On the Logic of Causal Explanation

Vladimir Lifschitz
Department of Computer Sciences
University of Texas at Austin
Austin, TX 78712

Abstract

The McCain-Turner semantics of causal rules is based on a fixpoint construction similar to the one found in the definition of default logic. In the special case when the heads of the rules are literals, it can be equivalently expressed by a translation from sets of rules into sets of propositional formulas. In this note, we define a translation from causal logic into classical logic that characterizes the semantics of arbitrary causal rules, without any restrictions on their syntactic form. This translation suggests a way to extend the McCain-Turner logic to nonpropositional causal theories.

Keywords: Actions; Causal logic; Circumscription

1 Introduction

The nonmonotonic logic of causal explanation proposed by Norman McCain and Hudson Turner [1997a] provides a simple and powerful formalism for representing properties of actions. Their work builds on the analysis of causality due to Hector Geffner [1990], on work by Fangzhen Lin [1995] and on their own research described in [McCain and Turner, 1995]. The main idea of this theory is to distinguish between the claim that a proposition is true and the stronger claim that there is a *cause* for it to be true. Causal dependencies are described by "causal rules" of the form

$$F \leftarrow G \tag{1}$$

where F and G are propositional formulas.[1] Rule (1) expresses that F has a cause if G is true, or that G provides a "causal explanation" for F.

The definition of the semantics of causal rules (reproduced in Section 7 below) uses a fixpoint construction similar to the one familiar from the definition of an extension in default logic [Reiter, 1980]. In the special case when the heads of the rules are literals, the semantics can be equivalently expressed by a translation from sets of rules into sets of propositional formulas. This translation, called "literal completion," is similar to the completion process used in logic programming [Clark, 1978]. In [McCain and Turner,

[1]The syntax used in [McCain and Turner, 1997a] is $G \Rightarrow F$.

1997b], literal completion is used to apply satisfiability planning [Kautz and Selman, 1992] to action domains described by causal theories.

In this note, we define a translation from causal logic into classical logic that characterizes the semantics of *arbitrary* causal rules, without any restrictions on their syntactic form. This translation opens the possibility of applying satisfiability planning to causal theories with arbitrary formulas in the heads of rules. Furthermore, the new approach to the semantics of causal theories suggests a way to extend the McCain-Turner definition to nonpropositional rules. The treatment of free variables in nonmonotonic formalisms defined by fixpoint conditions is a notoriously difficult problem,[2] and the use of a translational semantics is crucial for overcoming this difficulty. This extension of the McCain-Turner system allows us to describe causal dependencies in the domains characterized by nonpropositional theories. In the new framework, large (or even infinite) propositional causal theories can be sometimes replaced by compact nonpropositional representations.

The translation that we propose may introduce second-order quantifiers, but in many cases of interest these quantifiers can be eliminated.[3] An elimination process leads, in particular, to an extension of the process of literal completion to nonpropositional causal theories.

The McCain-Turner semantics is based on the idea that a causal theory should "explain" the truth value of every propositional symbol in its language. They characterize this "principle of universal causation" as a rather strong philosophical commitment that is rewarded by the mathematical simplicity of the theory. In the modification described below, we distinguish between "explainable" symbols—those whose values must be causally explained by the given rules—and "nonexplainable" ones.

The formulation of causal logic presented in this note serves as the basis of a new action language in the style of [Gelfond and Lifschitz, 1993], in which the concurrent execution of actions can be described, and both actions and fluents can be represented by expressions with parameters. This is the subject of a forthcoming paper.

The definition and examples of nonpropositional causal theories are discussed in the next two sections. In Section 4 we show how the effects of actions over an infinite set of time instants can be described by a finite collection of causal rules with variables. A nonpropositional version of literal completion is introduced in Section 5, and its correctness is proved in Section 6. Finally, in Section 7 we show that the McCain-Turner semantics of causal rules is indeed equivalent to a special case of the semantics proposed in this note.

---

[2] It is known, for instance, that the semantics of open defaults proposed in [Reiter, 1980] leads to serious difficulties [Baader and Hollunder, 1995].

[3] Quantifiers over propositional variables can be always eliminated, although this can make the formula significantly longer. There exists an implementation of a process of eliminating quantifiers over predicate variables based on the ideas of [Ackermann, 1935] (`http://www.ida.liu.se/labs/kplab/projects/dls/`). An implementation of another algorithm for the elimination of second-order quantifiers can be found at `http://www.mpi-sb.mpg.de/guide/staff/ohlbach/scan/scan.html` .

