Causal Default Reasoning: Principles and Algorithms

Hector Geffner
Departamento de Computación
Universidad Simón Bolívar
Aptdo 89000, Caracas 1080-A
Venezuela
hector@usb.ve

Abstract

The minimal model semantics is a natural interpretation of defaults yet it often yields a behavior that is too weak. This weakness has been traced to the inability of minimal models to reflect certain *implicit* preferences among defaults, in particular, preferences for defaults grounded on more 'specific' information and preferences arising in causal domains. Recently, 'specificity' preferences have been explained in terms of conditionals. Here we aim to explain causal preferences. We draw mainly from ideas known in Bayesian Networks to formulate and formalize *two principles* that explain the basic preferences that arise in causal default reasoning. We then define a semantics based on those principles and show how variations of the algorithms used for inheritance reasoning and temporal projection can be used to compute in the resulting formalism.

Motivation

The semantics of minimal models provides a direct interpretation of defaults: to determine the acceptable consequences of a classical theory W augmented with a set D of formulas labeled as 'defaults', the semantics selects the models of W that violate minimal subsets of D. The resulting semantics is simple and captures the basic intuition that 'as many defaults should be accepted as it is consistently possible". Yet the conclusions sanctioned are too weak. Consider for example the theory comprised of the defaults[1]

$$(\star) \qquad r_1 : \mathsf{a} \rightarrow \mathsf{d} \ , \ r_2 : \mathsf{a} \wedge \mathsf{b} \rightarrow \neg \mathsf{d} \ , \ r_3 : \mathsf{c} \rightarrow \mathsf{b}$$

The defaults may stand for the rules 'if I turn the key, the car will start', 'if I turn the key and the battery is dead, the car won't start', and 'if I left the lights on last night, the battery is dead'. They can also be thought as a simplified representation of the Yale Shooting scenario (Hanks & McDermott 1987): 'if Fred is alive, he will be alive', 'if Fred is alive and I shoot, Fred won't be alive', and 'if I load the gun, I will shoot'.

[1]Throughout the paper, defaults $p \rightarrow q$ are regarded as material implications, not as rules of inference. Material implications which are firmly believed will be denoted with the symbol '$\Rightarrow$'.

Given the facts a and c, we would expect the conclusion $\mathsf{b} \wedge \neg \mathsf{d}$. Yet the facts produce three minimal models M_i, each M_i violating a single default r_i, $i = 1, 2, 3$, with two of those models, M_2 and M_3, sanctioning exactly the opposite conclusion.

Part of the reason the expected conclusion is not sanctioned is that we haven't explicated the precedence of the rule $\mathsf{a} \wedge \mathsf{b} \rightarrow \neg \mathsf{d}$ over the conflicting but less 'specific' rule $\mathsf{a} \rightarrow \mathsf{d}$. This precedence can be expressed in a number of ways: by giving the first rule priority over the second, by making the first rule a strict non-defeasible rule, or by adding explicit axioms to 'cancel' the second rule when the first rule is applicable. Each of these options have different merits. For our purposes what matters is that they all prune the minimal model M_2 that violates the 'superior' rule r_2, but leave the other two minimal models M_1 and M_3 intact. Hence, the expected conclusion $\mathsf{b} \wedge \neg \mathsf{d}$ still fails to be sanctioned. The question this raises, is: *on what grounds can the 'unintended' minimal model M_3, where rule r_3 is violated, be pruned?*

The default theory $(\star)$ is not very different from the theories handled by inheritance and temporal projection algorithms (Horty, Thomason, & Touretzky. 1987; Dean & McDermott 1987). The idea in these algorithms is to use a default like $A \rightarrow p$ to conclude p from A, when either there are no 'reasons' for $\neg p$ or when those reasons are 'weaker' than $A \rightarrow p$. The essence of these algorithms is a careful form of forward chaining that can be expressed as follows:

<u>Procedure P_0</u>

$\vdash p$ if p in W or $\vdash A$, $A \rightarrow p$ in D, and all *forward arguments* for $\sim p$ not weaker than $A \rightarrow p$ contain a rule $B \rightarrow q$ s.t. $\vdash \sim q$

As it is common in these procedures, rules are definite, A and B are conjunctions of atoms, p and $\sim p$ are incompatible propositions, and $\vdash A$ holds when $\vdash a_i$ holds for each $a_i \in A$. Likewise, forward arguments for $\sim p$ are collection of rules Δ that permit us to establish $\sim p$ from the facts W by *reasoning along the direction of the rules.* That is, such collection of rules Δ must contain a rule $C \rightarrow \sim p$ such that C logically follows from Δ and W. The strength of such rules determine the strength of the argument.

