Chapter 1

Nonmonotonic Causal Logic

Hudson Turner

This chapter describes a nonmonotonic causal logic designed for representing knowledge about the effects of actions. A causal rule

$$\phi \Leftarrow \psi$$

where ϕ and ψ are formulas of classical logic, is understood to express that (the truth of) ψ is a sufficient condition for ϕ's being caused. A causal theory T is a set of causal rules, and is assumed to describe all such sufficient conditions. Thus, given an interpretation I, the set of formulas

$$T^I = \{\ \phi \mid \phi \Leftarrow \psi \in T \text{ and } I \models \psi\ \}$$

can be understood to describe everything caused in a world such as I (according to T). The *models* of causal theory T are those interpretations for which what is true is exactly what is caused: that is, the interpretations I such that I is the unique model of T^I. This fixpoint condition makes the logic nonmonotonic; adding causal rules to T may produce new models.

Causal theories allow for convenient formalization of such challenging phenomena as indirect effects of actions (ramifications), implied action preconditions, concurrent interacting effects of actions, and things that change by themselves. These capabilities stem from a robust solution to the frame problem [33]: one can write causal rules

$$\begin{array}{rcl} F_{t+1} & \Leftarrow & F_{t+1} \wedge F_t \\ \neg F_{t+1} & \Leftarrow & \neg F_{t+1} \wedge \neg F_t \end{array} \tag{1.1}$$

saying that a sufficient cause for fluent F's being true (respectively, false) at time $t+1$ is its being so at time t and remaining so at time $t+1$. In this way, persistent facts are effectively said to have inertia as their cause. On the other hand, the fixpoint definition of a model of a causal theory T requires that everything have a cause (according to T), and so any facts not explained by inertia must have some other explanation. Consequently, in the context of the fixpoint condition, the inertia rules (1.1) capture the commonsense notion

that things don't change unless they're made to. (The frame problem is briefly discussed in Chapter 6 as one of the motivations for nonmonotonic logics, and other solutions to the frame problem are presented in Chapters 16–18.)

In a causal theory, one easily describes not only direct effects of actions, but also their indirect effects, and even interacting effects of concurrent actions. For example, suppose there is a large bowl of soup, with left and right handles. We can describe the direct effect of lifting each side as follows.

$$\begin{array}{rcl} \mathit{up}(\mathit{left})_{t+1} & \Leftarrow & \mathit{lift}(\mathit{left})_t \\ \mathit{up}(\mathit{right})_{t+1} & \Leftarrow & \mathit{lift}(\mathit{right})_t \end{array} \tag{1.2}$$

More interestingly, we can then describe, without reference to any action, that if only one side of the bowl is up, then there is a cause for the soup's being spilled.

$$\mathit{spilled}_t \Leftarrow \mathit{up}(\mathit{left})_t \not\equiv \mathit{up}(\mathit{right})_t \tag{1.3}$$

Under this description, if the soup is not already spilled and both sides of the bowl are lifted at once, the soup will remain unspilled. But if only one side is lifted, the soup will be spilled. Let's consider these two scenarios in more detail. Assume that initially neither side is up and the soup is unspilled. In the first scenario, both handles are lifted concurrently, and at the next time: by (1.2) there is a cause for both sides' being up, and by inertia—that is, by (1.1) with *spilled* in place of F—there is a cause for the soup's being unspilled. For the second scenario, suppose instead that only the left handle is lifted. Then, at the next time: by (1.2) there is a cause for the left side's being up, by inertia there is a cause for the right side's remaining down, and by (1.3) there is a cause for the soup's being spilled. Notice that the formalization does not support the incorrect alternative outcome in which, after lifting only the left handle, both sides are up and the soup remains unspilled. Why not? By (1.2) there would be a cause for the left side's being up, and by inertia there would be a cause the soup's remaining unspilled, but there would be *no* cause for the right side to be up (and the definition of a model requires that everything have a cause).

The prior example is meant to help with intuitions about the fixpoint definition of a model and its role in the expressive possibilities of causal theories; in the interest of clarity, it is important to emphasize that the example causal theory is incompletely specified. Indeed, the fixpoint definition requires that *everything* have a cause according to the theory. Accordingly, about occurrences and nonoccurrences of actions A, we often write that they are caused in either case, as follows.

$$\begin{array}{rcl} A_t & \Leftarrow & A_t \\ \neg A_t & \Leftarrow & \neg A_t \end{array} \tag{1.4}$$

That is, if A occurs at time t, there is a cause for this, and if, on the other hand, A doesn't occur at time t, there is a cause for that. Similarly, we typically say that initial facts about fluents F are caused.

$$\begin{array}{rcl} F_0 & \Leftarrow & F_0 \\ \neg F_0 & \Leftarrow & \neg F_0 \end{array} \tag{1.5}$$

So, although everything in a model must be caused, it is convenient to take the view that some causes are exogenous to our description, and we can simply stipulate their existence, by writing rules such as those in (1.4) and (1.5). Moreover, this is only the most extreme

version of the general case. After all, the nonmonotonic logic described here is *causal* only in a limited sense: causal rules allow a distinction between being true and having a cause. Causal theories do not grapple with the question of what a cause may be, and do not support direct reasoning about what causes what. Fortunately, many questions related to reasoning about actions require only that we be able to determine what are the causally possible histories (of the world being described), and for this purpose it seems enough to be able to describe the conditions under which facts are caused.

This strategic emphasis on the distinction between what is caused and what is (merely) true can be understood to follow Geffner [14, 15], whose work was influenced by Pearl [35]. (Pearl's ideas on causality have been extensively developed since then. See, for instance, [36].) In the reasoning about action community, this line of work was motivated primarily by the ramification problem—the problem of representing and reasoning about indirect effects of actions. For some time there were attempts to describe relationships like those described by (1.3) through the use of "state constraints": formulas of classical logic such as

$$(up(left)_t \not\equiv up(right)_t) \supset spilled_t\,. \tag{1.6}$$

It seems that the crucial shortcoming of a formula like (1.6), for the purpose of reasoning about ramifications, is that it simply doesn't say enough about what can (and cannot) cause what. Indeed, it is equivalent to its contrapositive

$$\neg spilled_t \supset (up(left)_t \equiv up(right)_t)\,.$$

In the last 15 years, there have been *many* reasoning about action proposals incorporating more explicitly causal notions. The nonmonotonic causal logic described in this chapter was introduced in [30]. The most relevant prior work appears in [28, 29, 44]. A much fuller account of causal theories was published in 2004 [18], although a number of results presented in this chapter do not appear there. Causal theories have been studied, applied and extended in [24, 25, 17, 31, 45, 22, 27, 47, 2, 3, 4, 6, 7, 11, 20, 1, 5, 10, 21, 46, 13, 9, 42].