2 Causal Theories

Consider any language of classical logic (propositional, first-order or higher-order, possibly many-sorted). A *causal rule* is an expression of form (1), where F and G are formulas of this language, called the *head* and the *body* of the rule. If the body of a rule is the logical constant *True*, we will drop it and write the rule as $F \leftarrow$.

A *causal theory* is defined by

- a finite set of nonlogical constants[4] of the underlying language, called the *explainable symbols* of the theory, and
- a finite set of causal rules.

In the definition of the semantics of causal theories below, we use the substitution of variables for the explainable symbols in a formula. In connection with this, it is convenient to denote formulas by expressions like $F(E)$, where E is the list of all explainable symbols. Then, for any tuple e of variables that is similar[5] to E, the result of replacing all occurrences of the constants E in $F(E)$ by the variables e can be denoted by $F(e)$. The use of the notation $F(E)$ for a formula does not presume that the formula actually contains all explainable symbols; it may even contain no explainable symbols at all, in which case $F(e)$ equals $F(E)$.

Consider a causal theory T with the explainable symbols E and the causal rules

$$F_i(E, x^i) \leftarrow G_i(E, x^i) \qquad (i = 1, \ldots),$$

where x^i is the list of all free variables of the i-th rule. Take a tuple e of new variables similar to E. By $T^*(e)$ we denote the formula

$$\bigwedge_i \forall x^i (G_i(E, x^i) \supset F_i(e, x^i)).$$

Note that the occurrences of explainable symbols in the heads are replaced here by variables, and the occurrences in the bodies are not. We will view T as shorthand for the sentence

$$\forall e(T^*(e) \equiv e = E). \tag{2}$$

(The expression $e = E$ stands for the conjunction of the equalities between the members of e and the corresponding members of E.) For instance, by a *model* of T we mean a model of (2); a *theorem* of T is a sentence that is entailed by (2).

Intuitively, the condition $T^*(e)$ expresses that the possible values e of the explainable symbols E are "explained" by the rules of T. Sentence (2) says that the actual values of these symbols are the only ones that are explained by the rules of T.

[4]A *nonlogical constant* is a function constant or a predicate constant (other than equality). This includes, in particular, object constants (function constants of arity 0) and propositional constants (predicate constants of arity 0).

[5]The similarity condition means that (i) e has the same length as E, (ii) if the k-th member of E is a function constant then the k-th member of e is a function variable of the same type, and (iii) if the k-th member of E is a predicate constant then the k-th member of e is a predicate variable of the same type.

Formula (2) can be equivalently written as

$$T^*(E) \wedge \forall e(T^*(e) \supset e = E). \tag{3}$$

In this expression, the term $T^*(E)$ is the conjunction of the universal closures of the implications

$$G_i(E, x^i) \supset F_i(E, x^i)$$

corresponding to the rules of T.

3 Examples

Example 1. The only causal rule of T is

$$P(x) \leftarrow Q(x), \tag{4}$$

where P and Q are unary predicate constants, and P is explainable. Then $T^*(p)$ is

$$\forall x(Q(x) \supset p(x)),$$

where p is a unary predicate variable, and (2) turns into

$$\forall p(\forall x(Q(x) \supset p(x)) \equiv p = P).$$

It is easy to check that this second-order formula is equivalent to

$$\forall x P(x) \wedge \forall x Q(x). \tag{5}$$

What is the intuitive meaning of this result? Since the predicate constant P is declared explainable, the value of P at every point x should be "explained" by the causal rules of T. The only rule of T explains $P(x)$, but it does not explain $\neg P(x)$. So all values of P can be explained only if P is identically true. Furthermore, this rule explains $P(x)$ only for the values of x that satisfy Q. So all values of P can be explained only if Q is identically true as well.

Example 2. The causal rules of T are (4) and

$$\neg P(x) \leftarrow \neg P(x), \tag{6}$$

where P and Q are as in Example 1. Now $T^*(p)$ is

$$\forall x(Q(x) \supset p(x)) \wedge \forall x(\neg P(x) \supset \neg p(x)),$$

and (2) becomes

$$\forall p\{[\forall x(Q(x) \supset p(x)) \wedge \forall x(\neg P(x) \supset \neg p(x))] \equiv p = P\}.$$

Proposition 1 below shows that this causal theory is equivalent to

$$\forall x(P(x) \equiv Q(x)). \tag{7}$$

Intuitively, rule (6) says that, whenever P is false at a point x, there is a cause for this. If this rule is included in a causal theory then the other rules should allow us to explain every case when P is true. Since $Q(x)$ is the only possible explanation for $P(x)$, P turns out to be equivalent to Q.

