Causal Theories for Nonmonotonic Reasoning

Hector Geffner
hector@ibm.com
T. J. Watson Research Center
P. O. Box 704, Room H1-K10
Yorktown Heights, N.Y. 10598

Abstract

Causal theories are default theories which explicitly accommodate a distinction between 'explained' and 'unexplained' propositions. This is accomplished by means of an operator 'C' in the language for which propositions α are assumed explained when literals of the form $C\alpha$ hold. The behavior of causal theories is determined by a preference relation on models based on the minimization of unexplained abnormality. We show that causal networks, general logic programs and theories for reasoning about change can be all naturally expressed as causal theories. We also develop a proof-theory for causal theories and discuss how they relate to autoepistemic theories, prioritized circumscription, and Pearl's C-E calculus.

Introduction

Preferential entailment has emerged as a powerful means for specifying non-monotonic behavior. An order on interpretations determines the preferred models of a given theory, and those models determine in turn the propositions that the theory non-monotonically entails. Shoham [1988], who most clearly articulated the framework, used a preference order on interpretations to specify the behavior of temporal theories dealing with change. Similar proposals have been advanced for logic programs with negation [Przymusinski, 1987] and defeasible inheritance hierarchies [Krishnaprasad *et al.*, 1989] among others.

In all these proposals, the preference order adopted is tailored to the particular task or domain addressed. Still, the appeal of preferential entailment as a unifying framework for non-monotonic inference could be enhanced if a general domain-independent criterion for inferring preferences from theories could be developed. A general proof-theory, for instance, if available, could then be imported by any individual domain as well. Similarly, the resulting framework would subsume the expressive power of individual domains, enabling a user to express, say, theories which combine patterns of inference characteristic of theories for reasoning about change with those of general logic programs.

An attempt in this direction was recently reported in [Geffner, 1989b], where it was shown that non-monotonic inference in several domains of interest in AI could be understood in terms of a model preference criterion based on the minimization of *unexplained* abnormality. 'Explanations' were defined as logical derivations constrained by the *form* of the formulas in the knowledge base. A formula $\neg p \Rightarrow q$, for example, was assumed to permit an explanation of q in terms of an assumption $\neg p$, but to exclude an explanation of p in terms of an assumption $\neg q$. Thus, the different semantics associated with two logically equivalent logic programs $q \leftarrow \neg p$ and $p \leftarrow \neg q$, for instance, could be accounted for.

In this paper we pursue the same goal and intuition as in [Geffner, 1989b] but proceed with a different formalization. Rather than appealing to the syntactic form of the rules in the knowledge base to distinguish between logically equivalent theories, we appeal to a richer language. Such a language includes a new operator 'C' for which propositions α are assumed explained when literals of the form $C\alpha$ hold. Default theories over the new language are called *causal theories*. As we will show, not only causal theories do abide by the principle that equal models imply equal behavior, but they also provide a significant additional expressive power. Causal networks, general logic programs and theories about change can be all naturally expressed as causal theories. Moreover causal theories lend themselves to a simple sound and complete proof-theory.

Causal Theories

A causal theory is a default theory augmented with a 'causal' operator 'C.' Default theories are essentially what McCarthy's [1986] refers to as "abnormality" theories: classical first order theories in which certain literals, abnormalities (resp. assumptions), are expected to be false (resp. true).[1] The language of causal theories is closed under all standard connectives, while ex-

[1] Note that unlike McCarthy, we will be concerned with abnormality *literals* as opposed to abnormal *individuals*. The trade-offs involved in that choice are discussed in [Geffner, 1989a, section 4.3].

cluding expressions with embedded causal operators. We usually denote abnormalities by atoms of the form $\mathbf{ab_i}(a)$, and use the symbol α possibly indexed as a variable ranging over abnormalities. Additionally, we often partition the formulas in a causal theory T into two sets: a background context K containing the formulas which express generic knowledge (e.g. "penguins are birds"), and an evidence set E containing the formulas which express knowledge specific to the situation at hand (e.g. "Tweety is a bird"; see [Geffner and Pearl, 1987]).

The operator C is most commonly used to encode causal or explanatory rules of the form "if a then b" as sentences of the form $a \Rightarrow \mathrm{C}b$ (see [Pearl, 1988a]). A rule such as "rain causes the grass to be wet" may thus be expressed as a sentence **rain** $\Rightarrow$ C**grass_wet**, which can then be read as saying that if **rain** is true, **grass_wet** is explained. We assume that the operator C obeys certain minimal constraints which correspond to the postulates of system T in modal logic [Hughes and Cresswell, 1968]:

[C1] $\mathrm{C}\alpha \Rightarrow \alpha$

[C2] $\mathrm{C}(\alpha \Rightarrow \beta) \Rightarrow (\mathrm{C}\alpha \Rightarrow \mathrm{C}\beta)$

[C3] If $\vdash_K \alpha$ then $\mathrm{C}\alpha$

[C1] forces every explained proposition to be true, while [C2] and [C3] guarantee the set of explained proposition to be closed under deduction. The expression '$\vdash_K \alpha$' in [C3] is an abbreviation of $K \vdash \alpha$, which says that in any causal theory every proposition that logically follows from its background context is explained.