An implementation of causal theories—the Causal Calculator (CCALC)—is publicly-available, and many of the above-cited papers describe applications of it. The key to this implementation is an easy reduction from (a subclass of) causal theories to classical propositional logic, by a method introduced in [30], closely related to Clark's completion [8] for logic programs (discussed in Chapter 7). Thus, automated reasoning about causal theories can be carried out via standard satisfiability solvers. (The initial version of CCALC was due primarily to Norm McCain, and is described in his PhD thesis [32]. Since then it has been maintained and developed by Vladimir Lifschitz and his students at the University of Texas at Austin.)

Subsequent sections of this chapter are organized as follows.

- Section 1.1 defines causal theories (more adequately), considers a few examples, and remarks on several easy mathematical properties.
- Section 1.2 presents a "strong equivalence" result, which justifies a general replacement property, despite the nonmonotonicity of the logic.
- Section 1.3 specifies the reduction to classical propositional logic that makes automated reasoning about causal theories convenient.

- Section 1.4 demonstrates further expressive possibilities of causal theories, such as nondeterminism and things that change by themselves.
- Section 1.5 briefly describes the high-level action language $\mathcal{C}+$ that is based on causal theories.
- Section 1.6 characterizes the close mathematical relationship betweeen causal theories and Reiter's default logic, and notes the remarkable fact that inertia rules essentially like (1.1) appear already in Reiter's 1980 paper.
- Section 1.7 presents Lifschitz's reformulation of causal theories in higher-order classical logic, somewhat in the manner of circumscription.
- Section 1.8 presents the modal nonmonotonic logic UCL that includes causal theories as a special case.

1.1 Fundamentals

We first define a slight extension of usual (Boolean) propositional logic, convenient when formulas are used to talk about states of a system. Then we define the syntax and semantics of causal theories, make some observations, and consider a few examples, including a more precise account of the soup-bowl example already discussed.

1.1.1 Finite domain propositional logic

A (finite-domain) *signature* is a set σ of symbols, called *constants*, with each constant c associated with a nonempty finite set $Dom(c)$ of symbols, called the *domain* of c. An *atom* of signature σ is an expression of the form

$$c{=}v$$

("the value of c is v") where $c \in \sigma$ and $v \in Dom(c)$. A *formula* of σ is a propositional combination of atoms.

To distinguish formulas of usual propositional logic from those defined here, we call them "classical."

An *interpretation* of σ is a function mapping each element of σ to an element of its domain. An interpretation I *satisfies* an atom $c{=}v$ (symbolically, $I \models c{=}v$) if $I(c) = v$. The satisfaction relation is extended from atoms to arbitrary formulas according to the usual truth tables for the propositional connectives.

Also as usual, a *model* of a set Γ of formulas is an interpretation that satisfies all formulas in Γ. If Γ has a model, it is said to be *consistent*, or *satisfiable*. If every model of Γ satisfies a formula F, then we say that Γ *entails* F and write $\Gamma \models F$. Formulas, or sets of formulas, are *equivalent* if they have the same models.

A *Boolean* constant is one whose domain is the set $\{\mathsf{t}, \mathsf{f}\}$ of truth values. A *Boolean* signature is one whose constants are all Boolean. If c is a Boolean constant, we sometimes write c as shorthand for the atom $c{=}\mathsf{t}$. When the syntax and semantics defined above are restricted to Boolean signatures and to formulas that do not contain f, they turn into the usual syntax and semantics of classical propositional formulas. Even when the signature is not Boolean, there are easy reductions from finite domain propositional logic to classical propositional logic. (For more on this, see [18].)

1.1.2 Causal theories

Syntax

A *causal rule* (of signature σ) is an expression of the form

$$\phi \Leftarrow \psi \tag{1.7}$$

where ϕ and ψ are formulas (of σ). We call ϕ the *head* of the rule, and ψ the *body*. The intuitive reading of (1.7) is "ϕ is caused if ψ".

A *causal theory* (of σ) is a set of causal rules (of σ).

Semantics

Consider any causal theory T and interpretation I (of σ). We define

$$T^I = \{\phi \mid \phi \Leftarrow \psi \in T, I \models \phi\}\,.$$

Intuitively, T^I describes what is caused in a world like I, according to T.

An interpretation I is a *model* of a causal theory T if I is the only model of T^I.

Observation 1 For any causal theory T and interpretation I, I is a model of T iff, for all formulas ϕ,

$$I \models \phi \text{ iff } T^I \models \phi\,.$$

This observation corresponds precisely to the informal characterization we began with—the models of a causal theory T are the interpretations for which what is true is exactly what is caused (according to T).

Observation 2 If I is a model of a causal theory T, then $I \models \psi \supset \phi$ for every $\phi \Leftarrow \psi \in T$.

Examples

Take

$$\sigma = \{p\}\,,\ \ \mathit{Dom}(p) = \{1,2,3\}\,,\ \ T_1 = \{\,p{=}1 \Leftarrow p{=}1\,\}\,.$$

There are three interpretations (of σ), as follows: $I_1(p) = 1$, $I_2(p) = 2$, $I_3(p) = 3$. Notice that $T_1^{I_1} = \{p{=}1\}$. Clearly $I_1 \models T_1^{I_1}$, while $I_2 \not\models T_1^{I_1}$ and $I_3 \not\models T_1^{I_1}$, which shows that I_1 is a model of T_1. On the other hand, $T_1^{I_2} = T_1^{I_3} = \emptyset$. So I_2 is not the unique model of $T_1^{I_2}$, nor is I_3 the unique model of $T_1^{I_3}$. Consequently, neither I_2 nor I_3 is a model of T_1.

Consider the causal theory T_2 obtained by adding the causal rule $p{=}2 \Leftarrow p{=}2$ to T_1. One easily verifies that both I_1 and I_2 are models of T_2, which shows that causal theories are indeed nonmmonotonic: adding a rule to T_1 produced a new model.