The theory ($\star$) can be processed by this procedure once the literal $\neg\mathtt{d}$ is replaced by a new atom $\mathtt{d}'$ declared to be incompatible with $\mathtt{d}$. P_0 then yields $\mathtt{a}$ and $\mathtt{c}$, as $W = \{\mathtt{a}, \mathtt{c}\}$, then $\mathtt{b}$, as r_3 is a reason for $\mathtt{b}$ and there are no forward arguments for $\sim\mathtt{b}$, and finally $\mathtt{c}$, as r_2 is a reason for $\mathtt{d}'$ and the only forward argument for $\sim\mathtt{d}'$ ($\mathtt{d}$) rests on the rule r_1 that is weaker than r_2.

The procedure P_0 is simple and intuitive, and as the example shows, captures inferences that escape the minimal model semantics. To understand why this happens it is useful to look at the latter from its proof-theoretic perspective. The proof-theory of minimal models can be understood in terms of *arguments:*

Definition 1 *A subset Δ of D is an* argument *against a default r in D given a set of formulas W, when $W+\Delta$ is logically consistent but $W + \Delta + \{r\}$ is not.*

If there is an argument against r, there will be a minimal model that violates r, and vice versa. In the presence of priorities, arguments need to be ordered by 'strength' and this criterion has to be modified slightly (Baker & Ginsberg 1989; Geffner 1992).

In the theory ($\star$) the 'spurious' model M_3 pops up because the rules r_1 and r_2 provide an argument against r_3. This argument is not considered by P_0 because it is not a *forward* argument. Since this is the right thing to do in this case, one may wonder whether non-forward arguments can always be ignored. The answer is no. For example, if $\mathtt{d}$ is observed, P_0 would still derive $\mathtt{b}$, implicitly favoring rule r_2 over r_3 with no justification. The same is true for approaches in which defaults are represented by one-directional rules of inference. Interestingly, in such a case, the minimal model semantics behaves correctly.

Causal Rule Systems

Preliminary Definitions

The problem of defining a semantics for causal default theories can be seen as the problem of distinguishing legitimate arguments from spurious ones. In this section we will look closer at this distinction in the context of a class of simple causal default theories that we call *causal rule systems* or CRSs for short. The language of causal rule systems is sufficiently powerful to model interesting domains and most scenarios analyzed in the literature but purposely excludes non-causal rules and disjunctions. We will report the extensions to handle these and other constructs elsewhere.

A *causal rule system* T comprises a set D of *defeasible causal rules* $A \rightarrow p$ where p is an atom and A is a conjunction of zero or more atoms, a set F of atomic *facts,* and a set C of *constraints* expressing that a given conjunction of atoms B cannot be true. Variables are assumed to be universally quantified. Rules and constraints will refer to members of D and C respectively, or to ground instances of them. Likewise, constraints will be divided between *background constraints* and *observed constraints.* The former will express domain constraints and will be denoted by rules $B \rightarrow$ with no head (e.g., $\mathtt{alive}(p,t) \wedge \mathtt{dead}(p,t) \rightarrow$), while the latter will express contingent constraints and will be denoted as $\neg B$ (e.g., $\neg[\mathtt{on}(\mathtt{a},\mathtt{b},\mathtt{t_1}) \wedge \mathtt{on}(\mathtt{b},\mathtt{c},\mathtt{t_1})]$). This distinction between background and evidence is implicit in probability theory and in Bayesian Networks (Pearl 1988b) and it used in several theories of default reasoning (Geffner 1992; Poole 1992). For simplicity, we will assume that background constraints $B \rightarrow$ involve exactly two atoms. We will say that such pairs of atoms are *incompatible* and use the expression $\sim p$ to denote atoms q *incompatible* with p. The generalization to n-ary background constraints is straightforward but makes the notation more complex and it is seldom needed.