An interpretation M that satisfies a causal theory T and the constraints [C1]–[C3] will be said to be a model of T. Moreover, we will refer to the set of abnormalities rendered true by an interpretation M as the *gap* of the interpretation and denote it as $A[M]$. Since the preference relation on models will be an exclusive function of the model gaps we will find useful to group models into classes of models. Formally, the *class $\mathcal{C}$ of T with a gap* $A[\mathcal{C}]$ will represent the non-empty collection of models M of T with a gap $A[M] \subseteq A[\mathcal{C}]$. Intuitively, since the negation of abnormalities are *assumptions* expected to hold, a class $\mathcal{C}$ with a gap $A[\mathcal{C}]$ represents the collection of models which validate all assumptions logically compatible with $A[\mathcal{C}]$. Models and classes of T with *minimal gaps* will be said to be *minimal.* Moreover, we will say that a proposition p *holds* in a class $\mathcal{C}$ of T, when p holds in every model in $\mathcal{C}$. Proof-theoretically this is equivalent to require that p be derivable from T and a set of assumptions compatible with $A[\mathcal{C}]$.[2]

The operator C is used as a device to order the *classes* of models of the theories T of interest. As in [Geffner, 1989b], such a preference order is defined as a function of the abnormalities and the *explained* abnormalities in the different classes. An abnormality α is *explained* in a class $\mathcal{C}$ when the literal $\mathrm{C}\alpha$ holds in $\mathcal{C}$. If we denote the set of explained abnormalities in a class $\mathcal{C}$ by $A^c[\mathcal{C}]$, then the preference relation on classes of a theory T can be described as follows:

Definition 1 *A class $\mathcal{C}$ is* as preferred *as a class $\mathcal{C}'$ iff* $A[\mathcal{C}] - A^c[\mathcal{C}] \subseteq A[\mathcal{C}']$. *$\mathcal{C}$ is* preferred *to* *$\mathcal{C}'$ iff $\mathcal{C}$ is as preferred as $\mathcal{C}'$ but $\mathcal{C}'$ is* not *as preferred as $\mathcal{C}$.*

In words, a class $\mathcal{C}$ is preferred to a class $\mathcal{C}'$ when every abnormality in $\mathcal{C}$ but not in $\mathcal{C}'$ has an explanation, but not vice versa. Notice that this preference relation on classes is not necessarily transitive, and thus, certain care is required for defining the propositions which a given theory *causally entails.*

Let us say that a collection $\mathcal{B}$ of classes constitute a *basis* for a theory T if for every class $\mathcal{C}$ of T not in $\mathcal{B}$, there is a class $\mathcal{C}'$ in $\mathcal{B}$ such that $\mathcal{C}'$ is preferred to $\mathcal{C}$. Moreover let us say that a basis $\mathcal{B}$ *supports* a proposition p if p holds in every class in $\mathcal{B}$. Then, we will say that a proposition p is *causally entailed* by a causal theory T when there is a basis for T which supports p.

Most theories T of interest will be well-founded in the sense that for any *non-minimal* model M of T there will be a *minimal* model M' of T such that $A[M'] \subset A[M]$.[3] In such cases, causal entailment can be computed by considering the minimal classes of T only. Indeed, $\mathcal{B}$ will be a basis for a well-founded theory T iff for every *minimal* class $\mathcal{C}$ of T *not* in $\mathcal{B}$, $\mathcal{B}$ includes a *minimal* class $\mathcal{C}'$ preferred to $\mathcal{C}$.

Moreover, T will often possess a single *minimal* basis $\mathcal{B}$. In such cases, we will refer to the classes in $\mathcal{B}$ as the *preferred classes* of T. For such theories causal entailment can be cast in the more familiar form in which a proposition is causally entailed when it holds in all the preferred classes.

Example 1 Let us consider first a simple causal theory T given by the single sentence $\neg \mathbf{ab_1} \Rightarrow \mathrm{C}\mathbf{ab_2}$, where $\mathbf{ab_1}$ and $\mathbf{ab_2}$ are two different abnormalities. Such a theory admits two minimal classes: a class $\mathcal{C}_1$, comprised by the models of T which only sanction the abnormality $\mathbf{ab_1}$, and a class $\mathcal{C}_2$, comprised of the models which only sanction the abnormality $\mathbf{ab_2}$. Thus $\mathcal{C}_1$ has an associated gap $A[\mathcal{C}_1] = \{\mathbf{ab_1}\}$, while $\mathcal{C}_2$ has an associated gap $A[\mathcal{C}_2] = \{\mathbf{ab_2}\}$. Both classes represent the minimal classes of T, as there is no model of T that satisfies both $\neg \mathbf{ab_1}$ and $\neg \mathbf{ab_2}$, together with the restriction $\mathrm{C}\alpha \Rightarrow \alpha$. The abnormalities α explained in each class $\mathcal{C}$ can be determined by testing which literals $\mathrm{C}\alpha$ hold in $\mathcal{C}$. As we said, this amounts checking whether there is a set of assumptions legitimized by $\mathcal{C}$ which together with T implies $\mathrm{C}\alpha$. Thus, in the class $\mathcal{C}_2$, the abnormality $\mathbf{ab_2}$ is

[2] The notions of derivability and consistency are to be understood relative to the postulates [C1]–[C3].