Note that (5) is a theorem of the theory from Example 1, but not a theorem of the theory obtained from it by adding rule (6). This fact illustrates the nonmonotonic character of causal theories. The nonmonotonicity of the formalism is not surprising, because the formulas representing the rules of T occur negatively in the second conjunctive term of (3). This term is similar, in this sense, to the minimality condition in the definition of circumscription.

Including rule (6) in a causal theory is similar to postulating that P is "false by default," or, in other words, to circumscribing P. The following proposition makes this precise:

Proposition 1. *A causal theory of the form*

$$\begin{array}{l} F_i \leftarrow \quad (1 \leq i \leq m), \\ \neg P(x) \leftarrow \neg P(x), \end{array}$$

where P is the only explainable symbol of the theory and x is a tuple of distinct variables, is equivalent to the circumscription of P in $\bigwedge_{i=1}^{m} \widetilde{\forall} F_i$.[6]

To apply this proposition to Example 2, observe first that (4) can be rewritten as

$$Q(x) \supset P(x) \leftarrow$$

without changing the formula $T^*(p)$. It follows that the causal theory of Example 2 is equivalent to the circumscription of P in $\forall x(Q(x) \supset P(x))$, that is, to (7).

Proof of Proposition 1. Denote the given causal theory by T, and let $F(P)$ stand for $\bigwedge_{i=1}^{m} \widetilde{\forall} F_i$. Then

$$T^*(p) \equiv F(p) \wedge \forall x(\neg P(x) \supset \neg p(x)) \equiv F(p) \wedge (p \leq P)$$

(for notation, see [Lifschitz, 1994], Section 2.1), and consequently

$$T \equiv F(P) \wedge \forall p(F(p) \wedge (p \leq P) \supset p = P) \equiv F(P) \wedge \forall p(p < P \supset \neg F(p)).$$

Examples 1 and 2 are different from causal theories in the sense of [McCain and Turner, 1997a] in two ways. First, they are nonpropositional. Second, in these examples we distinguish between the explainable symbols (P) and the nonexplainable symbols (Q); as pointed out in the introduction, McCain and Turner require a causal explanation for *every* symbol. Technically, the availability of nonexplainable symbols in our version of the

[6] By $\widetilde{\forall} F$ we denote the universal closure of F.

theory is not very essential: Reclassifying a predicate constant Q as explainable could be neutralized by adding the causal rules

$$\begin{array}{l} Q(x) \leftarrow Q(x), \\ \neg Q(x) \leftarrow \neg Q(x), \end{array} \tag{8}$$

which say that, whatever the truth value of $Q(x)$ is, there is a cause for this formula to have that value. More precisely:

Proposition 2. *Let T be a causal theory containing rules (8), where Q is an explainable predicate constant and x is a tuple of distinct variables. The causal theory obtained from T by deleting Q from the set of explainable symbols and dropping rules (8) is equivalent to T.*

For instance, the causal theory whose rules are

$$\begin{array}{l} P(x) \leftarrow Q(x), \\ \neg P(x) \leftarrow \neg P(x), \\ Q(x) \leftarrow Q(x), \\ \neg Q(x) \leftarrow \neg Q(x), \end{array} \tag{9}$$

where both P and Q are explainable, is equivalent to the theory of Example 2.

Proof of Proposition 2. Let E be the list of all explainable symbols other than Q, and let T_1 be the causal theory obtained from T by deleting Q from the set of explainable symbols and dropping rules (8). Then

$$\begin{array}{rcl} T & \equiv & T^*(E,Q) \wedge \forall eq(T^*(e,q) \supset (e = E \wedge q = Q)) \\ & \equiv & T_1^*(E,Q) \wedge \forall eq[(T_1^*(e,q) \wedge \forall x(Q(x) \supset q(x)) \wedge \forall x(\neg Q(x) \supset \neg q(x))) \\ & & \qquad \supset (e = E \wedge q = Q)] \\ & \equiv & T_1^*(E,Q) \wedge \forall eq((T_1^*(e,q) \wedge q = Q) \supset (e = E \wedge q = Q)) \\ & \equiv & T_1^*(E,Q) \wedge \forall e(T_1^*(e,Q) \supset e = E) \\ & \equiv & T_1. \end{array}$$

4 Describing Actions by Causal Rules

Example 2 has a simple interpretation in terms of actions and their effects. Imagine that $Q(x)$ represents an action (more precisely, the assertion that a certain action is executed during a particular interval of time), and that $P(x)$ represents a fluent (more precisely, the assertion that a certain truth-valued fluent holds after performing the action). Then rule (4) tells us that, after performing the action, there is a cause for the fluent to hold. Rule (6) expresses that this fluent is "false by default"—that it is "momentary" in the sense of [Lifschitz and Rabinov, 1989].

