq i \leq m$,*

$$P(X_i = b_i | X_1 = b_1, \ldots, X_{i-1} = b_{i-1}, X_{i+1} = b_{i+1}, X_m = b_m) = P(X_i = b_i).$$

2. *for all $Ind \in \mathbf{Ind}$, $P(Ind = \text{TRUE}) = P(Ind = \text{FALSE}) = 1/2$.*

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

A precise interpretation and discussion of Definition E.2 will be given in Section 2. Briefly, item (1) is an assumption that is implicit in the use of abnormalities. Item (2) expresses our ignorance about the variables in $\mathbf{Ind}$. Item (3) expresses that $\neg Ab$ is preferred over Ab. We will be interested in the following case: our preference ordering over models can be modeled by some simple P, but it is not known which. Hence the specific values of $P(Ab = \text{TRUE}) = \epsilon_{Ab}$ are unknown.

We call $\mathcal{M}$ a *probabilistically preferred* model for theory T under P if $\mathcal{M}$ is one of the models that maximize $P(\mathcal{M}|T)$ ('the probability of $\mathcal{M}$ given that the set of propositions T holds'). We call $\mathcal{M}$ a *potentially probabilistically preferred model* for T if there exists a simple P over $\mathbf{X}$ under which $\mathcal{M}$ is a preferred model for T: if we are given theory T but the only thing we know about P is that it is simple, then it seems reasonable to consider *every* model that is potentially preferred. We write PP-preferred as short for 'potentially probabilistically preferred'. The proposition below shows that picking the PP-preferred models is equivalent to picking the models which are minimal in the usual sense of nonmonotonic logic:

Proposition E.3 *$\mathcal{M}$ is a PP-preferred model for theory T iff $\mathcal{M}$ is an L-preferred model for theory T.*

Proof: (only-if) Suppose $\mathcal{M}$ is a PP-preferred model for T. Let us assume, by means of contradiction, that $\mathcal{M}$ is not L-preferred. Then by definition there exists a model $\mathcal{M}' \models T$ such that (E.1) holds for $\mathcal{M}'$ and $\mathcal{M}$. We will show that, for every simple P, $P(\mathcal{M}'|T)$ must be larger than $P(\mathcal{M}|T)$. This implies that $\mathcal{M}$ is *not* a PP-preferred model of T and hence gives the desired contradiction.

[6]Some people think it strange that preferred models should coincide with 'the most probable' models. From both a Bayesian (subjective) and an MDL-point of view this makes perfect sense as we shall discuss further on.

To show this, note first that, since both $\mathcal{M} \models T$ and $\mathcal{M}' \models T$, we have $P(\mathcal{M}'|T) > P(\mathcal{M}|T) \Leftrightarrow P(\mathcal{M}') > P(\mathcal{M})$. Define $\mathcal{M}[X_i]$ to be the $b_i \in \mathbf{B}$ for which $\mathcal{M} \models X_i \equiv b_i$. Since P is simple, we have by Definition E.2:

$$P(\mathcal{M}') = \prod_{1 \leq i \leq m} P\{X_i = \mathcal{M}'[X_i]\} = P(\mathcal{M}) \cdot \prod_{Ab \in \mathbf{D}} \frac{1 - P(Ab)}{P(Ab)} = P(\mathcal{M}) \cdot \prod_{Ab \in \mathbf{D}} \frac{1 - \epsilon_{Ab}}{\epsilon_{Ab}}$$

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

(if) Assume that $\mathcal{M}$ is an L-preferred model. Let P be such that for all $Ab \in \mathbf{Ab}$ with $\mathcal{M} \models Ab$, $P(Ab) = 1/2 - \epsilon$ where $\epsilon > 0$ is some small number, while $P(Ab') = \epsilon$ for all $Ab' \in \mathbf{Ab}$ with $\mathcal{M} \not\models Ab'$. Clearly, P is simple. We are assuming that $\mathcal{M}$ is L-preferred so there is no model $\mathcal{M}'$ for T such that (E.1) holds. This means that all $\mathcal{M}' \neq \mathcal{M}$ with $\mathcal{M}' \models T$ and $P(\mathcal{M}') \neq P(\mathcal{M})$ satisfy $\mathcal{M}' \models Ab'$ for some $Ab' \in \mathbf{Ab}$ with $\mathcal{M} \not\models Ab'$. This implies that by picking $\epsilon > 0$ small enough, we can get $P(\mathcal{M}) > P(\mathcal{M}')$ and, since $\mathcal{M} \models T$ and $\mathcal{M}' \models T$, also $P(\mathcal{M}|T) > P(\mathcal{M}'|T)$. This implies that $\mathcal{M}$ is a PP-preferred model. □

Summarizing, if we are willing to assume that P is simple but otherwise unknown, then the two approaches to handling abnormalities can be reconciled. We will discuss how exactly P should be interpreted and whether or not the assumption that P is simple can be justified in the next section. First, we show how to extend the definition of 'simple' P to domains involving several levels of abnormalities and to the first-order setting we considered in Chapter 9.