Every rule will have a *priority* which will be represented by a non-negative integer; the higher the number associated with a rule, the higher its priority. This scheme is sufficiently simple and will not introduce the distortions common to total orderings of defaults because *the scope of priorities will be local:* priorities will only establish an order among rules $A \rightarrow p$ and $B \rightarrow \sim p$ whose consequents are incompatible. Priorities thus play a role analogous to local conditional probability matrices in Bayesian Nets allowing us to determine the net effect of conflicting causal influences acting on the same proposition. Unless otherwise stated, all priorities will be assumed equal.

Finally, as in other representations involving causal relations (e.g., (Shoham 1988; Pearl 1988b)), we will require that causal rules be *acyclic.* To make this precise, let us define the dependency graph of a CRS as a directed graph whose nodes are ground atoms, and where for every instance of a causal rule $A \rightarrow p$ there is a c-link relating every atom a in A to p, and for every instance of a background constraint $p \wedge q \rightarrow$, there are two k-links, one from p to q and another from q to p. Then, a CRS is *acyclic* iff its dependency graph does not contain cycles involving c-links. It's very simple to check that the encoding of acyclic inheritance network in CRSs is acyclic, like the encoding of theories about change (see below).

Semantics

The key to distinguishing legitimate arguments from spurious ones lies in an idea advanced by Pearl in a number of places which is at the basis of the model of intercausal relations captured by Bayesian Networks. It says that *two events should not interact through a common variable that they causally affect, if that variable or a consequence of it has not been observed* (Pearl 1988b, pp. 19). As we will see below, there are arguments in causal rule systems that violate this criterion. To identify and prune such arguments, it will be useful to recall the distinction between forward and backward arguments:

Definition 2 *An argument Δ against a rule $A \rightarrow p$ is a* forward *argument when Δ contains a rule $B \rightarrow \sim p$ such that B is supported by Δ.*[2] *An argument which is not a forward argument is a* backward *argument.*

In causal rule systems, all forward and backward arguments arise from rules violating some constraint. That is, a consistent collection of rules Δ will be an argument against a default $A \rightarrow p$ when the rules $\Delta + \{A \rightarrow p\}$ support a conjunction of atoms B ruled out by some constraint. Moreover, such a constraint can be either part of the evidence or part of the background. In a Bayesian Network, the former would be represented by an observation and the latter by the network itself.[3] It is simple to check then that the arguments that violate Pearl's criterion are precisely the *the backward arguments that do not originate in evidential constraints but in background constraints.* Such arguments establish a relation on events merely because they have *conflicting* causal influences on unobserved variables. To identify and prune those arguments, we will first make precise the distinction between background and evidential constraints.

Definition 3 *An* evidential constraint *is a formula $\neg B$, where B is a conjunction of atoms that is consistent with the background constraints but* inconsistent *with the background constraints* and *the evidence (facts and observed constraints).*

Basically, $\neg B$ is an evidential constraint when $\neg B$ is an observed constraint, or when $B \wedge A \rightarrow$ or $\neg(B \wedge A)$ are background or observed constraints, and A is a conjunction of one or more *facts.*

The backward arguments that arise from evidential constraints will be called *evidential arguments,* and the ones arising from background constraints will be called *spurious arguments.*

Definition 4 *A collection of rules Δ is* refuted by the evidence *or is an* evidential nogood *when Δ supports B, for an evidential constraint $\neg B$. A* backward argument Δ *against a rule r is* evidential *when $\Delta + \{r\}$ is refuted by the evidence, and is* spurious *otherwise.*

For example, in the theory that results from ($\star$) by replacing $\neg$d by an atom d$'$ incompatible with d, there are no evidential constraints and the backward argument against rule r_3 comprised of rules r_1 and r_2 is spurious. On the other hand, if d is observed, there will be an evidential constraint $\neg$d$'$ and r_2 will provide an evidential argument against r_3.

[2] A formula w is supported by Δ, when w logically follows Δ and the facts.

[3] The constraint that p and q cannot be both true can be captured in a Bayesian Network either by encoding p and q as values of a single variable, or by encoding them as values of two different variables with a third variable, *observed to be 'false'*, which is true when both p and q are true.