This interpretation is appropriate, for instance, if $Q(x)$ represents plucking string x on the guitar, and $P(x)$ means that string x sounds. Conclusion (7), sanctioned by the semantics of causal rules introduced above, reflects the commonsense fact that a string sounds if and only if it has just been plucked. This applies, in particular, to the case when $Q(x)$ holds for more than one value of x, that is, when several strings are plucked simultaneously.

The next example differs from Example 2 in two ways. First, instead of momentary fluents, we consider fluents governed by the commonsense law of inertia. Second, instead of considering the effect of a single action, we trace the evolution of the fluents along an infinite sequence of instants of time. The language will include a second sort of variables, for natural numbers, and each of the predicates P, Q will get a second argument of this new sort. We will follow the style of formalization proposed in [McCain and Turner, 1997a], except that the availability of variables will allow us to replace an infinite set of axioms by a finite one.

Example 3. The causal rules of T are

$$
\begin{array}{l}
n' \neq 0 \leftarrow\ , \\
m' = n' \supset m = n \leftarrow\ , \\
p(0) \wedge \forall m(p(m) \supset p(m')) \supset p(n) \leftarrow\ , \\
P(x, n') \leftarrow Q(x, n), \\
P(x, n') \leftarrow P(x, n) \wedge P(x, n'), \\
\neg P(x, n') \leftarrow \neg P(x, n) \wedge \neg P(x, n'), \\
P(x, 0) \leftarrow P(x, 0), \\
\neg P(x, 0) \leftarrow \neg P(x, 0),
\end{array}
$$

and P is explainable. The first three rules characterize the system of natural numbers by the standard Peano axioms, using the predicate variable p to express induction. The fourth rule describes the effect of Q on P; $P(x, n)$ expresses that $P(x)$ holds at time n, and $Q(x, n)$ expresses that the action $Q(x)$ is executed between times n and n'. The next pair of rules expresses the law of inertia for P by saying that there is a cause for $P(x)$ to be true (false) at time n' if it is true (respectively, false) both at time n and at time n'. The last two rules are similar to (8); their effect is to obviate the need to explain the values of P at the initial instant of time.

In the next section we will see that this theory is equivalent to the conjunction of the Peano axioms (that is, the heads of the first three rules) and the condition

$$P(x, n') \equiv P(x, n) \vee Q(x, n). \tag{10}$$

5 Literal Completion

Consider a causal theory T such that every explainable symbol of T is a predicate constant. A causal rule is *definite* relative to T if its head

- is a literal that does not contain explainable symbols of T in argument positions, or
- does not contain explainable symbols of T.

A causal theory T is *definite* if every rule of T is a definite rule.

The causal theories from Examples 1–3 are definite. The need to have formulas more complex than literals in the heads of causal rules arises, for instance, when an explicitly defined symbol is added to the language [McCain and Turner, 1997a]; a definitional extension of a causal theory is not definite.

The process of completion described below turns any definite causal theory into an equivalent set of sentences of classical logic. If the heads and the bodies of all rules of the theory are first-order formulas, as in Examples 1 and 2, then the resulting sentences are first-order also. Moreover, we will see that even in Example 3, where one of the rules includes a predicate variable, completion does not introduce any new higher-order variables and thus leads to a simple result.

The main new feature of literal completion in comparison with the completion process defined in [Clark, 1978] is that the "completed definition" of a predicate constant P consists of *two* equivalences, one "positive" and one "negative."

The algorithm takes any definite causal theory T as input. By P we denote any explainable symbol of T; P^ϵ stands for P if $\epsilon = 1$ and for $\neg P$ if $\epsilon = -1$. Using this notation, we can write any literal whose predicate symbol is explainable as $P^\epsilon(t)$, where t is a tuple of terms. We perform successively the following steps.

Step 1: For each P, take a tuple of new variables x and replace each rule of the form $P^\epsilon(t) \leftarrow G$ by $P^\epsilon(x) \leftarrow (x = t) \wedge G$.