1.1 Extension to Several Levels of Abnormalities

We may extend our concept of 'simple probability distributions' to several 'levels' of abnormality by further partitioning $\mathbf{X}$. For example, if we need two levels, we may set $\mathbf{X} = \mathbf{Ind} \cup \mathbf{Ab}_1 \cup \mathbf{Ab}_2$. In this new setting, a 'simple' distribution P becomes a P such that all variables in $\mathbf{X}$ are still independent, while

$$\begin{aligned} &\text{for all } Ab_1 \in \mathbf{Ab}_1: \quad && P(Ab_1) < 1/2\ ; \\ &\text{for all } Ab_2 \in \mathbf{Ab}_2: \quad && 0 < P(Ab_2) < \min_{Ab_1 \in \mathbf{Ab}_1} (P(Ab_1))^k. \end{aligned} \tag{E.2}$$

where k is the number of propositional variables in $\mathbf{Ab}_1$. One easily verifies that in this way, models without abnormalities in $\mathbf{Ab}_2$ will always be PP-preferred over models with abnormalities in $\mathbf{Ab}_2$, no matter the number of abnormalities in $\mathbf{Ab}_1$. Such a two-step hierarchy of abnormalities is often needed in nonmonotonic reasoning; for example, in Chapter 9, Section 9.6.5 we needed one level of abnormality to model surprises and another level to model the occurrence of unexpected events, since 'surprises' where considered more 'abnormal' than unexpected events (recall that a surprise was defined as a change of fluent value without an event according for the change). Clearly, if needed we can also introduce a set $\mathbf{Ab}_3$ of abnormalities associated with even smaller probabilities, and in the same way a set $\mathbf{Ab}_4$, $\mathbf{Ab}_5$ etc. In this way we can probabilistically represent every finite hierarchy of abnormalities.

The First-Order Case Thus far, our analysis has only been for propositional domains. In general, it is not clear how to extend the present mapping of preference orderings to probability distributions to the full first-order case (for an attempt to establish a connection between probability theory and first-order logic, see [5]). Nevertheless, as soon as a domain is essentially finite (and hence could in principle be modeled by a propositional theory), an appropriate version of 'simple' probability distributions can be defined and the mapping holds again.

Therefore, if we restrict the first-order theories that we considered in Chapter 9 to domains containing only a finite number of fluents, events and time points, the minimization of abnormality introduced in Section 9.6.5 can once more be interpreted probabilistically. In fact, the minimization procedure given by Definition 9.23 of page 230 can be given a probabilistic interpretation completely analogously to the minimization procedure defined in Definition E.1 of this Epilogue. We will not show this in full detail; instead, we will only roughly indicate how it may be achieved. Let T be a causal theory defined as in Definition 9.27, page 232. By restricting ourselves to a finite number of time points and a finite set of fluents, we can treat the whole domain by propositional logic and regard each instantiation of each predicate as a separate atom. We can then regard each instantiation of Ab_1 as a separate weak abnormality and each instantiation of Ab_2 as a separate strong abnormality. Definition 9.23 then reduces to the following preference scheme: select the causal models for T with the smallest (in the subset sense) number of abnormalities $Ab_2 \in \mathbf{Ab}_2$ and, among these, further select the models with the smallest (in the subset sense) number of $Ab_1 \in \mathbf{Ab}_1$. It is easy to show (analogously to Proposition E.3) that this is equivalent to selecting the potentially preferred causal models for T, if we assume a 'simple' probability distribution where 'simple' is now defined also for $Ab_2 \in \mathbf{Ab}_2$ (as in Equation E.2).