Before defining a semantics that reflects this distinction, let us define *causal arguments* as forward arguments with the appropriate strength:

Definition 5 *A forward argument Δ against a default $A \rightarrow p$ is a* causal argument *when Δ contains a rule $B \rightarrow \sim p$ not weaker than $A \rightarrow p$ such that B is supported by Δ.*

The basic principle in causal default reasoning can then be expressed as follows:

Only causal and evidential arguments need to be considered in causal default reasoning. In particular, rules not facing causal or evidential arguments should be accepted.

Although the given definitions of causal and evidential arguments are tied to the language of causal rule system, these notions, like the notions of causal and evidential support in Bayesian Networks, are more general. It should thus be possible to devise analogous definitions for more powerful languages.

The semantics of causal rule systems will reflect this principle. We will define it in two steps. Let us first say that *an argument Δ is validated by a model M* when M does not violate any rule in Δ, and that a rule r violated by a model M is *causally (evidentially) justified* when there is a causal (evidential) argument against r validated by M. Let us also say that a causal rule system is *predictive* when it does not give rise to evidential arguments; i.e., when no collection of rules is refuted by the evidence. Then, since in the absence of evidential arguments only causal arguments need to be considered, the semantics of predictive systems can be defined as follows:

Definition 6 *The* causal models *of a* predictive *causal rule system are the models in which every rule violated is causally justified.*

To extend this semantics to an arbitrary causal rule system T, we will use the expression T/Δ to denote the result of removing the rules in Δ from T. The minimal collection of rules Δ that render T/Δ a *predictive* system will be called *culprit sets.* It's easy to check that these culprit sets are exactly the minimal sets of rules that 'hit' all collections of rules refuted by the evidence (the evidential nogoods). The semantics of *arbitrary* causal rule systems can then be defined as follows:

Definition 7 *The causal models of an* arbitrary *causal rule system T are the causal models of the* predictive *systems T/Δ for any culprit set Δ.*

The system that results from the theory ($\star$) after changing $\neg$d by d$'$ is a predictive system with a single class of causal models where r_1 is the only violated rule. On the other hand, if the fact d is added, two culprit sets $\{r_2\}$ and $\{r_3\}$ arise, and the causal models of the resulting theory will satisfy r_1 and violate one of r_2 or r_3.

Definition 8 *A ground atom p is a* causal consequence *of a causal rule system T iff p holds in all causal models of T.*

Some Properties

Proposition 1 *Causal models always exist when the facts and constraints are logically consistent.*

Now let $\Delta[M]$ denote the collection of rules violated by a causal model M of a causal rule system T. Then, $T/\Delta[M]$ is a logically consistent set of Horn clauses, and thus, has a unique minimal Herbrand model M_H. Clearly M_H is a causal model of T as well, and moreover, it is a *canonical model* in the sense that only models like M_H need to be considered:

Proposition 2 *A ground atom p is a causal consequence of a causal rule system T if p is true in all its* canonical *causal models.*

Every consistent causal rule system will have one or more canonical causal models. Finding one of them can be done efficiently:

Theorem 1 *The problem of finding a canonical causal model for a finite propositional* CRS *is tractable.*

The importance of this result is that many theories of interest have a *unique* canonical causal model. Let us say that a theory is *deterministic* when for every pair of rules $A \rightarrow p$ and $B \rightarrow \sim p$ with incompatible consequents and compatible antecedents, one rule has priority over the other. Then for predictive theories that are deterministic we get:

Theorem 2 *Causal rule systems which are predictive and deterministic possess a single canonical causal model.*

Corollary 1 *The problem of determining whether a given ground atom is a causal consequence of a finite, predictive and deterministic causal rule system is tractable.*

A class of causal theories that are predictive and deterministic are the theories for reasoning about change which include no information about the 'future' (Lin & Shoham 1991; Sandewall 1993). They can be expressed as CRSs of the form:

Persistence:	$T(f,s) \rightarrow T(f,r(a,s))$
Action:	$T(p_1,s) \wedge \cdots \wedge T(p_n,s) \rightarrow T(f,r(a,s))$
Facts:	$T(f_1,s_0)$; $T(f_2,s_0)$; ...
Constraints:	$T(f,s) \wedge T(g,s) \rightarrow$