Now let's reconsider the soup-bowl domain, discussed somewhat informally in the introductory remarks. Take the Boolean signature

$$\sigma = \{\mathit{upL}_0, \mathit{upR}_0, \mathit{sp}_0, \mathit{upL}_1, \mathit{upR}_1, \mathit{sp}_1, \mathit{liftL}_0, \mathit{liftR}_0\}\,.$$

Then following causal theory T represents the soup-bowl domain, with constants abbreviated: upL_0 for $up(left)_0$, sp_1 for $spilled_1$, $liftL_0$ for $lift(left)_0$, and so forth.

$$\begin{array}{rcl@{\qquad}rcl@{\qquad}rcl}
upL_0 & \Leftarrow & upL_0 & upL_1 & \Leftarrow & upL_1 \wedge upL_0 & & & \\
\neg upL_0 & \Leftarrow & \neg upL_0 & \neg upL_1 & \Leftarrow & \neg upL_1 \wedge \neg upL_0 & liftL_0 & \Leftarrow & liftL_0 \\
upR_0 & \Leftarrow & upR_0 & upR_1 & \Leftarrow & upR_1 \wedge upR_0 & \neg liftL_0 & \Leftarrow & \neg liftL_0 \\
\neg upR_0 & \Leftarrow & \neg upR_0 & \neg upR_1 & \Leftarrow & \neg upR_1 \wedge \neg upR_0 & liftR_0 & \Leftarrow & liftR_0 \\
sp_0 & \Leftarrow & sp_0 & sp_1 & \Leftarrow & sp_1 \wedge sp_0 & \neg liftR_0 & \Leftarrow & \neg liftR_0 \\
\neg sp_0 & \Leftarrow & \neg sp_0 & \neg sp_1 & \Leftarrow & \neg sp_1 \wedge \neg sp_0 & & &
\end{array} \tag{1.8}$$

$$\begin{array}{rcl@{\qquad}rcl}
upL_1 & \Leftarrow & liftL_0 & sp_0 & \Leftarrow & upL_0 \not\equiv upR_0 \\
upR_1 & \Leftarrow & liftR_0 & sp_1 & \Leftarrow & upL_1 \not\equiv upR_1
\end{array} \tag{1.9}$$

Most of the causal rules in T are of the "standard" kinds already discussed: the first column of (1.8) says that facts about the initial time are exogenous; the second column of (1.8) says that fluents upL, upR, sp are inertial; and the third column says that causes of occurrence and nonoccurrence of the actions $liftL, liftR$ are exogenous. The "interesting" rules appear in (1.9): the two on the left describe the direct effects of lifting the left and right handles of the bowl; and the other two say that, at both times 0 and 1, if only one side of the bowl is up, then there is a cause for the soup's being spilled.

Without much difficulty, one can verify that the models of this formalization are as previously described. For instance, consider the interpretation in which both sides are initially down, the soup is initially unspilled, the left handle is lifted at time 0, and at time 1 the left side is up, the right side is not, and the soup is spilled. That is, take

$$I = \{\neg upL_0, \neg upR_0, \neg sp_0, liftL_0, \neg liftR_0, upL_1, \neg upR_1, sp_1\}\,.$$

Then

$$T^I = \{\neg upL_0, \neg upR_0, \neg sp_0, liftL_0, \neg liftR_0, upL_1, \neg upR_1, sp_1\}\,,$$

and so I is a model of T. (That is, I is the unique model of T^I.) By comparison, consider the interpretation

$$I = \{\neg upL_0, \neg upR_0, \neg sp_0, liftL_0, \neg liftR_0, upL_1, upR_1, \neg sp_1\}\,.$$

Then

$$T^I = \{\neg upL_0, \neg upR_0, \neg sp_0, liftL_0, \neg liftR_0, upL_1, \neg sp_1\}\,,$$

and so this interpretation I is not a model of T, since it is not the unique model of T^I. (Intuitively, there is no explanation for the right side's being up at time 1.)

Constraints

A causal rule with head $\bot$ is called a *constraint*. A constraint

$$\bot \Leftarrow \phi \tag{1.10}$$

can be understood to say that $\neg\phi$ must be the case, but without asserting the existence of a cause for $\neg\phi$. Constraints behave monotonically; that is, adding (1.10) to a causal theory simply rules out those models that satisfy ϕ.

Definitional extensions and defaults

It is straightforward to add a new Boolean constant d to the signature σ and define it using a formula ϕ of the original signature; simply add the rule

$$d \equiv \phi \Leftarrow \top\,. \tag{1.11}$$

Indeed, let T be a causal theory (of σ), with T_d the causal theory (of $\sigma \cup \{d\}$) obtained by adding (1.11) to T. For any interpretation I of σ, let I_d be the interpretation of $\sigma \cup \{d\}$ such that (i) $I_d(c) = I(c)$, for all $c \in \sigma$, and (ii) $I_d(d) = \mathsf{t}$ iff $I \models \phi$. Then, I is a model of T iff I_d is a model of T_d. Moreover, every model of T_d has the form I_d, for some interpretation I of σ.

More generally, we can add to σ a new constant d with $\mathit{Dom}(d) = \{v_1, \ldots, v_n\}$ and define d using formulas $\phi_2, \ldots, \phi_n$ of σ, no two of which are jointly satisfiable, by adding the following causal rules.

$$\begin{array}{rcl} d{=}v_1 & \Leftarrow & d{=}v_1 \\ d{=}v_2 & \Leftarrow & \phi_2 \\ & \vdots & \\ d{=}v_n & \Leftarrow & \phi_n \end{array} \tag{1.12}$$

Indeed, let T be a causal theory (of σ), with T_d the causal theory (of $\sigma \cup \{d\}$) obtained by adding the rules (1.12) to T. For any interpretation I of σ, let I_d be the interpretation of $\sigma \cup \{d\}$ such that (i) $I_d(c) = I(c)$, for all $c \in \sigma$, and (ii) for all $i \in \{2, \ldots, n\}$, $I_d(d) = v_i$ iff $I \models \phi_i$. Then, I is a model of T iff I_d is a model of T_d. Moreover, every model of T_d has the form I_d, for some interpretation I of σ.