2 Interpretation in the Spirit of MDL

2.1 Bayesian interpretation

From the Bayesian (subjectivist) view on probability theory, probabilities can be used to represent degrees of belief. If an agent assigns a high probability to proposition A, this means that he (it) has a high degree of belief in A. The notion of 'degree of belief' can be given a concrete interpretation in terms of betting games: the probability $P(A)$ that an agent assigns to A determines at what odds the agent would accept the bet 'A is true': if $P(A) = p$, then the agent should accept a bet at odds 1-to-q if $q < p$ and he should refuse such a bet if $q > p$ (this interpretation has been emphasized by De Finetti; see [38, 114]).

From the point of view of the MDL Principle, the connection between probabilities and betting is somewhat different and more direct: the MDL Principle allows us to *identify* probability distributions with betting strategies. This turns the probabilistic interpretation of default reasoning into a much less abstract 'gambling' interpretation, as we will now show.

2.2 MDL Interpretation

As we discussed in Chapter 2, Section 2.2, the probability distributions that MDL is concerned with are to be interpreted as *models* of data, not to be confused with the traditional notion of a probability distribution 'according to which data are drawn'. We saw that each such a 'modeling' distribution P can be interpreted as a code with codeword lengths $L(D) = -\log P(D)$ for all D; the better P fits D, the shorter the codelength $L(D)$ (and conversely, each code can be interpreted as a probability distribution). We also saw (Chapter 2, Section 2.7) that every 'modeling' probability distribution P may be interpreted as defining a *betting strategy.* We will now explain what this means in a logical framework.

Betting Strategies in a Logical Framework In a logical framework, P is defined over the models of a logical theory T. The betting strategy concerns situations in which the agent does not know the true model $\mathcal{M}$ and it can express its beliefs about the actual situation by a probability distribution P over the set of possible models $\mathbf{M}$. The phrase 'an agent adopts distribution P' means the following: imagine the agent is asked to take part in a game in which it would have to place bets on the different models at equal odds (i.e. it has to divide its capital over all models $\mathcal{M}$; then, the true model $\mathcal{M}$ is revealed and an amount of c times the bet placed on $\mathcal{M}$ is paid to the agent, where c is equal for all $\mathcal{M}$). In such a game, the agent would divide its capital over the set of possible models such that the capital put on model $\mathcal{M}$ is proportional to $P(\mathcal{M})$. If the agent is rational, then $P(\mathcal{M}_1) > P(\mathcal{M}_2)$ iff the agent thinks that $\mathcal{M}_1$ is more likely than $\mathcal{M}_2$. We will henceforth assume that an agent always adopts the betting strategy that it considers optimal.

It is important to note that the agent invests some of its capital in each x with $P(x) > 0$. This MDL-way of associating bets with probabilities is different from the Bayesian way we described above. To see this, let $\mathbf{E} = \{1, 2\}$. In the Bayesian (De Finetti's) view, if an agent's belief is expressed by $P(X = 1) = 1/2$, this means it is willing to invest *every* fraction $q < 1/2$ of its capital in the game with pay-off 1 if X turns out to be 1 and pay-off 0 otherwise. From the MDL point of view (at least in our interpretation of MDL), $P(X = 1) = 1/2$ is primarily identified *only* with a willingness to invest 50 % of one's capital on outcome 1 and 50 % on outcome 2, where both bets are placed at odds 1-to-2 (see Chapter 2, Section 2.7). Under this bet the total amount of capital will not change whatever the actual outcome. It thus constitutes an 'empty' bet, that can be properly used to express complete ignorance about which of the two outcomes will actually take place.

Armed with our interpretation of probability distributions as gambling strategies, we will now interpret the definition of 'simple' probability distributions (Definition E.2) in gambling terms. Using Proposition E.3 this will also give us an interpretation of the models 'with the least abnormalities' in gambling terms.

Definition E.2 Revisited Assume an agent uses a 'simple' distribution P (Definition E.2) to express its knowledge about a domain involving (1) abnormalities (variables in **Ab**) and (2) atoms about which it does not want to make any assumptions (variables in **Ind**). The question is whether P, viewed as a betting strategy, correctly represents the agent's qualitative knowledge about the domain.

Simple distributions P were defined in three steps. In the first step, all variables in $\mathbf{X}$ were assumed to be independent. In betting terms, this says that the agent's bets on the outcome of variable X_i are *not influenced* by knowledge of the outcomes of (any combination of) the other variables $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m$. Whether this 'subjective independence' assumption can be justified or not will be discussed below.