Step 2: Replace each rule $P^\epsilon(x) \leftarrow G$ obtained in the previous step by $P^\epsilon(x) \leftarrow \exists y G$, where y are the free variables of the original rule.

Step 3: For each P and each ϵ, replace all rules $P^\epsilon(x) \leftarrow G_i$ by the rule $P^\epsilon(x) \leftarrow \vee_i G_i$.

Step 4: For each P and each ϵ, replace the rule $P^\epsilon(x) \leftarrow G$ obtained in the previous step by the sentence $\widetilde{\forall}(P^\epsilon(x) \equiv G)$.

Step 5: Replace each remaining rule $F \leftarrow G$ by the sentence $\widetilde{\forall}(G \supset F)$.

The sentences obtained in this way are called the *completion sentences* for T.

Completion Theorem. *Every definite causal theory is equivalent to the conjunction of its completion sentences.*

In the examples of literal completion below, we use classical logic to simplify the bodies of the rules after some of the steps. The formulas obtained at the end of this process are logically equivalent to the completion sentences.

Example 1, continued. The result of Step 1 is

$$P(x_1) \leftarrow x_1 = x \wedge Q(x).$$

Step 2 leads to

$$P(x_1) \leftarrow \exists x(x_1 = x \wedge Q(x)),$$

which can be rewritten as

$$P(x_1) \leftarrow Q(x_1).$$

Step 3 adds the second rule

$$\neg P(x_1) \leftarrow \mathit{False}.$$

Step 4 produces the equivalences

$$\begin{array}{l} \forall x_1(P(x_1) \equiv Q(x_1)), \\ \forall x_1(\neg P(x_1) \equiv \mathit{False}). \end{array}$$

It is clear that the conjunction of these formulas is logically equivalent to (5).

Example 2, continued. After the first two steps, we get

$$\begin{array}{l} P(x_1) \leftarrow Q(x_1), \\ \neg P(x_1) \leftarrow \neg P(x_1). \end{array}$$

Step 3 does not change these rules, and the result of Step 4 is

$$\begin{array}{l} \forall x_1(P(x_1) \equiv Q(x_1)), \\ \forall x_1(\neg P(x_1) \equiv \neg P(x_1)). \end{array}$$

It is clear that the conjunction of these formulas is logically equivalent to (7).

Example 3, continued. After Steps 1 and 2, we get

$$\begin{array}{l} n' \neq 0 \leftarrow \,, \\ m' = n' \supset m = n \leftarrow \,, \\ p(0) \land \forall m(p(m) \supset p(m')) \supset p(n) \leftarrow \,, \\ P(x_1, m_1) \leftarrow \exists n(m_1 = n' \land Q(x_1, n)), \\ P(x_1, m_1) \leftarrow \exists n(m_1 = n' \land P(x_1, n) \land P(x_1, n')), \\ \neg P(x_1, m_1) \leftarrow \exists n(m_1 = n' \land \neg P(x_1, n) \land \neg P(x_1, n')), \\ P(x_1, m_1) \leftarrow m_1 = 0 \land P(x_1, 0), \\ \neg P(x_1, m_1) \leftarrow m_1 = 0 \land \neg P(x_1, 0), \end{array}$$

and Step 3 gives

$$\begin{array}{l} n' \neq 0 \leftarrow \,, \\ m' = n' \supset m = n \leftarrow \,, \\ p(0) \land \forall m(p(m) \supset p(m')) \supset p(n) \leftarrow \,, \\ P(x_1, m_1) \leftarrow \exists n(m_1 = n' \land Q(x_1, n)) \\ \qquad\qquad\qquad \lor \exists n(m_1 = n' \land P(x_1, n) \land P(x_1, n')) \\ \qquad\qquad\qquad \lor (m_1 = 0 \land P(x_1, 0)), \\ \neg P(x_1, m_1) \leftarrow \exists n(m_1 = n' \land \neg P(x_1, n) \land \neg P(x_1, n')) \\ \qquad\qquad\qquad \lor (m_1 = 0 \land \neg P(x_1, 0)). \end{array}$$