Notice that this latter technique can also be understood as a way of giving new constant d a default value which is overridden just in case one of $\phi_2, \ldots, \phi_n$ is true.

1.2 Strong equivalence

Equivalence of causal theories is defined in the usual way—as having the same models—but since the logic is nonmonotonic, equivalence does not yield a replacement property. That is, it is not generally safe to replace a subset of the rules of a causal theory with an equivalent set of rules. For example, assume that the signature is Boolean, with two constants, p and q. Let S be the causal theory with rules

$$\begin{array}{rcl} p & \Leftarrow & q\,, \\ q & \Leftarrow & p\,, \end{array}$$

and let T be the causal theory with rules

$$\begin{array}{rcl} p & \Leftarrow & p\,, \\ q & \Leftarrow & q\,. \end{array}$$

Causal theories S and T are equivalent; for each the unique model is $\{p, q\}$. Now, let R consist of the single rule

$$\neg p \Leftarrow \neg p\,.$$

Notice that $S \cup R$ still has only the model $\{p, q\}$, while $T \cup R$ has a second model, $\{\neg p, q\}$. Thus, in the context of $S \cup R$ it is not safe to replace S with T, even though S and T are equivalent.

Of course it is clear that we can always safely replace a rule $\phi \Leftarrow \psi$ with a rule $\phi' \Leftarrow \psi'$ if ϕ is equivalent to ϕ' and ψ is equivalent to ψ'. But we can do better.

The crucial notion is "strong equivalence", introduced (for logic programming) in [23]. (See Chapter 7.) We say causal theories S and T are *strongly equivalent* if, for every causal theory R, $S \cup R$ is equivalent to $T \cup R$.

It is clear that strong equivalence yields the replacement property we want. If S and T are strongly equivalent, we can safely replace S with T no matter the context: doing so will not affect the set of models. But the definition is inconvenient to check, since it requires reasoning about all possible contexts $S \cup R$. For convenience, we want a rather different characterization of strong equivalence.

An *SE-model* of causal theory T is a pair (I, J) of interpretations such that

- $I \models T^I$, and
- $J \models T^I$.

Strong Equivalence Theorem [46]
Causal theories are strongly equivalent iff their SE-models are the same.

While deciding equivalence of causal theories is a Π_2^P-complete problem, deciding *strong* equivalence is co-NP-complete.

1.3 Completion

A causal theory is *definite* if

- the head of every rule is either an atom or $\bot$, and
- no atom occurs in the head of infinitely many rules.

We say an atom $c = v$ is *trivial* if $Dom(c) = \{v\}$. If a causal theory is definite, its *completion* consists of the following formulas.

- For each constraint $\bot \Leftarrow \phi$, include the formula $\neg\phi$.
- For each nontrivial atom A of the signature, include the formula

$$A \equiv (\phi_1 \vee \cdots \vee \phi_n)$$

where $\phi_1, \ldots, \phi_n$ are the bodies of the rules with head A. (If $n = 0$, the rhs is $\bot$.)

Completion Theorem [30, 18]
If a causal theory is definite, its models are exactly the models of its completion.

Assume that the signature has three atoms, p_0, p_1, and a_0, with $\mathit{Dom}(p_0) = \mathit{Dom}(p_1) = \{1, 2, 3\}$ and $\mathit{Dom}(a_0) = \{\mathsf{t}, \mathsf{f}\}$. Consider the following (definite) causal theory.

$$\begin{array}{lll}
p_0{=}0 \Leftarrow p_0{=}0 & & \\
p_0{=}1 \Leftarrow p_0{=}1 & & a_0 \Leftarrow a_0 \\
p_0{=}2 \Leftarrow p_0{=}2 & & a_0{=}\mathsf{f} \Leftarrow \neg a_0 \\
\\
p_1{=}0 \Leftarrow p_1{=}0 \wedge p_0{=}0 & & p_1{=}1 \Leftarrow p_0{=}0 \wedge a_0 \\
p_1{=}1 \Leftarrow p_1{=}1 \wedge p_0{=}1 & & p_1{=}2 \Leftarrow p_0{=}1 \wedge a_0 \\
p_1{=}2 \Leftarrow p_1{=}2 \wedge p_0{=}2 & & p_1{=}0 \Leftarrow p_0{=}2 \wedge a_0
\end{array} \tag{1.13}$$

Its completion is as follows.

$$\begin{array}{lll}
p_0{=}0 \equiv p_0{=}0 & & a_0 \equiv a_0 \\
p_0{=}1 \equiv p_0{=}1 & & a_0{=}\mathsf{f} \equiv \neg a_0 \\
p_0{=}2 \equiv p_0{=}2 & & \\
p_1{=}0 \equiv (p_1{=}0 \wedge p_0{=}0) \vee (p_0{=}2 \wedge a_0) & & \\
p_1{=}1 \equiv (p_1{=}1 \wedge p_0{=}1) \vee (p_0{=}0 \wedge a_0) & & \\
p_1{=}2 \equiv (p_1{=}2 \wedge p_0{=}2) \vee (p_0{=}1 \wedge a_0) & &
\end{array}$$

All but three of these formulas are tautological. Those three together can be simplified as follows.

$$\begin{array}{l}
p_0{=}0 \supset (p_1{=}0 \wedge \neg a_0) \vee (p_1{=}1 \wedge a_0) \\
p_0{=}1 \supset (p_1{=}1 \wedge \neg a_0) \vee (p_1{=}2 \wedge a_0) \\
p_0{=}2 \supset (p_1{=}2 \wedge \neg a_0) \vee (p_1{=}0 \wedge a_0)
\end{array}$$

The Completion Theorem would not be correct if we did not restrict the completion process to nontrivial atoms. Consider, for instance, the causal theory with no rules whose signature σ has only one constant c, with $\mathit{Dom}(c) = \{0\}$. This causal theory has one model—the only interpretation of σ. But if the definition of completion did not exclude trivial atoms, then the completion of this theory would be $c{=}0 \equiv \bot$, which is unsatisfiable.

The Completion Theorem implies that the satisfiability problem for definite causal theories belongs to class NP. In fact, it is NP-complete. Indeed, given any formula ϕ of Boolean signature σ, the causal theory $\{\bot \Leftarrow \neg\phi\} \cup \{c{=}v \Leftarrow c{=}v \mid c \in \sigma, v \in \{\mathsf{t}, \mathsf{f}\}\}$ is definite and has the same models as ϕ.