[3] A sufficient condition for T to be well-founded is that T gives rise to a finite number of bound assumptions, where an assumption is bound when it is in conflict with other assumptions (see [Geffner, 1989a]).

explained as the literal $\text{C}\mathbf{ab_2}$ logically follows from T and the assumption $\neg\mathbf{ab_1}$. On the other hand, the abnormality $\mathbf{ab_1}$ is *not* explained in $\mathcal{C}_1$, as there is no set of assumptions validated by $\mathcal{C}_1$ which supports the literal $\text{C}\mathbf{ab_1}$. It follows then, that the class $\mathcal{C}_2$ is preferred to $\mathcal{C}_1$, as $A[\mathcal{C}_2] - A^c[\mathcal{C}_2] = \emptyset \subseteq A[\mathcal{C}_1]$, but $A[\mathcal{C}_1] - A^c[\mathcal{C}_1] = \{\mathbf{ab_1}\} \not\subseteq A[\mathcal{C}_2] = \{\mathbf{ab_2}\}$. Furthermore, since the theory T is well-founded, and $\mathcal{C}_1$ and $\mathcal{C}_2$ are the only minimal classes of T, it follows then that $\mathcal{B} = \{\mathcal{C}_2\}$ is the single minimal basis of T, and thus, that $\mathcal{C}_2$ is the single preferred class of T. As a result, the propositions $\neg\mathbf{ab_1}$ and $\mathbf{ab_2}$ which hold in $\mathcal{C}_2$ are (causally) entailed by T.

Applications

In this section we consider the use of causal theories for reasoning about change and for specifying and extending the semantics of general logic programs. For the use of causal theories for inheritance and abductive reasoning see [Geffner, 1989a].

Reasoning about Change

Theories for reasoning about change need to represent the effects of actions, the conditions which can prevent actions from achieving their normal effects, and the tendency of certain aspects of the world (fluents) to remain stable (see [McDermott, 1982]). Here we will refer to the first type of rules as *change* rules, to the second type as *cancellation* rules, and to the third type as *persistence* rules.

Change, cancellation and persistence rules can interact in various ways. The Yale shooting scenario [Hanks and McDermott, 1987] illustrates a problem that results from spurious interactions between change and persistence rules. We now present general guidelines to locally map general theories for reasoning about change into causal theories which avoid those problems. The guidelines are uncommitted about the particular temporal notation used. For simplicity, we use a simple reified temporal language sufficient to illustrate the relevant issues. Other notations could be used as well. The notation $p(x)_t$ below, where p is a predicate and t is a time point, is used as an abbreviation of the sentence $\text{Holds}(p(x), t)$, to read "fluent $p(x)$ holds at time t." We also assume for simplicity a discrete time where t precedes $t+1$.

First we specify the encoding of rules about change. A rule describing the effect $\mathtt{e}(x)$ of an action $\mathtt{a}(x)$ with precondition $\mathtt{p}(x)$ is encoded as a causal rule of the form:

$$\mathtt{p}(x)_t \wedge \mathtt{a}(x)_t \Rightarrow \text{C}\mathtt{e}(x)_{t+1}$$

where x is a tuple of variables and both x and t are universally quantified. Such a rule can be read as stating that given the precondition $\mathtt{p}(x)$, $\mathtt{a}(x)$ causes or explains $\mathtt{e}(x)$.

Often, however, rules about change are defeasible. Defeasible rules about change are encoded by means of a unique abnormality predicate $\mathtt{ab_i}$ and a *pair* of causal rules:

$$\mathtt{p}(x)_t \wedge \mathtt{a}(x)_t \wedge \neg\mathtt{ab_i}(x)_t \Rightarrow \text{C}\mathtt{e}(x)_{t+1}$$
$$\mathtt{p}(x)_t \wedge \mathtt{a}(x)_t \wedge \text{C}\neg\mathtt{e}(x)_{t+1} \Rightarrow \text{C}\mathtt{ab_i}(x)_t$$

where the second rule simply asserts that the violation of an expected change is explained when there is an explanation for the negation of the expected effect (a similar rule is needed for modeling inheritance hierarchies [Geffner, 1989a]).

The persistence of a fluent f (e.g. $\mathtt{on(a,b)}$), on the other hand, is encoded by the expressions:

$$f_t \wedge \neg\mathtt{ab_i}(f)_t \Rightarrow f_{t+1}$$
$$\text{C}\neg f_{t+1} \Rightarrow \text{C}\mathtt{ab_i}(f)_t$$

where $\mathtt{ab_i}(f)_t$ is an abbreviation of the atom $\mathtt{ab_i}(f, t)$, read "the persistence of f holds at time t." Thus, while the first rule expresses the tendency of fluents to remain stable, the second rule expresses that changes are explained when the negation of the projected fluent is explained.