The completion sentences are

$$\begin{array}{l} PA, \\ \forall x_1 m_1 [P(x_1, m_1) \equiv \exists n(m_1 = n' \land Q(x_1, n)) \\ \qquad\qquad\qquad \lor \exists n(m_1 = n' \land P(x_1, n) \land P(x_1, n')) \\ \qquad\qquad\qquad \lor (m_1 = 0 \land P(x_1, 0))], \\ \forall x_1 m_1 [\neg P(x_1, m_1) \equiv \exists n(m_1 = n' \land \neg P(x_1, n) \land \neg P(x_1, n')) \\ \qquad\qquad\qquad \lor (m_1 = 0 \land \neg P(x_1, 0))], \end{array}$$

where PA stands for the Peano axioms

$$\begin{array}{l}\forall n(n' \neq 0),\\ \forall mn(m' = n' \supset m = n),\\ \forall pn[p(0) \wedge \forall m(p(m) \supset p(m')) \supset p(n)].\end{array}$$

The fact that every number other than 0 has a unique successor, which is a logical consequence of PA, can be used to rewrite the conjunction of the last two completion sentences as the universal closure of (10).

6 Proof of the Completion Theorem

The proof of the completion theorem is based on the fact that a definite theory can be decomposed into "simple" parts. The decomposition process is of interest in its own right, because it can be also used to simplify some theories that are not definite.

About causal theories T_1, T_2 with sets E_1, E_2 of explainable symbols we say that they are *disjoint* if

- E_1 and E_2 are disjoint sets,
- the symbols in E_1 do not occur in the heads of the rules of T_2, and the symbols in E_2 do not occur in the heads of the rules of T_1.

For any pairwise disjoint causal theories $T_1, \dots, T_m$, define their *union* to be the causal theory obtained by combining their rules and their explainable symbols.

Lemma 1. *The union of pairwise disjoint causal theories $T_1, \dots, T_m$ is equivalent to the conjunction $T_1 \wedge \cdots \wedge T_m$.*

Proof:

$$\begin{array}{l}T_1(E_1) \wedge \cdots \wedge T_m(E_m)\\ \qquad \wedge \forall e_1 \cdots e_m((T_1(e_1) \wedge \cdots \wedge T_m(e_m)) \supset (e_1 = E_1 \wedge \cdots \wedge e_m = E_m))\\ \equiv T_1(E_1) \wedge \cdots \wedge T_m(E_m)\\ \qquad \wedge \forall e_1 \cdots e_m((T_1(e_1) \wedge \cdots \wedge T_m(e_m)) \supset e_1 = E_1)\\ \qquad \wedge \cdots\\ \qquad \wedge \forall e_1 \cdots e_m((T_1(e_1) \wedge \cdots \wedge T_m(e_m)) \supset e_m = E_m)\\ \equiv T_1(E_1) \wedge \cdots \wedge T_m(E_m)\\ \qquad \wedge \forall e_1((T_1(e_1) \wedge \exists e_2 T_2(e_2) \wedge \cdots \wedge \exists e_m T_m(e_m)) \supset e_1 = E_1)\\ \qquad \wedge \cdots\\ \qquad \wedge \forall e_m((\exists e_1 T_1(e_1) \wedge \cdots \wedge \exists e_{m-1} T_{m-1}(e_{m-1}) \wedge T_m(e_m)) \supset e_m = E_m)\\ \equiv T_1(E_1) \wedge \cdots \wedge T_m(E_m)\\ \qquad \wedge \forall e_1(T_1(e_1) \supset e_1 = E_1) \wedge \cdots \wedge \forall e_m(T_m(e_m) \supset e_m = E_m)\\ \equiv T_1 \wedge \cdots \wedge T_m.\end{array}$$

A causal theory T is *simple* if it consists of two rules

$$\begin{array}{l}P(x) \leftarrow G_1(x),\\ \neg P(x) \leftarrow G_2(x)\end{array} \tag{11}$$

such that

- P is the only explainable symbol of T,
- x is a tuple of distinct variables,
- $G_1(x)$, $G_2(x)$ are formulas containing no free variables other than the members of x.

For instance, the theory from Example 2 is simple. It is clear that every simple theory is definite. The following lemma shows that any simple theory is equivalent to the pair of sentences obtained from it by applying Step 4 of the completion process.