It's simple to verify that the resulting theories are acyclic, predictive and deterministic (assuming rules about change have priority over rules about persistence). An equivalent formulation based on time rather than on situations would have similar properties. Most semantics for reasoning about change coincide for theories like the one above: the preferred model is the one which is 'chronologically' minimal (Shoham 1988), where every 'exception' is explained (Gelfond & Lifschitz 1988), where actions are minimized (Lifschitz 1987), etc. Moreover, the proof-procedure P_0 presented in Section 1 is adequate for such systems (the facts F should take the place of W in P_0):

Theorem 3 *The proof-procedure P_0 is sound and complete for finite causal rule systems that are both predictive and deterministic. The procedure P_0 remains sound but not necessarily complete for systems which are predictive but not deterministic.*

Priorities Revisited

The results of the previous section seem to confirm that for predictive theories causal models are adequate. For more general theories however, they are not. A simple example illustrates the problem. Consider the chain of rules:

$$r_1 : \mathtt{a} \rightarrow \mathtt{b} \ , \ r_2 : \mathtt{b} \rightarrow \mathtt{c} \ , \ r_3 : \mathtt{c} \rightarrow \mathtt{d}$$

together with the rule $r_4 : \mathtt{f} \rightarrow \mathtt{c}'$, where $\mathtt{c}$ and $\mathtt{c}'$ are incompatible and r_4 has priority over r_2. Given $\mathtt{a}$ and $\mathtt{f}$, we get a predictive, deterministic theory, whose single (canonical) causal model M sanctions $\mathtt{b}$ and $\mathtt{c}'$. If the prediction $\mathtt{c}'$ is confirmed, however, another causal model M' pops up, which refutes the prediction $\mathtt{b}$. This is because the observation $\mathtt{c}'$ yields an evidential constraint $\neg\mathtt{c}$ that refutes the rules r_1 and r_2, producing two culprit sets $\{r_1\}$ and $\{r_2\}$.

The explanation for this behavior can also be found by looking at probability theory and Bayesian Networks. We have a 'gate' composed of two conflicting rules $r_4 : \mathtt{f} \rightarrow \mathtt{c}'$ and $r_2 : \mathtt{b} \rightarrow \mathtt{c}$, with r_4 having priority over r_2. The semantics partially reflects this priority by validating r_4 in all causal models. Yet, while it makes $\mathtt{b}$ irrelevant to $\mathtt{c}'$, it makes $\mathtt{c}'$ *relevant* to $\mathtt{b}$; i.e., $\mathtt{c}'$ refutes $\mathtt{b}$ even though $\mathtt{b}$ does not refute $\mathtt{c}'$. This is anomalous because irrelevance should be symmetric (Pearl 1988b). A common solution to this problem is to add 'cancellation axioms' (like making rule r_2 *inapplicable* when $\mathtt{f}$ holds). Here we develop a different solution that avoids making such axioms explicit.

Let us say that a rule $A \rightarrow p$ is a *defeater* of a conflicting rule $B \rightarrow \sim p$ when its priority is higher and that $A \rightarrow p$ is *verified* in a model when both A and p hold in the model. Then, we will be able to regard a collection of rules Δ as *irrelevant* when those rules are *defeated* as follows:[4]

Definition 9 *A collection of rules Δ is* defeated *or* preempted *in a causal rule system T when every causal model of T/Δ verifies a defeater for each rule in Δ.*

Then the second principle needed is:

[4]The notions of defeat and preemption are common in inheritance algorithms and argument systems (e.g., (Horty, Thomason, & Touretzky. 1987; Pollock 1987)). Here the corresponding notions are slightly more involved because rules are used to reason both forward and backward.