For the causal encoding of a version of the Yale Shooting problem, see [Geffner, 1989a]. Here we will consider a slightly richer example due to Ginsberg and Smith [1988].

Example 2 Let us assume that there is a room with some ducts that maintain the room ventilated. Moreover, an object sitting on a duct, blocks the duct, and if all ducts get blocked, the room becomes stuffy. This information is encoded in a causal theory with background:

$$\mathtt{duct}(x) \wedge \exists y.\mathtt{on}(y,x)_t \Rightarrow \text{C}\mathtt{blocked}(x)_t$$
$$[\forall x.\mathtt{duct}(x) \Rightarrow \mathtt{blocked}(x)_t] \Rightarrow \text{C}\mathtt{stuffy}_{t+1}$$
$$\mathtt{move\_to}(x,y)_t \wedge \neg\mathtt{ab_1}(x,y)_t \Rightarrow \text{C}\mathtt{on}(x,y)_{t+1}$$
$$\mathtt{move\_to}(x,y)_t \wedge \text{C}\neg\mathtt{on}(x,y)_{t+1} \Rightarrow \text{C}\mathtt{ab_1}(x,y)_t$$

The persistence of the fluents $\mathtt{on}(x,y)$, $\mathtt{stuffy}$, and $\mathtt{blocked}(x,y)$, and their negations, is expressed as stipulated above. To keep in mind that all these fluents are really *terms*,[4] we use the notation $\overline{f}$ to denote the fluent which is the complement of f. Thus, for instance, $\overline{\mathtt{on(a,b)}}$ stands for the 'negation' of $\mathtt{on(a,b)}$. Namely, if $\mathtt{on(a,b)}$ holds at time t, $\overline{\mathtt{on(a,b)}}$ will not, and vice versa. This is expressed by a constraint

$$f_t \Rightarrow \neg\overline{f}_t$$

which renders f and $\overline{f}$ incompatible, provided that the complement of $\overline{f}$ is f itself.

Finally, we need to express that an object cannot be on two different places at the same time:

$$\mathtt{on}(x,y)_t \wedge \mathtt{on}(x,z)_t \Rightarrow y = z$$

Given this background K, we consider a theory $T = \langle K, E\rangle$ describing a room with two ducts $\mathtt{d_1}$ and $\mathtt{d_2}$. Furthermore, at time $\mathtt{t} = \mathtt{0}$ it is known that the room

[4] Recall that $\mathtt{blocked}(x,y)_t$ is an abbreviation of the *atom* $\text{Holds}(\mathtt{blocked}(x,y), t)$.

is not stuffy, that a block $\mathtt{a}$ is sitting on top of duct $\mathtt{d_1}$, and that a block $\mathtt{b}$ is sitting on a place different than $\mathtt{d_2}$. Namely, $E = \{\mathtt{duct}(x) \Rightarrow x = \mathtt{d_1} \vee x = \mathtt{d_2}, \overline{\mathtt{stuffy_0}}, \mathtt{on(a, d_1)_0}, \overline{\mathtt{on(b, d_2)_0}}\}$.

In the context T, the fluents $\overline{\mathtt{stuffy_0}}$, $\mathtt{on(a, d_1)_0}$, and $\overline{\mathtt{on(b, d_2)_0}}$ project both forward and backward in time. If block $\mathtt{b}$ is moved to duct $\mathtt{d_2}$ at time $\mathtt{t} = 0$, however, conflicts among these projections arise, resulting in three classes of minimal models: the intended class $\mathcal{C}$ where the action is successful and, as a result, the two ducts get blocked and the room becomes stuffy; the class $\mathcal{C}'$, where the action is successful but somehow the block $\mathtt{a}$ has been removed from duct $\mathtt{d_1}$; and the class, $\mathcal{C}''$, where the action is unsuccessful and the block $\mathtt{b}$ remains in a place different than $\mathtt{d_2}$. Nonetheless, the interpretation of causal theories singles out the intended class $\mathcal{C}$ as the only preferred class, capturing the intuition that block $\mathtt{a}$ stays on $\mathtt{d_1}$ and that the room becomes stuffy. Note that such a behavior arises without the presence of explicit cancellation axioms.

Logic Programming

While the adequacy of the framework presented for reasoning about change rests mainly on empirical grounds —how natural it is to express knowledge about these domains and how closely the resulting behavior resembles the behavior intended by the user— a growing body of work on the semantics of general logic programs will permit us to assess the expressivity and semantics of causal theories on more formal grounds.

As it is standard, we consider only the Herbrand models of programs. Moreover, since for answering existential queries, a program involving variables can be shown to be equivalent to a program without variables, we will be dealing mainly with variable-free logic programs. More precisely, we will analyze the semantics of general logic programs in terms of two mappings $C_i[\cdot]$, $i = 1, 2$, each converting a program P into a causal theory $C_i[P]$. Each mapping associates a different "meaning" with P. For the purposes of logic programming, $C_1[P]$ is the most relevant. The mapping $C_2[\cdot]$ will be used mainly to illustrate the relation between the interpretation of general logic programs and the semantics of causal theories. We assume every atom in a logic program to be an "abnormality" and write $\mathcal{C}_A$ to represent the class of Herbrand models M whose "abnormalities" are among those of A. Namely, if $\mathcal{L}$ denotes the formulas not involving the causal operator, $\mathcal{C}_A$ will stand for the collection of models M such that $M \cap \mathcal{L} \subseteq A$.