Lemma 2. *Any simple theory (11) is equivalent to*

$$\forall x(P(x) \equiv G_1(x)) \wedge \forall x(\neg P(x) \equiv G_2(x)). \tag{12}$$

Proof. For a simple theory (11), sentence (3) turns into

$$\begin{array}{l}\forall x(G_1(x) \supset P(x)) \wedge \forall x(G_2(x) \supset \neg P(x))\\ \qquad \wedge \forall p[(\forall x(G_1(x) \supset p(x)) \wedge \forall x(G_2(x) \supset \neg p(x))) \supset p = P],\end{array}$$

or, equivalently,

$$\begin{array}{l}\forall x(G_1(x) \supset P(x)) \wedge \forall x(G_2(x) \supset \neg P(x))\\ \wedge \forall p[(\forall x(G_1(x) \supset p(x)) \wedge \forall x(p(x) \supset \neg G_2(x))) \supset \forall x(p(x) \supset P(x))]\\ \wedge \forall p[(\forall x(G_1(x) \supset p(x)) \wedge \forall x(p(x) \supset \neg G_2(x))) \supset \forall x(P(x) \supset p(x))].\end{array} \tag{13}$$

To eliminate the second-order quantifiers in the last two conjunctive terms, observe that these terms can be written as the negations of the sentences

$$\exists p[\forall x(G_1(x) \supset p(x)) \wedge \forall x(p(x) \supset \neg G_2(x)) \wedge \neg \forall x(p(x) \supset P(x))] \tag{14}$$

and

$$\exists p[\forall x(G_1(x) \supset p(x)) \wedge \forall x(p(x) \supset \neg G_2(x)) \wedge \neg \forall x(P(x) \supset p(x))]. \tag{15}$$

The second conjunctive term of (14) contains a negative occurence of p, and the occurrences of p in the other two conjunctive terms are positive. Consequently, (14) is equivalent to

$$\forall x(G_1(x) \supset \neg G_2(x)) \wedge \neg \forall x(\neg G_2(x) \supset P(x)).$$

In the presence of the first two conjunctive terms of (13), this can be further rewritten as

$$\neg \forall x(\neg G_2(x) \supset P(x)).$$

The first conjunctive term of (15) contains a positive occurence of p, and the occurrences of p in the other two conjunctive terms are negative. Consequently, (15) is equivalent to

$$\forall x(G_1(x) \supset \neg G_2(x)) \wedge \neg \forall x(P(x) \supset G_1(x)).$$

In the presence of the first two conjunctive terms of (13), this can be further rewritten as

$$\neg\forall x(P(x) \supset G_1(x)).$$

Consequently, (13) is equivalent to

$$\forall x(G_1(x) \supset P(x)) \wedge \forall x(G_2(x) \supset \neg P(x)) \wedge \forall x(\neg G_2(x) \supset P(x)) \wedge \forall x(P(x) \supset G_1(x)),$$

which is equivalent to (12).

Proof of the Completion Theorem. Let T be a definite theory with the explainable symbols $P_1, \ldots, P_m$, and let T' be the theory obtained from T after Steps 1–3 of the completion process. It is easy to see that T' is equivalent to T. For each $i = 1, \ldots, m$, let T_i be the theory whose rules are the rules of T' that contain P_i in the head and whose only explainable symbol is P_i. Let T_{m+1} be the theory whose rules are the rules of T' that are not included in any of the theories $T_1, \ldots, T_m$ and whose set of explainable symbols is empty. It is clear that the theories $T_1, \ldots, T_m, T_{m+1}$ are pairwise disjoint, and that their union is T'. By Lemma 1, it follows that T' is equivalent to $T_1 \wedge \cdots \wedge T_m \wedge T_{m+1}$. Each of the theories $T_1, \ldots, T_m$ is simple. By Lemma 2, it follows that the formula $T_1 \wedge \cdots \wedge T_m$ is equivalent to the conjunction of the completion sentences obtained in Step 4. Since the heads of the rules of T_{m+1} do not include explainable symbols, it follows that T' is equivalent to the conjunction of all completion sentences for T.

7 Relation to the McCain-Turner Definition

The semantics of causal rules according to [McCain and Turner, 1997a] can be defined as follows. Let T be a set of propositional causal rules. For any interpretation I of the language of T, let T^I be the set of the heads of all rules of T whose bodies are satisfied by I. We say that I is *causally explained* according to T if I is the only model of T^I. A formula F is a *consequence* of T if every interpretation that is causally explained according to T satisfies F.

Consider, for instance, the propositional counterpart of rules (9):

$$\begin{aligned} &P \leftarrow Q, \\ &\neg P \leftarrow \neg P, \\ &Q \leftarrow Q, \\ &\neg Q \leftarrow \neg Q. \end{aligned} \tag{16}$$

For the interpretation I that makes both P and Q true, T^I consists of the formulas P and Q. Since I is the only model of these formulas, it is causally explained according to T. The interpretation that makes both P and Q false is causally explained also. The two other interpretations of the language of T are not causally explained.