As mentioned previously, the Causal Calculator uses completion to automate reasoning about definite causal theories via standard satisfiability solvers for propositional logic. In relation to this, there are two complications to consider: (i) the completion formulas are not in clausal form, and (ii) the completion formulas are not Boolean. Both these obstacles can be efficiently overcome, as long as we are willing to extend the signature, with the result that all models of the resulting set of clauses correspond to models of the definite causal theory, and vice versa. (For more detail, see [18].)

1.4 Expressiveness

We have already seen an example involving conditional (direct) effects of actions (1.13), and have discussed at some length an example with indirect effects of actions and interacting effects of concurrent actions (1.8,1.9). We have mentioned other possibilities, such as

nondeterminism, implied action preconditions, and things that change by themselves. A few examples follow. Many additional examples, of these and other kinds, can be found in the cited papers on causal theories, including some developing the theme of "elaboration tolerance", which is crucial to the long-term success of approaches to reasoning about action, but beyond the scope of this chapter.

1.4.1 Nondeterminism: coin tossing

The nondeterminism of coin tossing is easily represented. Let the signature consist of three constants—$coin_0, coin_1, toss_0$—where the first two constants have domain $\{heads, tails\}$ and the third constant is Boolean. Let T be as follows.

$$\begin{aligned} coin_0 = v &\Leftarrow coin_0 = v \\ toss_0 &\Leftarrow toss_0 \\ \neg toss_0 &\Leftarrow \neg toss_0 \\ coin_1 = v &\Leftarrow coin_1 = v \wedge coin_0 = v \\ coin_1 = v &\Leftarrow coin_1 = v \wedge toss_0 \end{aligned}$$

(Each line above in which v appears represents two causal rules, one for each appropriate value of the metavariable v.) As discussed previously, the first line expresses that causes for initial facts are exogenous; the next two that causes for action occurrences (and nonoccurrences) are exogenous; the fourth that the value of *coin* is inertial. The two rules represented by the fifth line rather resemble the inertia rules, except that they say: if the coin is heads after toss, then there is a cause for this, and, on the other hand, if the coin is tails after toss, then there is a cause for that. One easily verifies that T has six models. In two of them, the coin is not tossed and so remains either heads or tails. In the other four, the coin is initially heads or tails, and after being tossed it is again either heads or tails.

1.4.2 Implied action preconditions: moving an object

There are k Boolean constants $put(v)_0$, for $v \in \{1, \ldots, k\}$ ($k > 1$), along with two additional constants, loc_0, loc_1, whose domains are $\{1, \ldots, k\}$.

$$\begin{aligned} loc_0 = v &\Leftarrow loc_0 = v \\ put(v)_0 &\Leftarrow put(v)_0 \\ \neg put(v)_0 &\Leftarrow \neg put(v)_0 \\ loc_1 = v &\Leftarrow loc_1 = v \wedge loc_0 = v \\ loc_1 = v &\Leftarrow put(v)_0 \end{aligned}$$

(Here, v is a metavariable ranging over $\{1, \ldots, k\}$, so the five lines above represent $5k$ causal rules.) The first three lines express the exogeneity of initial facts and actions in the standard way; the fourth line expresses inertia; the fifth says that putting the object in location v causes it to be there. This causal theory has $k(k+1)$ models: there are k possible initial locations of the block, and for each of these, there are $k+1$ possible continuations—either zero or one of the k put actions occurs, with the appropriate outcome at time 1. Although concurrent actions are, in general, allowed in causal theories, in this case the conflicting outcomes of the k put actions make it impossible to execute two of them at once. It is not necessary to include the $\frac{k(k-1)}{2}$ causal rules that would be required to explicitly state these impossibilities. Instead, they are implied by the description.

1.4.3 Things that change by themselves: falling dominos

We wish to describe the chain reaction of k dominos falling over one after the other, after the first domino is tipped over at time 0. In this description, for simplicity, we will stipulate that all dominos are initially up, and we will describe only the possibility of the tip action occurring at time 0. So the signature consists of tip_0 along with $\mathit{up}(d)_t$ for $d \in \{1, \ldots, k\}$ and $t \in \{0, \ldots, k\}$.

$$
\begin{array}{rcll}
\mathit{up}(d)_0 & \Leftarrow & \top & (1 \leq d \leq k) \\
\mathit{tip}_0 & \Leftarrow & \mathit{tip}_0 & \\
\neg\mathit{tip}_0 & \Leftarrow & \neg\mathit{tip}_0 & \\
\mathit{up}(d)_{t+1} & \Leftarrow & \mathit{up}(d)_{t+1} \wedge \mathit{up}(d)_t & (1 \leq d \leq k, 0 \leq t \leq k-1) \\
\neg\mathit{up}(d)_{t+1} & \Leftarrow & \neg\mathit{up}(d)_{t+1} \wedge \neg\mathit{up}(d)_t & (1 \leq d \leq k, 0 \leq t \leq k-1) \\
\neg\mathit{up}(1)_1 & \Leftarrow & \mathit{tip}_0 & \\
\neg\mathit{up}(d+1)_{t+2} & \Leftarrow & \neg\mathit{up}(d)_{t+1} \wedge \mathit{up}(d)_t & (1 \leq d \leq k-1, 0 \leq t \leq k-2)
\end{array}
$$

The first line says that initially all dominos are up; the second and third that the tip action may or may not occur at time 0; the fourth and fifth lines posit inertia for the dominos' being up or down; and the sixth line describes the direct effect of the tip action (executed at time 0). The seventh and last line is of particular interest. It says that if domino d is up at time t and down at time $t + 1$, then there is a cause for domino $d + 1$ to be down at time $t + 2$. Notice that this rule mentions three successive time points, but no actions. Once the first domino is tipped, the others fall successively with no further action taken. This causal theory has two models. In the first, the tip action does not occur at time 0 and all dominos are up at all times. In the second, all dominos are initially up, and at each time point i ($1 \leq i \leq k$) the ith domino has fallen. (That is, in this model I, for all $t \in \{0, \ldots, k\}$ and $d \in \{1, \ldots, k\}$, $I \models \mathit{up}(d)_t$ iff $d > t$.)

