Computer Science Department Technical Report
Cognitive Systems Laboratory
University of California
Los Angeles, CA 90024-1596

DEFAULT REASONING, MINIMALITY AND COHERENCE

Hector Geffner

June 1989
CSD-890037

To appear in Proceedings, 1st Intl. Conf. on Principles of Knowledge Representation and Reasoning (KR'89), Toronto, May 1989.

TECHNICAL REPORT R-125 February 1989.

Default Reasoning, Minimality and Coherence

Hector Geffner*
Cognitive Systems Lab.
Computer Science Department
UCLA
Los Angeles, CA 90024

Abstract

A preferential semantics for default reasoning is presented. A partial order is defined over *classes* of models which establishes a preference for classes with a minimal set of *unexplained* exceptions. Exceptions are explained in terms of justifications which are syntactically extracted from the knowledge base. The resulting semantics succeeds in pruning the spurious models which arise in minimal model semantics, legitimizing a behavior in closer correspondence with intuition. Likewise, the proposed framework unifies and extends ideas stemming from work in default reasoning, logic programming and abductive reasoning.

1 Introduction

In [McCarthy, 80;86] McCarthy proposed circumscription as a simple but powerful second order axiom capable of endowing first order logic with non-monotonic features. In model theoretic terms, circumscription can be understood as replacing the traditional notion of entailment as truth in all models by a weaker, defeasible form of entailment in which only a subset of minimal models is considered. [McCarthy, 80; Lifschitz, 85; Etherington, 88].

Since then, several studies have analyzed the mathematical properties of circumscription.[1] Less attention, however, has been given to the circumscriptive framework as a framework for representing commonsense knowledge. In this regard, recent work has illustrated that, more often than not, the inferences sanctioned by circumscription from relatively simple conceptualizations turn out to be weaker than expected (e.g. [Hanks and McDermott, 86; Haugh, 88]). Minimal, unintended models often arise which prevent certain intended conclusions from being certified.

This mismatch between intended meaning and the meaning uncovered by circumscription has recently prompted Shoham [88] to propose a close alternative to circumscription in which the notion of *minimal* model is replaced by an appropriate notion of *preferred* model. Shoham convincingly argues in favor of this semantic shift, and illustrates its convenience by considering a troubling problem in the domain of temporal reasoning raised by Hanks and McDermott [86].

More recently, these ideas have been further developed by Makinson [89] and Kraus *et al.* [88], who prove some interesting soundness and completeness results. Sandewal [88] has also proposed a preferential semantics for non-monotonic entailment, which he defines in terms of partial interpretations.

Nonetheless, no 'preferential semantics' attempting to capture *the intended meaning* of general default theories has yet been proposed. Defining such an account is the main goal of this paper. Our approach draws on McCarthy's [86] suggestion that default reasoning be formalized in terms of the minimization of 'abnormality.' We depart, however, from McCarthy's minimal model semantics in two ways. First, the preference ordering does not apply directly to models, but to *classes* of models, with each class embedding a commitment to certain set of assumptions. Second, the preference ordering favors classes of models which minimize *unexplained* abnormality, rather than plain abnormality. These explanations are assembled in terms of justifications which are syntactically extracted from the knowledge base. The result is an account which succeeds in eliminating the spurious models that arise in minimal model semantics, permitting a behavior in closer correspondence with intuition. In addition, the resulting framework unifies and extends ideas stemming from work in default reasoning, logic programming and abductive inference.

The paper is organized as follows. In section 2, we introduce the preference ordering. Such ordering applies to sets of models, which we call *classes*. We define the conditions under which an abnormality is regarded as *explained* in a given class, and the conditions which make a class *admissible*. In section 3, we illustrate the appeal of the proposed account by ana-

*This work was supported in part by National Science Foundation Grant # IRI-8610155.

[1] See [Reiter, 87] for a relevant bibliography.

lyzing examples from the domains of reasoning about action, inheritance hierarchies, logic programming and abductive reasoning, and by comparing the results to related proposals. Finally in section 4, we summarize the main ideas, discuss some of the controversial points and point out some of the remaining problems.

2 A Preferential Semantics for Default Reasoning

2.1 Definitions

The default theories we shall consider are comprised of two components: a background context K and an evidence set E. The background context corresponds to an intensional characterization of the domain of interest in the form of a set of defeasible and undefeasible rules, while the evidence set corresponds to an extensional characterization of the particular situation of interest in the form of factual assertions [Geffner and Pearl, 87].

Among the predicate symbols occurring in the theories of interest, a distinguished set AB of predicates is used to express assumptions and abnormality conditions. In the context of default reasoning such set would contain 'abnormality' predicates [McCarthy, 86], while in the context of logic programming it would contain all the predicate symbols of interest. For a predicate $\mathtt{ab_i}$ in AB, we shall refer to atoms of the form $\mathtt{ab_i(a)}$, where $\mathtt{a}$ stands for a vector of ground terms, as *exceptions* or *abnormalities,* and to their negations, $\neg\mathtt{ab_i(a)}$, as *assumptions.*

The background context K of a given theory is itself structured into four components. There is a *terminological* component given by a set of strict rules (e.g. 'penguins are birds'), a *default* component given by a set of defeasible rules (e.g. 'birds fly'), a set of user-supplied *explicit exceptions* or *justifications* (e.g. 'injured birds are abnormal birds with respect to flying'), and a set of *implicit justifications* derived from the defaults in K, in a way to be described below.