We consider first the mapping $C_1[\cdot]$ which converts each rule

$$\gamma \leftarrow \alpha_1, \ldots, \alpha_n, \neg\beta_1, \ldots, \neg\beta_m$$

in P, where $n \geq 0$ and $m \geq 0$, and α's, β's and γ are atoms, into a *causal* rule of the form

$$C\alpha_1 \wedge \ldots \wedge C\alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow C\gamma \quad .$$

Example 3 Consider a program P given by the following rules:

$$\mathtt{c} \leftarrow \mathtt{a}, \neg\mathtt{b}$$
$$\mathtt{d} \leftarrow \neg\mathtt{c}$$
$$\mathtt{a} \leftarrow$$

P gives rise to two minimal models: $M_1 = \{\mathtt{a}, \mathtt{c}\}$ and $M_2 = \{\mathtt{a}, \mathtt{b}, \mathtt{d}\}$, the former of which is the single *canonical* or *perfect* model of P [Apt *et al.*, 1987; Przymusinski, 1987].

The mapping $C_1[\cdot]$ maps P into the causal theory $C_1[P]$:

$$\mathtt{Ca} \wedge \neg\mathtt{b} \Rightarrow \mathtt{Cc}$$
$$\neg\mathtt{c} \Rightarrow \mathtt{Cd}$$
$$\mathbf{true} \Rightarrow \mathtt{Ca}$$

$C_1[P]$ gives rise to two minimal classes $\mathcal{C}_{M_1}$ and $\mathcal{C}_{M_2}$, with gaps M_1 and M_2 as above. Furthermore, in the former class, the atoms $\mathtt{a}$ and $\mathtt{c}$ are both explained, as $C_1[P], \neg\mathtt{b} \vdash \mathtt{Ca} \wedge \mathtt{Cc}$ holds, and $\neg\mathtt{b}$ is a legitimate assumption in $\mathcal{C}_{M_1}$. On the other hand, only the atom $\mathtt{a}$ is explained in $\mathcal{C}_{M_2}$. Thus, the class $\mathcal{C}_{M_1}$ is the single preferred class of $C_1[P]$. As a result, the canonical model M_1 of P and the preferred class $\mathcal{C}_{M_1}$ of $C_1[P]$ sanction the same non-causal (free from 'C') literals.

As the example suggests, for *stratified* programs the following correspondence between the canonical model of P and the single preferred class of the theory $C_1[P]$ can be established:[5]

Theorem 1 *M is the canonical model of a stratified program P if and only if $\mathcal{C}_M$ is the single preferred class of $C_1[P]$.*

Moreover, if we say that a class of models is *perfectly coherent* when it explains *every* abnormality that it sanctions, we obtain a correspondence between the *stable* models of a program P [Gelfond and Lifschitz, 1988; Fine, 1989] and the perfectly coherent classes of the causal theory $C_1[P]$, even when P is *not* stratified.

Theorem 2 *M is a* stable *model of an arbitrary program P if and only if $\mathcal{C}_M$ is a perfectly coherent class of the causal theory $C_1[P]$.*

In spite of this correspondence, however, the semantics of causal theories $C_1[P]$ and the stable semantics of logic programs P diverge outside the family of stratified programs. On the one hand, programs may lack stable models (e.g. $\{\mathtt{p} \leftarrow \neg\mathtt{p}\}$); on the other, causal theories $C_1[P]$ may give rise to multiple preferred classes even when P accepts a single stable model (e.g. $P = \{\mathtt{a} \leftarrow \neg\mathtt{b}, \mathtt{b} \leftarrow \neg\mathtt{a}, \mathtt{p} \leftarrow \neg\mathtt{b}, \mathtt{p} \leftarrow \neg\mathtt{p}\}$).

Logic Programs and Causal Networks We now investigate the semantics associated with a second mapping $C_2[\cdot]$ of logic programs into causal theories.

[5] Proofs can be found in [Geffner, 1989a].

For a logic program P, $C_2[P]$ represents the collection of rules which result from mapping each rule

$$\gamma \leftarrow \alpha_1, \ldots, \alpha_n, \neg\beta_1, \ldots, \neg\beta_m$$

in P, into a causal rule of the form:

$$\alpha_1 \wedge \ldots \wedge \alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow C\gamma$$

The difference with the previous translation is that the positive antecedents of the resulting causal rules do not need to be "causally" established. This renders the models of the causal theory $C_2[P]$ as models of $C_1[P]$, though not the other way around. As a result, the semantics of causal theories of the form $C_2[P]$ differs from that of $C_1[P]$ even within the family of stratified programs.