1.4.4 Things that tend to change by themselves: pendulum

So far all examples have postulated commonsense inertia in the standard way: things don't change unless made to. But we can easily take a more general view: some things will change unless made not to. For example, we can describe a pendulum that, if left alone, will oscillate between being on the left and not being on the left. But if held, the pendulum will not change position.

$$
\begin{array}{rcl}
\mathit{left}_0 & \Leftarrow & \mathit{left}_0 \\
\neg\mathit{left}_0 & \Leftarrow & \neg\mathit{left}_0 \\
\mathit{hold}_t & \Leftarrow & \mathit{hold}_t \\
\neg\mathit{hold}_t & \Leftarrow & \neg\mathit{hold}_t \\
\mathit{left}_{t+1} & \Leftarrow & \mathit{left}_{t+1} \wedge \neg\mathit{left}_t \\
\neg\mathit{left}_{t+1} & \Leftarrow & \neg\mathit{left}_{t+1} \wedge \mathit{left}_t \\
\mathit{left}_{t+1} & \Leftarrow & \mathit{hold}_t \wedge \mathit{left}_t \\
\neg\mathit{left}_{t+1} & \Leftarrow & \mathit{hold}_t \wedge \neg\mathit{left}_t
\end{array}
$$

The first four lines express the usual assumptions about exogeneity. The next two lines express that the pendulum will tend to change position from one time to the next. That is, if it changes position between times t and $t + 1$, then there is a cause for its position at time $t+1$. The last two lines say that when the pendulum is held, it will not change position.

In all models of this causal theory, the pendulum changes position between times t and $t+1$ iff it is not held at time t.

1.5 High-level action language $\mathcal{C}+$

High-level action languages feature a concise, restricted syntax for describing the effects of actions, and often benefit from a relatively simple, well-understood semantics, typically defined in terms of a transition system whose nodes are the possible states and whose directed edges are labelled with the actions whose execution in the source state can result in the target state. STRIPS [12] and ADL [37, 38] can be seen as the first high-level action languages.

Language $\mathcal{C}+$ [18] is a descendant of the action language $\mathcal{A}$ [16]. The semantics of $\mathcal{C}+$ is given by reduction to causal theories: for each action description D of $\mathcal{C}+$ and each natural number n, there is a corresponding causal theory $T(D,n)$. The states of the transition system described by D are given by the models of $T(D,0)$, and the possible transitions are given by the models of $T(D,1)$. The paths of length n in the transition system correspond to the models of $T(D,n)$. For instance, the causal theory (1.13) is $T(D,1)$ for the following domain description D

inertial p
exogenous a
a causes $p=1$ if $p=0$
a causes $p=2$ if $p=1$
a causes $p=0$ if $p=2$

where p is designated a "simple fluent constant" with domain $\{0,1,2\}$ and a a Boolean "action constant". Similarly, the causal theory (1.8,1.9) for the soup bowl example corresponds to the domain description

inertial *upL*, *upR*, *sp*
exogenous *liftL*, *liftR*
liftL causes *upL*
liftR causes *upR*
caused *sp* if *upL* $\not\equiv$ *upR*

where *upL*, *upR* and *sp* are Boolean "simple fluent constants" and *liftL* and *liftR* are Boolean "action constants". Here, as elsewhere, the high-level action language provides an especially nice syntax for representing action domains. Indeed, many of the published applications of the Causal Calculator use $\mathcal{C}+$, or its immediate predecessor $\mathcal{C}$ [17], as the "input language".

1.6 Relationship to default logic

A causal theory of a Boolean signature can be viewed as a default theory in the sense of Reiter [41]. (The syntax and semantics of propositional default theories are reviewed in Chapter 6.) Let us agree to identify a causal rule $\phi \Leftarrow \psi$ with the default

$$\frac{:\ \psi}{\phi}\ .$$

In the statement of the following theorem, we identify an interpretation I with the set of formulas satisfied by I.

Default Logic Theorem Let T be a causal theory of a Boolean signature. An interpretation I is a model of T iff I is an extension of T in the sense of default logic.

This theorem shows that causal rules are essentially prerequisite-free defaults with a single justification, so long as we are interested only in those extensions that correspond to interpretations (that is to say, in the extensions that are consistent and complete).

For instance, the causal theory

$$\{p \Leftarrow q,\ q \Leftarrow q,\ \neg q \Leftarrow \neg q\} \tag{1.14}$$

corresponds to the default theory

$$\left\{ \frac{:\ q}{p},\ \frac{:\ q}{q},\ \frac{:\ \neg q}{\neg q} \right\},$$

which has two extensions: the set of all consequences of p, q, and the set of all consequences of $\neg q$. The first extension is complete, and corresponds to the only model of (1.14).

Remarkably, the causal rules (1.1) used in the solution to the frame problem that has been adopted in causal theories were, in essence, proposed already in Reiter's original 1980 paper on default logic. But an account of how to successfully "use" such default rules to express commonsense inertia in reasoning about action did not appear until [43, 44]. (This use of such default rules was derived from a similar application in the setting of knowledge update [39, 40].) In this connection, it may be helpful to mention the historically important "Yale Shooting" paper of Hanks and McDermott [19], who argued that neither default logic nor circumscription were suitable formalisms for reasoning about action. The Yale Shooting paper considered a rather different attempt to solve the frame problem in default logic, and demonstrated that *that attempt* was unsatisfactory. A brief account of this appears in [26]. For more detail, see [44].

1.7 Causal theories in higher-order classical logic

Lifschitz [24] extended causal theories to the nonpropositional case using higher-order classical logic to express the fixpoint condition on a model, much in the manner of circumscription. His definition can be understood to provide a general method for translating finite Boolean causal theories into classical propositional logic.

Begin with a signature of classical logic, with a finite subset of the nonlogical constants designated "explainable". A *nonpropositional causal rule* is an expression of the form

$$\phi \Leftarrow \psi$$

where ϕ and ψ are classical formulas. A *nonpropositional causal theory* is a finite set of nonpropositional causal rules.

The special case of nonpropositional causal theories in which all nonlogical constants are explainable propositional constants coincides with the special case of the previously defined (finite-domain propositional) causal theories in which all constants are Boolean and causal theories are assumed to be finite.