The second step in the definition of simple P assigns the uniform distribution over the variables the agent wants to make no assumptions about. It is a form of the Principle of Insufficient Reason (and as such, a special case of Maximum Entropy, and as such, a version of MDL; see Chapter 3). It says that the agent should place an 'empty' bet (see above) on the variables it wants to make no assumptions about: it places half of its capital on Ind = TRUE and at the same time half of its capital on Ind = FALSE at odds 1-to-2. This is clearly the best it can do if it wants to make no assumptions about Ind, since every other assignment of probabilities (bets) leads to a potentially higher loss in betting.

The third item in Definition E.2 says that the agent is prepared to bet a larger percentage of its capital on outcome $\neg Ab$ than on outcome Ab; since ϵ_{Ab} is not specified, the item says nothing about the exact percentage. This item merely translates the agent's qualitative beliefs about a domain in terms of betting strategies. Items 1 and 2 need more explanation which follows below. We start with item 1.

Independence Assumption Item 1 expresses that our 'degree of belief' in proposition $X_i = b_i$ is not influenced by variables different from X_i. This will of course not be justified in all domains. In domains where it is not, it will lead to 'stupid' bets. But Proposition E.3 tells us that in such domains, minimization of abnormalities will also lead to a preference of 'stupid' models. A good example is the Yale Shooting domain in its original formulation (page 190), where the 'law of persistence' was formalized as follows:

$$\forall f, t \,.\, \neg Ab(f,t) \supset [Ho(f,t) \equiv Ho(f,t+1)]$$

If we use a probability distribution P' over the atoms in the language that corresponds to our intuitive beliefs about the domain, then the following holds: while $Ab(f,t)$ (denoting that fluent f changes value) has a very small probability *a priori*, if no observations are available, it can nevertheless receive a very *high* probability once additional information becomes available. For example, if it is given that $Ho(Shoot, 2) \wedge Ho(Loaded, 2)$ (*Shoot* takes place with a loaded gun) then $Ab(Alive, 2)$ suddenly receives a very high probability: conditioned on $Ho(Shoot, 2) \wedge Ho(Loaded, 2)$, our degree of belief in $Ab(Alive, 2)$ becomes much higher. Hence the percentage of our capital we would be prepared to put on $Ab(Alive, 2)$ is *not* independent of additional knowledge we may have about the domain: the distribution P' which corresponds to our intuitive beliefs does *not* render the abnormalities $Ab(f,t)$ probabilistically independent of the outcomes of the other atoms in the domain. *Such a situation cannot be modeled by a simple minimization of abnormality.*

Two Mistakes In the early days of nonmonotonic reasoning, it was believed that a single mechanism (like minimizing abnormality) could be used to handle most if not all kinds of different defaults. The probabilistic interpretation of abnormalities shows that minimizing abnormality only works under strong independence assumptions. The advantage of the probabilistic formulation is that it makes these assumptions explicit. As such, it 'explains' why the basic kind of default reasoning – based on minimizing abnormality – turned out to be so problematic and brittle.

While the original mistake in the nonmonotonic reasoning community had been to use the mechanism of minimizing abnormality in domains where it cannot be applied, the resulting difficulties led some researchers to make (what we view as) the opposite mistake: they started to distinguish between concepts which, from the probabilistic/betting strategy point of view, are identical. We give a little example. In Geffner [51] we find two kinds of abnormalities: 'explained' and 'unexplained' abnormalities. For example, the abnormality that *Alive* becomes $\neg$*Alive* when a loaded gun is fired is 'explained' while the abnormality of a gun unloading itself is 'unexplained'. In the probabilistic view, an abnormality stands for a proposition which is assigned low probability according to some distribution P. In Chapter 2, Section 2.2, we discussed the MDL interpretation of probability distributions as *models* (not in the logical sense of a mathematical structure): a probability distribution P assigns to each possible realization of data (in this case, to each structure $\mathcal{M}$) a number that indicates how well that data (structure) is *explained* by the model P. As such, an 'abnormality' is by definition 'something that is not well-explained'. In a more proper terminology, what was called an 'explained' abnormality above would then be no abnormality at all. More generally, in the MDL view of probabilities as models, the expression $P(X|Y)$ can be informally regarded as measuring 'how well Y explains X under the model P', or, equivalently, 'how expected X is given Y under model P' or yet equivalently, 'how surprising $\neg X$ is given Y under model P'. Note that the 'event' Y itself may represent a probabilistic model, not necessarily a set of outcomes (logical models). If there is no conditioning event Y, then $P(X)$ can be regarded as expressing how well the model P itself explains X.