For instance, the stratified program $P = \{\mathtt{q} \leftarrow \neg\mathtt{p}, \mathtt{p} \leftarrow \mathtt{r}, \mathtt{r} \leftarrow \mathtt{p}\}$, possesses a single canonical model $M = \{\mathtt{q}\}$, and thus $\mathcal{C}_M$ is the single causally preferred class of $C_1[P]$. On the other hand, the mapping $C_2[\cdot]$ renders the causal theory $C_2[P] = \{\neg\mathtt{p} \Rightarrow C\mathtt{q}, \mathtt{r} \Rightarrow C\mathtt{p}, \mathtt{p} \Rightarrow C\mathtt{r}\}$ which accepts *two* preferred classes $\mathcal{C}_M$ and $\mathcal{C}_{M'}$, with $M = \{\mathtt{q}\}$ and $M' = \{\mathtt{p}, \mathtt{r}\}$.

In this example, the 'anomalous' behavior of the theory $C_2[P]$ is a consequence of the circularity relating the atoms $\mathtt{p}$ and $\mathtt{r}$. 'Circular' explanations are precluded in $C_1[P]$ but not in $C_2[P]$. What is interesting, however, is that once these circularities are removed, the 'anomalous' behavior is guaranteed to disappear.

Let us say that a program P is *acyclic* when its dependency graph does not contain cycles. Acyclic programs are thus stratified. Moreover, acyclic programs, not only preclude 'recursion trough negation,' but *every* type of recursion. For acyclic programs, the following result applies.

Theorem 3 *Let P be an acyclic program. Then the class $\mathcal{C}_M$, where M is the canonical model of P, is the unique preferred class of the theories $C_1[P]$ and $C_2[P]$.*

In other words, once recursion is removed the two mappings examined result into an identical behavior, in correspondence with the received semantics of logic programs. While the requirement of acyclicity is unacceptably strong in the domain of programming, it is common among network representational languages, such as inheritance hierarchies [Touretzky, 1986] and Bayesian networks [Pearl, 1988b]. Indeed, causal theories of the form $C_2[P]$ for acyclic programs P, possibly augmented by integrity constraints, provide a sufficiently expressive language for reasoning in *causal networks*. [Geffner, 1989a] discusses the use of such representations for abductive reasoning as well as some of their advantages over the representations resulting from the mapping $C_1[\cdot]$.

Proof-Theory

The proof-theory of causal theories is structured in the form of an argument-based system (e.g. [Loui, 1987]). We assume the theories of interest are well-founded in the sense defined above. The notions of consistency and derivations are to be understood relative to postulates [C1]-[C3].

We start with some terminology. Assumptions are the complements of abnormalities. We denote the complement of a proposition p as $\bar{p}$, thus, if p is an abnormality $\bar{p}$ is an assumption, and vice versa. Furthermore, in a context T, a set Δ of assumptions constitutes an ***argument*** if the set Δ is consistent with T. Moreover, Δ is an argument *for* q, if q is derivable from the causal theory $T + \Delta$, and an argument *against* q, if $\neg q$ is derivable from $T + \Delta$. When the context T is understood, we also say that Δ *supports* q and $\neg q$ respectively. Now, if Δ is not an argument in T, then Δ is said to be a *conflict set.* Two arguments are *in conflict* when their union is a conflict set. In such a case we also say that one argument *refutes* the other.

For instance, in a theory T consisting of the expressions $\neg\mathbf{ab_1} \Rightarrow C\mathbf{ab_2}$ and $\neg\mathbf{ab_3} \Rightarrow C\mathtt{q}$, the arguments $\Delta_1 = \{\neg\mathbf{ab_1}\}$ and $\Delta_2 = \{\neg\mathbf{ab_2}\}$ are in conflict. On the other hand, no argument is in conflict with $\Delta_3 = \{\neg\mathbf{ab_3}\}$.

By the minimality of preferred classes, it easily follows that propositions supported by unrefuted arguments are entailed. Thus, for example, we can safely conclude that propositions $C\mathtt{q}$ and $\mathtt{q}$ above are entailed, as they are supported by the unrefuted argument Δ_3.

Often, however, refuted arguments may also provide legitimate support. Once such example is the proposition $\mathbf{ab_2}$ which is entailed by T in spite of having a single minimal supporting argument Δ_1 which is refuted by Δ_2. Intuitively, what is going on is that Δ_1 not only refutes Δ_2, but also explains its negation. We will say in that case that Δ_1 is *protected* from Δ_2. As we will see, propositions supported by protected arguments may also be entailed.

Formally, let us say that an argument Δ explains a proposition p when Δ is an argument for Cp. Then the notion of *protection* can be defined as follows:

Definition 2 *An argument Δ is* protected *from a conflicting argument Δ' iff there is a set $\Delta'' \subseteq \Delta' - \Delta$ such that $\Delta + \Delta' - \Delta''$ is not a conflict set, and every abnormality α, $\bar{\alpha} \in \Delta''$, is explained by $\Delta + \Delta' - \Delta''$.*

Similarly we will say that an argument Δ is *stronger* than a conflicting argument Δ' when Δ is protected from Δ' but Δ' is not protected from Δ. In the example above, for instance, Δ_1 is *stronger* than the conflicting argument Δ_2. If we say that an argument is *stable* when it is stronger than any conflicting argument, we obtain the following sufficient conditions for a proposition to be causally entailed:

Theorem 4 *If a proposition p is supported by an* stable *argument, then p is causally entailed.*

Yet, theorem 4 does not provide *necessary* conditions. For instance, the proposition $\mathbf{ab_3}$ is causally entailed by the theory comprised of the formulas $\neg\mathbf{ab_1} \vee$

$\neg\mathbf{ab}_2 \Rightarrow \mathrm{C}\mathbf{ab}_3$ and $\mathbf{ab}_1 \vee \mathbf{ab}_2$, and yet, $\mathbf{ab}_3$ is not supported by any stable argument.