Arguments involving defeated rules are irrelevant

Certainly, *defeat is closed under union:*

Theorem 4 *If Δ_1 and Δ_2 are two sets of rules defeated in T, then the union $\Delta_1 + \Delta_2$ of those sets is also defeated in T.*

This means that in any causal rule system T there is always a unique maximal set of defeated rules which we will denote as $\Delta_0[T]$. The strengthened causal consequences of a causal rule system T can then be defined as follows:

Definition 10 (Revised) *A ground atom p is a* causal consequence *of a causal rule system T if p is true in all causal models of $T/\Delta_0[T]$.*

Since the second principle was motivated by problems that result from the presence of evidential arguments, it's reassuring that the new semantics is equivalent to the old one when such arguments are not present:

Theorem 5 *A ground atom p is a causal consequence of a* predictive *causal rule system T iff p is true in all causal models of T.*

To check whether a given atom p is a causal consequence however, it's not necessary to first identify the maximal set of rules defeated; any such set suffices:

Theorem 6 *If a ground atom p is true in all* causal models *of T/Δ, for any set of rules Δ defeated in T, then p is a causal consequence of T.*

We address finally the task of *computing* in general theories:

Theorem 7 *Let $P = \{p_1, p_2, \ldots, p_n\}$ be a collection of atoms derived by the procedure P_0 from a system T by application of a collection of rules Δ. Then each p_i in P is a causal consequence of T if P* shields *Δ from every evidential argument Δ' against rules in Δ; i.e., if every such Δ' contains a rule $B \rightarrow \sim p_i$ for some p_i in P.*

The new proof-procedure works in two stages: in the first it uses P_0 to derive tentative conclusions, ignoring evidential counterarguments. In the second, it verifies that all such counterarguments are defeated, and therefore, can legitimately be ignored.

In the theory above, the procedure P_0 yields b and c′ by application of the rules r_1 and r_4. To verify whether b and c′ are actual causal consequences of the theory, Theorem 7 says we only need to consider the evidential arguments against r_1 or r_4. Since the only such (minimal) argument $\Delta' = \{r_2, r_3\}$ contains a rule $r_2 : \mathsf{b} \rightarrow \mathsf{c}$ whose consequent c is incompatible with c′, we are done. Indeed, r_2 is defeated in the theory, and b and c′ follow.

Related Work

The first principle is a reformulation of Pearl's idea that 'causes do not interact through the variables they influence if these variables have not been observed' (Pearl 1988b, pp 19). Other attempts to use this idea in the context of causal default reasoning are (Pearl 1988a; Geffner 1992; Goldszmidt & Pearl 1992). The contribution of this work is to explain the patterns of causal default reasoning in terms of some simple and meaningful principles that tie up a number of important notions in default reasoning: minimal and 'coherent' models (see below), background vs. evidential knowledge, forward vs. backward reasoning, etc. Interestingly, the need to distinguish background from evidence also appears in systems that handle 'specificity' preferences (Poole 1992; Geffner 1992), and in a slightly different form, in certain systems for causal reasoning (Sandewal 1991; Konolige 1992).

Many approaches to causal default reasoning are based on a preference criterion that rewards the 'most coherent' models; namely, the models where the set of violated rules without a causal justification is minimal or empty (e.g., (Gelfond 1989), as approaches based on non-normal defaults and the stable model semantics). For predictive theories, the most coherent models and causal models coincide; indeed, the causal models of predictive theories are *defined* as the models where *all* violated rules have a causal justification. Yet the two types of models differ in the general case. For example, in a theory comprised of three rules $r_i : p_i \rightarrow q_i$, $i = 1, \ldots, 3$, where q_2 is incompatible with both q_1 and q_3, the most coherent model given the facts p_1, p_2, p_3 and $\neg q_1$, is the model M that validates the rule r_2. Yet, the model M' that validates r_3 is also a causal model. Likewise, if r_3 is given priority over r_2, both M and M' would be the most coherent models but only the former would be a causal model. In the two cases, causal models behave as if the rule r_1, which is violated by the evidence, was excluded. The 'coherence' semantics, on the other hand, does not exclude the rule but rewards the models where it gets a causal explanation. The result is that sometimes the coherence semantics is stronger than causal models and sometimes is weaker. More important though is that the coherence semantics violates the first principle; indeed, in the second case, the rule r_3 fails to be sanctioned even though r_3 does not face either causal or evidential arguments. As a result, the coherence semantics makes the proof-procedure described in Theorem 7 unsound.