To relate the notion of a causally explained interpretation to the semantics of causal theories defined in Section 2, let us agree to identify a set of propositional causal rules with the causal theory which consists of these rules and in which every propositional constant is explainable. For instance, we will consider the rules (16) as the rules of the propositional causal theory T in which both P and Q are explainable.

Proposition 3. *For any set T of propositional causal rules and any interpretation I, I is causally explained according to T iff I is a model of T.*

For instance, the two intepretations that are causally explained according to (16) are the models of $P \equiv Q$.

As a corollary, we conclude that the consequences of a set of propositional causal rules as defined by McCain and Turner are the same as its consequences in the sense of Section 2.

Proof of Proposition 3. Let

$$F_i(P) \leftarrow G_i(P) \qquad (i = 1, \ldots)$$

be the rules of T, where P is the list of all propositional constants. Then (2) is

$$\forall p(T^*(p) \equiv p = P),$$

where $T^*(p)$ stands for $\bigwedge_i (G_i(P) \supset F_i(p))$. Take an interpretation I, and let $\hat{I}(P)$ be the conjunction of all literals satisfied by I. It is clear that I is a model of T iff the formula

$$\hat{I}(P) \wedge \forall p(T^*(p) \equiv p = P)$$

is satisfiable. In the presence of $\hat{I}(P)$, $T^*(p)$ can be replaced by $T^I(p)$, where $T^I(P)$ is the conjunction of all formulas in T^I, and $p = P$ can be replaced by $\hat{I}(p)$. Consequently, I is a model of T iff the formula

$$\hat{I}(P) \wedge \forall p(T^I(p) \equiv \hat{I}(p)) \tag{17}$$

is satisfiable. Since the first conjunctive term $\hat{I}(P)$ is satisfiable and the second conjunctive term

$$\forall p(T^I(p) \equiv \hat{I}(p)) \tag{18}$$

does not contain any nonlogical constants, (17) is satisfiable iff (18) is logically valid. Furthermore, the logical validity of (18) is equivalent to the logical validity of

$$T^I(P) \equiv \hat{I}(P).$$

This formula is logically valid iff I satisfies $T^I(P)$.

Acknowledgements

Many thanks to Michael Gelfond, Hudson Turner and the anonymous referees for useful comments. This work was partially supported by National Science Foundation under grant IRI-9306751.

References

[Ackermann, 1935] Wilhelm Ackermann. Untersuchungen über das eliminationsproblem der mathematischen logik. *Mathematische Annalen*, 110:390–413, 1935.

[Baader and Hollunder, 1995] Franz Baader and Bernard Hollunder. Embedding defaults into terminological knowledge representation formalisms. *Journal of Automated Reasoning*, 14:149–180, 1995.

[Clark, 1978] Keith Clark. Negation as failure. In Herve Gallaire and Jack Minker, editors, *Logic and Data Bases*, pages 293–322. Plenum Press, New York, 1978.

[Geffner, 1990] Hector Geffner. Causal theories for nonmonotonic reasoning. In *Proc. AAAI-90*, pages 524–530, 1990.

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

[Kautz and Selman, 1992] Henry Kautz and Bart Selman. Planning as satisfiability. In *Proc. ECAI-92*, pages 359–363, 1992.

[Lifschitz and Rabinov, 1989] Vladimir Lifschitz and Arkady Rabinov. Things that change by themselves. In *Proc. IJCAI-89*, pages 864–867, 1989.

[Lifschitz, 1994] Vladimir Lifschitz. Circumscription. In D.M. Gabbay, C.J. Hogger, and J.A. Robinson, editors, *The Handbook of Logic in AI and Logic Programming*, volume 3, pages 298–352. Oxford University Press, 1994.

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

[McCain and Turner, 1995] Norman McCain and Hudson Turner. A causal theory of ramifications and qualifications. In *Proc. IJCAI-95*, pages 1978–1984, 1995.

[McCain and Turner, 1997a] Norman McCain and Hudson Turner. Causal theories of action and change. In *Proc. AAAI-97*, pages 460–465, 1997.

[McCain and Turner, 1997b] Norman McCain and Hudson Turner. Satisfiability planning with causal theories. In *Working Notes of the AAAI-97 Workshop on Robots, Softbots, Immobots: Theories of Action, Planning and Control*, pages 61–81, 1997.

[Reiter, 1980] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.