A simple extension of the definitions above takes care of such cases. Let us refer to a collection of arguments as a *cover*, and let us say that a cover *supports* a proposition p if every argument in the cover supports p. Furthermore, let us say that an argument is in *conflict* with a cover when the argument is in conflict with every argument in the cover, and that a cover is *stronger* than a conflicting argument Δ when it contains an argument stronger than Δ. Moreover, let us also say that a cover is *stable* when it is stronger than every conflicting argument. Then, the following sound and complete characterization of causal theories results:

Theorem 5 (Main) *A proposition p is causally entailed if and only if it is supported by a stable* cover.

For the theory comprised of the sentences $\neg\mathbf{ab}_1 \vee \neg\mathbf{ab}_2 \Rightarrow \mathrm{C}\mathbf{ab}_3$ and $\mathbf{ab}_1 \vee \mathbf{ab}_2$, it is easy to show that the pair of arguments $\Delta_1 = \{\neg\mathbf{ab}_1\}$ and $\Delta_2 = \{\neg\mathbf{ab}_2\}$ constitute a stable cover. Since such a cover supports the proposition of $\mathbf{ab}_3$, it follows then that $\mathbf{ab}_3$ is entailed by the theory.

Related Work

Causal theories are an elaboration of ideas in [Geffner, 1989b], where the notions of *explanations*, *classes*, and *coherence* were originally presented. The adoption here of a *causal operator* as part of the object-level language, however, has simplified matters considerably, providing additional expressive power and permitting the construction of a proof-theory. Part of the motivation for the move came from a proposal due to Pearl to explicitly incorporate a causal language into default theories. Pearl's proposal [Pearl, 1988a] draws on work in causal probabilistic networks to suggest a distinction between defaults which *encode* explanations (e.g. `fire` → `smoke`) from defaults which *trigger* explanations (e.g. `smoke` → `fire`). He calls the former defaults *causal* and the latter *evidential*. He argues that the language of default theories should accommodate such a distinction, and in particular, that explanation 'giving' defaults should be prevented from triggering explanation 'seeking' defaults. Pearl's proposal to preclude such chains consists of three parts. First, he labels every default as either *causal*, e.g. `rain` $\rightarrow_{\mathrm{C}}$ `grass_wet`, or *evidential*, e.g. `grass_wet` $\rightarrow_{\mathrm{E}}$ `sprinkler_on`; second, he distinguishes the status of propositions p established on *causal* grounds, Cp, from those established on *evidential* grounds, Ep; and third, he introduces a calculus for reasoning with causal and evidential rules which purposely precludes deriving q from Cp and an evidential rule $p \rightarrow_E q$.

Though differing in detail and goals, the reading of the operator 'C' in causal theories follows Pearl's intuitions. Pearl, however, focuses on evidential reasoning, while we focus on default reasoning. A proposal for performing evidential reasoning in causal theories is discussed in [Geffner, 1989a].

Causal theories are also related to Moore's [1985] autoepistemic theories. The autoepistemic encoding $L[P]$ of a stratified logic program P [Gelfond, 1987], for instance, turns out to be the "dual" of the causal encoding $C_1[P]$ (i.e. in $L[P]$ every ***negated*** atom is preceded with the ***autoepistemic*** operator 'L', while in $C_1[P]$ every ***non-negated*** atom is preceded with the ***causal*** operator 'C') and they both legitimize the same behavior. Indeed, it is possible to understand the autoepistemic operator L as an *evidential* operator, with Lα meaning "there is evidence for α." Namely, instead of using the *causal* operator C under the conventions that

$\neg\alpha$ is an assumption
C$\alpha \Rightarrow \alpha$ must hold for every (plain) sentence α, and
α is *explained* in a class when Cα holds,

we could have used an *evidential* operator E under the conventions that

$\neg$Eα is an assumption
$\alpha \Rightarrow$ Eα must hold for every (plain) sentence α, and
Eα is *explained* in a class when α holds.

Under such an approach the *evidential* encoding of a logic program would be identical to the *autoepistemic* encoding, except for the presence of E's instead of L's. Moreover, both encodings would sanction an equivalent semantics for stratified programs. For non-stratified programs, however, as for most default theories, the duality between causal and autoepistemic disappears. First, default theories may lack stable models; second, the prefix ¬L, as no causal prefix, "generates" the assumptions needed.