In what follows, let $\overline{N}$ be a list of all explainable nonlogical constants. We say that a list of nonlogical constants or variables is *similar* to $\overline{N}$ if it has the same length as $\overline{N}$ and each of its members is of the same sort as the corresponding member of $\overline{N}$. We can denote a formula (in which none, some, or all explainable nonlogical constants appear) by $\phi(\overline{N})$. Then for any list $\overline{M}$ that is similar to $\overline{N}$, we can write $\phi(\overline{M})$ to denote the formula obtained by simultaneously replacing each occurrence of each member of $\overline{N}$ by the corresponding member of $\overline{M}$.

Consider a nonpropositional causal theory T with rules

$$\begin{array}{rcl} \phi_1(\overline{N}, \overline{x_1}) & \Leftarrow & \psi_1(\overline{N}, \overline{x_1}) \\ & \vdots & \\ \phi_k(\overline{N}, \overline{x_k}) & \Leftarrow & \psi_k(\overline{N}, \overline{x_k}) \end{array}$$

where $\overline{x_i}$ is the list of all free variables for the ith causal rule. Let $\overline{n}$ be a list of new variables that is similar to $\overline{N}$. By $T^*(\overline{n})$ we denote the formula

$$\bigwedge_{1 \leq i \leq k} \forall \overline{x_i} \left(\psi_i(\overline{N}, \overline{x_i}) \supset \phi_i(\overline{n}, \overline{x_i}) \right) .$$

An interpretation is a *model* of T if it is a model of

$$\forall \overline{n} \left(T^*(\overline{n}) \equiv \overline{n} = \overline{N} \right) \tag{1.15}$$

where $\overline{n} = \overline{N}$ stands for the conjunction of the equalities between members of $\overline{n}$ and the corresponding members of $\overline{N}$.

Where the definitions overlap syntactically, this definition of model of a causal theory agrees with the definition given earlier.

Notice that for finite Boolean propositional causal theories the corresponding sentence (1.15) is a quantified Boolean formula, from which quantifiers can be eliminated (with worst-case exponential increase in size). Thus this approach yields a general translation of finite Boolean propositional causal theories into classical propositional logic.

The completion method from Section 1.3 can be extended to "definite" nonpropositional causal theories, so that a "first-order" causal theory T has the same models as the corresponding completion (which is a classical first-order formula) [24].

1.8 A logic of universal causation

UCL is a modal nonmonotonic logic obtained from standard S5 modal logic (see Chapter 15) by imposing a simple fixpoint condition that reflects the "principle of universal causation"—the requirement that everything true in a model have a cause. In [45], UCL was defined not only in the (Boolean) propositional case, but also for nonpropositional languages, and was shown to subsume the nonpropositional causal theories described in the previous section. Here, we consider a different extension of (Boolean) propositional UCL, introduced in [46], built from finite-domain propositional formulas.

The fundamental distinction in UCL—between propositions that have a cause and propositions that (merely) obtain—is expressed by means of the modal operator C, read as "caused." For example, one can write

$$\psi \supset \mathsf{C}\phi \tag{1.16}$$

to say that ϕ is caused whenever ψ obtains. UCL formula (1.16) corresponds to the causal rule $\phi \Leftarrow \psi$. This claim is made precise in the UCL Theorem below.

UCL formulas are obtained by extending the recursive definition of a formula with an additional case for the modal operator C, in the usual way for modal logic:

- If ϕ is a UCL formula, then so is $\mathsf{C}\phi$.

A *UCL theory* is a set of UCL formulas.

An *S5-structure* is a pair (I, S) such that I is an interpretation and S is a set of interpretations (all of the same signature) to which I belongs. *Satisfaction* of a UCL formula by an S5-structure is defined by the standard recursions over the propositional connectives, plus the following two conditions:

- if ϕ is an atom, $(I, S) \models \phi$ iff $I \models \phi$,
- $(I, S) \models \mathsf{C}\phi$ iff for all $J \in S$, $(J, S) \models \phi$.

For a UCL theory T, if $(I, S) \models T$, we say that (I, S) is an *I-model* of T, thus emphasizing the distinguished interpretation I.

We say that I is *causal model* of T if $(I, \{I\})$ is the unique I-model of T.

UCL Theorem For any causal theory T, the models of T are precisely the causal models of the corresponding UCL theory

$$\{\, \psi \supset \mathsf{C}\phi \,:\, \phi \Leftarrow \psi \in T \,\} \ .$$

It is possible to characterize strong equivalence for UCL theories, much as was done in Section 1.2 for causal theories. Interestingly, this requires a slight strengthening of S5. The SE-models of a UCL theory are a subset of the S5 models: those S5 models (I, S) such that $(I, \{I\})$ is also an S5 model [46].

Bibliography

[1] Akman, V.; Erdoğan, S.; Lee, J.; Lifschitz, V.; and Turner, H. 2004. Representing the Zoo World and the Traffic World in the language of the Causal Calculator. *Artificial Intelligence* 153:105–140.

[2] Artikis, A.; Sergot, M.; and Pitt, J. 2003a. An executable specification of an argumentation protocol. In *Proc. of Artificial Intelligence and Law (ICAIL)*, 1–11.

[3] Artikis, A.; Sergot, M.; and Pitt, J. 2003b. Specifying electronic societies with the Causal Calculator. In Giunchiglia, F.; Odell, J.; and Weiss, G., eds., *Proc. of Workshop on Agent-Oriented Software III (AOSE)*, 1–15. Springer LNCS 2585.

[4] Bochman, A. 2003. A logic for causal reasoning. In *Proc. IJCAI-03*, 141–146.

[5] Bochman, A. 2004. A causal approach to nonmonotonic reasoning. *Artificial Intelligence* 160(1–2):105–143.

[6] Campbell, J., and Lifschitz, V. 2003. Reinforcing a claim in commonsense reasoning. In *Logical Formalizations of Commonsense Reasoning: Papers from 2003 AAAI Spring Symposium*, 51–56.

[7] Chopra, A., and Singh, M. 2004. Nonmonotonic committment machines. In Dignum, F., ed., *Workshop on Agent Communication Languages*, volume 2922 of *Lecture Notes in Computer Science*, 183–200. Springer.

[8] Clark, K. 1978. Negation as failure. In Gallaire, H., and Minker, J., eds., *Logic and Data Bases*. New York: Plenum Press. 293–322.