Dangerous Use of Probabilities This view of probability distributions as 'explanation measurers' has been criticized by logicians and philosophers [25]. Whereas we just advocated the use of probabilistic semantics on the grounds that it makes explicit some assumptions (about independence) which remain hidden in the abnormality approach, these researchers fear the contrary effect: they argue that by adopting a probability distribution over a set of models one implicitly makes additional, unwarranted assumptions about the domain under consideration. Their objections are often related to the 'Ex Nihilo Nihil' criticism concerning the Maximum Entropy Principle and its special case, the Principle of Insufficient Reason (Chapter 3, Section 3.6.2): it is feared that, by using $P(\text{Ind} = \text{TRUE}) = 1/2$ to express our ignorance about Ind, we are 'getting something (knowledge) out of nothing (ignorance)'. For example, if $\mathrm{Ind}_1, \mathrm{Ind}_2 \in \mathbf{Ind}$, then a simple probability distribution assigns $P(\mathrm{Ind}_1 = \text{TRUE}) = P(\mathrm{Ind}_2 = \text{TRUE}) = 1/2$ and renders Ind_1 and Ind_2 independent. It seems that one can then infer that

$$P((\mathrm{Ind}_1 \vee \mathrm{Ind}_2) = \text{TRUE}) = 3/4.$$

Hence the degree of belief in $\mathrm{Ind}_1 \vee \mathrm{Ind}_2$ would be higher than the degree of belief in $\neg(\mathrm{Ind}_1 \vee \mathrm{Ind}_2)$ and mere ignorance about Ind_1 and Ind_2, together with the knowledge that they are independent, would give us a kind of 'knowledge' about the proposition $\mathrm{Ind}_1 \vee \mathrm{Ind}_2$. However, we used P *only* to determine which models lead to the smallest loss in proportional betting, where the bets where placed sequentially on the outcomes of the variables in $\mathbf{X}$. We did *not* use it to place bets on (or make predictions about) propositions like $\mathrm{Ind}_1 \vee \mathrm{Ind}_2$. The bets we did place on the variables in **Ind** where 'empty', neither decreasing nor increasing our total capital whatever the actual outcome. Therefore, as long as we only use P to infer what model would have given the minimal loss in proportional betting (and *do not use it to infer anything else*), the assignment $P(\mathrm{Ind}_1 = \text{TRUE}) = 1/2$ is completely harmless. The assignment is, in fact, an example of the 'safe' version of Maximum Entropy we introduced in Chapter 4 (Figure 4.1): whatever the actual outcomes of the variables in **Ind**, our *expectation* according to P of the logarithm of our loss in proportional betting on the variables in **Ind** will be equal to the logarithm of the loss we will *actually* make; hence, if only used for proportional betting, P truely satisfies Leslie Ellis' Ex Nihilo Nihil dictum. This is explained at length in Chapter 4, Section 4.2.3; see also page 76.

3 Conclusion: Use Probability Theory, but not Always

We think that many of the domains considered in the nonmonotonic reasoning community can be appropriately modeled by a probabilistic semantics: as soon as the models of a domain are to be ordered in terms of 'how surprising they are', a probabilistic semantics seems appropriate; as should be clear from our 'betting strategy interpretation', such a semantics remains applicable in many cases where probabilities (such as $P(\mathrm{Ind} = \text{TRUE}) = 1/2$) cannot be related to frequencies, as long as we are careful about what inferences to make based on these probabilities. Even when defaults stand for conventions (see the example on page 186) they often induce an ordering on the models of a domain in terms of how 'surprising' they are. In such a case, it seems to us, a probabilistic semantics is appropriate. Above we showed that it can make explicit certain (independence) assumptions underlying the reasoning; many additional reasons can be found in [116] (for a general discussion and criticism, see [25]).

However, we disagree with those who think that there is no fundamental rôle for nonmonotonic formalisms whatsoever. As we pointed out at the beginning of this Epilogue, nonmonotonic mechanisms can also be used for *theory completion*: instead of writing down the full set of laws to which we want our domain to be subject, we only specify a subset of them and we use a nonmonotonic completion mapping to make these theories 'complete'. In fact, in the first-order version of our theory formulated in

Section 9.6.5, we first used the minimization of *Do* to do this theory completion; this, it seems to us, had nothing to do with probability theory. We then proceeded to select the least surprising models of the completed theory; here a probabilistic semantics is appropriate.