The fact that the operator 'C' establishes a preference for the abnormality $\mathbf{q}$ over the abnormality $\mathbf{p}$ in a theory like $\neg\mathbf{p} \Rightarrow \mathrm{C}\mathbf{q}$, raises the question of whether the semantics of causal theories can be understood in terms of prioritized circumscription [McCarthy, 1986]. The answer is a qualified no: there are causal theories for which no priority order on the abnormalities will render an equivalent behavior. The causal theory $\{\neg\mathsf{a} \wedge \neg\mathsf{b} \Rightarrow \mathrm{C}\mathsf{c} \wedge \mathrm{C}\mathsf{d}, \neg\mathsf{c} \wedge \neg\mathsf{d} \Rightarrow \mathrm{C}\mathsf{a} \wedge \mathrm{C}\mathsf{b}\}$ for abnormalities a, b, c, and d, is one such example. Still, the semantics of such theories could in principle be captured by defining priorities for *non-atomic* formulas.

Finally, another family of theories related to those treated in this paper is Shoham's [1988] causal theories. Shoham's causal theories are epistemic theories designed for efficient reasoning about change. They are interpreted by a preference criterion which rewards models in which "as little is known for as long as possible." While there is no direct correspondence between our causal theories and Shoham's, it seems possible to understand the intuition behind Shoham's chronological minimization in terms of the ideas of explanation and coherence. If we recall that we regard an abnormality α as explained in a class C when C validates a set of assumption Δ which supports the truth of $C\alpha$, chronological minimization assumes α explained by Δ instead, when Δ supports the truth of α, without involving assumptions about times past α.

Acknowledgements. For the topic of this paper I have benefited from conversations with many people. I'd like to thank here K. Fine, M. Gelfond, M. Ginsberg, B. Grosof, K. Konolige, J. Pearl and H. Przymusinska.

References

[Apt *et al.*, 1987] K. Apt, H. Blair, and A. Walker. Towards a theory of declarative knowledge. In J. Minker, editor, *Foundations of Deductive Databases and Logic Programming*, pages 89–148. Morgan Kaufmann, Los Altos, CA, 1987.

[Fine, 1989] K. Fine. The justification of negation as failure. *Proceedings of 8th International Congress of Logic Methodology and Philosophy of Science*. North Holland, 1989.

[Geffner and Pearl, 1987] H. Geffner and J. Pearl. A framework for reaasoning with defaults. Technical Report TR-94, Cognitive Systems Laboratory, UCLA, Los Angeles, CA., August 1987. To appear in *Knowledge Representation and Defeasible Inference*, H. Kyburg, R. Loui and G. Carlson (Eds), Kluwer, 1989.

[Geffner, 1989a] H. Geffner. *Default Reasoning: Causal and Conditional Theories.* PhD thesis, UCLA, Los Angeles, CA, November 1989.

[Geffner, 1989b] H. Geffner. Default reasoning, minimality and coherence. *Proceedings of the First International Conference on Principle of Knowledge Representation and Reasoning*, pages 137–148, Toronto, Ontario, 1989.

[Gelfond and Lifschitz, 1988] M. Gelfond and V. Lifschitz. The stable model semantics for logic programming. *Proceedings 1988 Symposium on Logic Programming*, pages 1070–1080, Cambridge, Mass., 1988. MIT Press.

[Gelfond, 1987] M. Gelfond. On stratified autoepistemic theories. *Proceedings AAAI-87*, pages 207–211, Seattle, Washington, 1987.

[Ginsberg and Smith, 1988] M. Ginsberg and Smith. Reasoning about action I: A possible worlds approach. *Artificial Intelligence*, 35:165–195, 1988.

[Hanks and McDermott, 1987] S. Hanks and D. McDermott. Non-monotonic logics and temporal projection. *Artificial Intelligence*, 33:379–412, 1987.

[Hughes and Cresswell, 1968] G. Hughes and Cresswell. *An Introduction to Modal Logic.* Methuen and Co. LTD, London, Great Britain, 1968.

[Krishnaprasad *et al.*, 1989] T. Krishnaprasad, Kiefer, and D. Warren. On the circumscriptive semantics of inheritance networks. In Z. Ras and L. Saitta, editors, *Methodologies for Intelligent Systems 4*. North Holland, New York, N.Y., 1989.

[Loui, 1987] R. Loui. Defeat among arguments: A system of defeasible inference. *Computational Intelligence*, 1987.

[McCarthy, 1986] J. McCarthy. Applications of circumscription to formalizing commonsense knowledge. *Artificial Intelligence*, 28:89–116, 1986.

[McDermott, 1982] D. McDermott. A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6:101–155, 1982.

[Moore, 1985] R. Moore. Semantical considerations on non-monotonic logics. *Artificial Intelligence*, 25:75–94, 1985.

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

[Pearl, 1988b] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, Los Altos, CA., 1988.

[Przymusinski, 1987] T. Przymusinski. On the declarative semantics of stratified deductive databases and logic programs. In J. Minker, editor, *Foundations of Deductive Databases and Logic Programming*, pages 193–216. Morgan Kaufmann, Los Altos, CA, 1987.

[Shoham, 1988] Y. Shoham. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence.* MIT Press, Cambridge, Mass., 1988.

[Touretzky, 1986] D. Touretzky. *The Mathematics of Inheritance Systems.* Pitman, London, 1986.