[9] Craven, R., and Sergot, M. 2005. Distant causation in C+. *Studia Logica* 79:73–96.

[10] Dogandag, S.; Ferraris, P.; and Lifschitz, V. 2004. Almost definite causal theories. In *Logic Programming and Nonmonotonic Reasoning: Proc. of Seventh Int'l Conf.*, 74–86.

[11] Eiter, T., and Lukasiewicz, T. 2003. Probabilistic reasoning about actions in nonmonotonic causal theories. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI-2003)*, 192–199.

[12] Fikes, R., and Nilsson, N. 1971. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence* 2(3–4):189–208.

[13] Finzi, A., and Eiter, T. 2005. Game-theoretic reasoning about actions in nonmontonic causal theories. In *Logic Programming and Nonmonotonic Reasoning: 8th International Conference*, 185–197. Springer LNCS 3662.

[14] Geffner, H. 1990. Causal theories of nonmonotonic reasoning. In *Proc. of AAAI-90*, 524–530.

[15] Geffner, H. 1992. *Reasoning with defaults: causal and conditional theories*. Cambridge, MA: MIT Press.

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

[17] Giunchiglia, E., and Lifschitz, V. 1998. An action language based on causal explanation: Preliminary report. In *Proc. AAAI-98*, 623–630.

[18] Giunchiglia, E.; Lee, J.; Lifschitz, V.; McCain, N.; and Turner, H. 2004. Nonmonotonic causal theories. *Artificial Intelligence* 153(1&2):49–104.

[19] Hanks, S., and McDermott, D. 1987. Nonmonotonic logic and temporal projection. *Artificial Intelligence* 33(3):379–412.

[20] Lee, J., and Lifschitz, V. 2003. Describing additive fluents in action language C+. In *Proc. IJCAI'03*, 1079–1084.

[21] Lee, J. 2004. Definite vs. nondefinite causal theories. In *Logic Programming and Nonmonotonic Reasoning: Proc. of Seventh Int'l Conf.*, 141–153.

[22] Lifschitz, V.; McCain, N.; Remolina, E.; and Tacchella, A. 2000. Getting to the airport: the oldest planning problem in AI. In *Logic-Based Artificial Intelligence*. Kluwer. 147–165.

[23] Lifschitz, V.; Pearce, D.; and Valverde, A. 2001. Strongly equivalent logic programs. *ACM Transactions on Computational Logic* 2:526–541.

[24] Lifschitz, V. 1997. On the logic of causal explanation. *Artificial Intelligence* 96:451–465.

[25] Lifschitz, V. 1998. Situation calculus and causal logic. In *Proc. of the Sixth Int'l Conf. on Principles of Knowledge Representation and Reasoning*, 536–546.

[26] Lifschitz, V. 1999. Success of default logic. In Levesque, H., and Pirri, F., eds., *Logical Foundations of Cognitive Agents: Contributions in Honor of Ray Reiter*. Springer-Verlag. 208–212.

[27] Lifschitz, V. 2000. Missionaries and cannibals in the Causal Calculator. In *Proc. of the 7th Int'l Conf. on Principles of Knowledge Representation and Reasoning*, 85–96.

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

[29] McCain, N., and Turner, H. 1995. A causal theory of ramifications and qualifications. In *Proc. of IJCAI-95*, 1978–1984.

[30] McCain, N., and Turner, H. 1997. Causal theories of action and change. In *Proc. of AAAI-97*, 460–465.

[31] McCain, N., and Turner, H. 1998. Satisfiability planning with causal theories. In *Principles of Knowledge Representation and Reasoning: Proc. of the Sixth Int'l Conference*, 212–223.

[32] McCain, N. 1997. *Causality in Commonsense Reasoning about Actions*. Ph.D. Dissertation, University of Texas at Austin, Department of Computer Sciences.

[33] McCarthy, J., and Hayes, P. 1969. Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B., and Michie, D., eds., *Machine Intelligence*, volume 4. Edinburgh: Edinburgh University Press. 463–502. Reproduced in [34].

[34] McCarthy, J. 1990. *Formalizing common sense: papers by John McCarthy*. Norwood, NJ: Ablex.

[35] Pearl, J. 1988. Embracing causality in default reasoning. *Artificial Intelligence* 35:259–271.

[36] Pearl, J. 2000. *Causality: Models, reasoning and inference*. Cambridge University Press.

[37] Pednault, E. 1989. ADL: Exploring the middle ground between STRIPS and the situation calculus. In Brachman, R.; Levesque, H.; and Reiter, R., eds., *Proc. of the First Int'l Conf. on Principles of Knowledge Representation and Reasoning*, 324–332.

[38] Pednault, E. 1994. ADL and the state-transition model of action. *Journal of Logic and Computation* 4:467–512.

[39] Przymusinski, T., and Turner, H. 1995. Update by means of inference rules. In *Proc. of the 3rd Int'l Conf. on Logic Programming and Nonmonotonic Reasoning*, 156–174.

[40] Przymusinski, T., and Turner, H. 1997. Update by means of inference rules. *Journal of Logic Programming* 30(2):125–143.

[41] Reiter, R. 1980. A logic for default reasoning. *Artificial Intelligence* 13(1,2):81–132.

[42] Sergot, M., and Craven, R. 2005. Some logical properties of nonmonotonic causal theories. In *Logic Programming and Nonmonotonic Reasoning: 8th International Conference*, 198–210. Springer LNCS 3662.

[43] Turner, H. 1996. Representing actions in default logic: A situation calculus approach. In *Working Papers of the Third Symposium on Logical Formalizations of Commonsense Reasoning*.

[44] Turner, H. 1997. Representing actions in logic programs and default theories: A situation calculus approach. *Journal of Logic Programming* 31(1–3):245–298.

[45] Turner, H. 1999. A logic of universal causation. *Artificial Intelligence* 113:87–123.

[46] Turner, H. 2004. Strong equivalence for causal theories. In *Logic Programming and Nonmonotonic Reasoning: Proc. of Seventh Int'l Conf.*, 289–301.

[47] White, G. 2002. A modal formulation of McCain and Turner's theory of causal reasoning. In *Logics in Artificial Intelligence: Proc. 8th European Conference (JELIA'02)*, 211–222.