Causal rule systems which are *predictive* can be easily compiled into logic programs with negation as failure. Every ground rule $r_i : A_i \rightarrow p_i$ can be mapped to a logic programming rule $p_i \leftarrow A_i, \neg ab_i$ along with rules $ab_j \leftarrow A_i, \neg ab_i$ for every conflicting rule $r_j : A_j \rightarrow p_j$ with priority equal or smaller than r_i. In addition, facts translate to rules with empty bodies. From the discussion above, it's clear that the causal consequences of the original theory would be exactly the atoms that are true in all the stable models of the resulting program (Gelfond & Lifschitz 1988). The same mapping will not work for non-predictive systems as the semantics and algorithms for logic programs are limited in the the way negative information is handled (Geffner 1991).

Causal rule systems are also related to ordered programs (Laenens & Vermeir 1990) where, using our language, facts are treated like rules, and all constraints are treated like background constraints. As a result, there are no evidential arguments and all reasoning is done along the direction of the rules. This is also common to approaches in which defaults are regarded as tentative but one-way rules of inference. We have argued that such interpretations of defaults may be adequate in the context of predictive theories but will fail in general: even if a default $a \rightarrow b$ does not provide a reason to conclude $\neg a$ from $\neg b$, it may well provide a reason to avoid concluding a when $\neg b$ is known.

Conclusions

We have presented a pair of principles that account for the basic preferences among defaults that arise in causal domains. We have also defined a semantics based on those principles and presented some useful proof-procedures to compute with it. The language of the formalism can be extended in a number of ways. Some extensions are more direct (e.g., n-ary background constraints, non-causal rules); others are more subtle (e.g., disjunctions). We are currently working on these extensions and will report them elsewhere. We are also looking for more efficient procedures that would avoid the need to precompute all 'evidential nogoods' when they exist.

Acknowledgments. Thanks to Wlodek Zadrozny, Pierre Siegel, Benjamin Grosof, and Kurt Konolige for discussions related to the content of this paper.

References

Baker, A., and Ginsberg, M. 1989. A theorem prover for prioritized circumscription. In *Proceedings IJCAI-89*, 463–467.

Dean, T., and McDermott, D. 1987. Temporal data base management. *Artificial Intelligence* 32:1–55.

Geffner, H. 1991. Beyond negation as failure. In *Proceedings of the Second International Conference on Principle of Knowledge Representation and Reasoning*, 218–229.

Geffner, H. 1992. *Reasoning with Defaults: Causal and Conditional Theories.* Cambridge, MA: MIT Press.

Gelfond, M., and Lifschitz, V. 1988. The stable model semantics for logic programming. In *Proceedings of the Fifth International Conference and Symposium on Logic Programming*, 1070–1080. Cambridge, Mass.: MIT Press.

Gelfond, M. 1989. Autoepistemic logic and formalization of commonsense reasoning. a preliminary report. In *et al.*, M. R., ed., *Proceedings of the Second International Workshop on Non-Monotonic Reasoning*, 177–186. Berlin, Germany: Springer Lecture Notes on Computer Science.

Goldszmidt, M., and Pearl, J. 1992. Stratified rankings for causal relations. In *Proceedings of the Fourth Workshop on Nonmonotonic Reasoning*, 99–110.

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

Horty, J.; Thomason, R.; and Touretzky., D. 1987. A skeptical theory of inheritance. In *Proceedings AAAI-87*, 358–363.

Konolige, K. 1992. Using default and causal reasoning in diagnosis. In *Proceedings of the Third International Conf. on Principles of Knowledge Representation and Reasoning.*

Laenens, E., and Vermeir, D. 1990. A fixpoint semantics for ordered logic. *Journal of Logic and Computation* 1(2):159–185.

Lifschitz, V. 1987. Formal theories of action. In *Proceedings of the 1987 Workshop on the Frame Problem in AI*, 35–57.

Lin, F., and Shoham, Y. 1991. Provably correct theories of action. In *Proceedings AAAI-91*, 349–354.

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

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

Pollock, J. 1987. Defeasible reasoning. *Cognitive Science* 11:481–518.

Poole, D. 1992. The effect of knowledge on belief: Conditioning, specificity and the lottery paradox. *Artificial Intelligence* 49:281–309.

Sandewal, E. 1991. Features and fluents. Technical Report R-91-29, CS Department, Linkoping University, Linkoping, Sweden.

Sandewall, E. 1993. The range of applicability of nonmonotonic logics for the inertia problem. In *Proceedings IJCAI-93*, 738–743.

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