Given a theory T, we are interested in characterizing the set of consequences it certifies. We shall achieve such characterization by determining the set of assumptions which can be accepted in T. For that purpose, we shall introduce the notion of a *class* C with *gap* G, given by a set of exceptions $\{\delta_1, \ldots, \delta_n\}$, as the non-empty collection of models of T which validate all the assumptions $\neg\delta$, for $\delta \notin G$.

We shall also say that a proposition holds or is validated in a class when the proposition holds in all the model members of the class. The set of all the assumptions validated by a class constitutes what we shall refer as the class *support*. Thus, in proof theoretic terms, a proposition α holds in a class C of T when the support of C comprises a set of assumptions AS such that $T, AS \vdash \alpha$.

From the complementarity of gap and support it follows that a class with a minimal gap will have a maximal support, and vice versa. Classes with minimal gaps will be referred as *minimal classes*.

As an illustration, consider for instance a theory T with a default $\mathtt{A} \wedge \neg\mathtt{ab_1} \Rightarrow \mathtt{B}$, an explicit justification $\mathtt{C} \wedge \neg\mathtt{ab_2} \Rightarrow \mathtt{ab_1}$, and a body of evidence $E = \{\mathtt{A}, \mathtt{C}\}$. For such a theory, there are no models which can make both assumptions $\neg\mathtt{ab_1}$ and $\neg\mathtt{ab_2}$ true simultaneously. Thus, there is no class of T with an empty gap. There are, however, two minimal classes C_1 and C_2, with gaps $\{\mathtt{ab_1}\}$ and $\{\mathtt{ab_2}\}$, respectively. But note that the two classes of models are not equally meritorious. Intuitively we would expect the assumption $\neg\mathtt{ab_1}$ to be defeated by the explicit justification $\mathtt{C} \wedge \neg\mathtt{ab_2} \Rightarrow \mathtt{ab_1}$, as the latter expresses a condition under which the default $\mathtt{A} \wedge \neg\mathtt{ab_1} \Rightarrow \mathtt{B}$ is not to be applicable.

Our task hereafter will be to establish a partial order among classes which will permit us to uncover the intended models of a given theory. Since each class reflects a choice of assumptions, such an ordering can be usefully regarded as a preference ordering among different assumptions sets. Thus, in the example above, we would expect the preference ordering to favor the class C_1, committed to the assumption $\neg\mathtt{ab_2}$, over the other minimal class C_2, committed to the inferior assumption $\neg\mathtt{ab_1}$. Such preference will be indeed established on the basis that the exception $\mathtt{ab_1}$ is *explained* in the class C_1, while the exception $\mathtt{ab_2}$ is not explained in the class C_2. Thus, we shall say that C_1 is *more coherent* than C_2 and is, therefore, preferred over C_2. The conditions under which an exception is explained in a class are elaborated below.

We regard default instances such as $\mathrm{A} \wedge \neg\delta \Rightarrow \mathrm{B}$ as expectations, and exceptions such as δ as expectation failures. Essentially, we assume that an exception such as δ can be explained in a class in one of two ways. Either the class validates a proposition C, in the presence of an explicit justification of the form $\mathrm{C} \Rightarrow \delta$ in K, or the class validates the propositions A and D and the assumption $\neg\delta'$, in the presence of a competing default expectation $\mathrm{D} \wedge \neg\delta' \Rightarrow \mathrm{B}'$ in K, where B′ stands for a proposition incompatible with B in K (fig. 1).

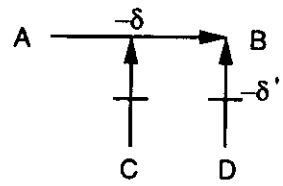

Figure 1: Explaining an exception δ

For the purposes of uniformity, however, the second case will be dealt with in a slightly different way. For each pair of defaults $\forall\mathtt{x}.\mathrm{A}(\mathtt{x}) \wedge \neg\mathtt{ab_i}(\mathtt{x}) \Rightarrow \mathrm{B}(\mathtt{x})$ and

$\forall x.D(x) \wedge \neg ab_k(x) \Rightarrow B'(x)$ with incompatible consequences in K, a formula of the form

$$\forall x.A(x) \wedge C(x) \Rightarrow ab_i(x) \vee ab_k(x),$$

referred as an *implicit justification*, will be added to K. Such formula is a deductive consequence of the rules in K, and has no effect on the models of T. Its addition, however, will permit a simple translation of the intuitions above into a formal definition.

Consider the ground instances of the rules in K. Each such ground rule might or might not involve certain assumptions in its body. For a set of assumptions AS we shall denote by K_{AS} the set of ground rules whose assumptions, if any, are among those of AS. We shall then say that an exception δ is *explained* in a class C of a context with background K and evidence E, if there is a set AS of assumptions validated by C, such that $E, K_{AS}, AS \vdash \delta$. In such case, we refer to the set AS of assumptions as an explanatory support of δ.

Note that an exception δ explained in a class C must belong to the class gap and, furthermore, it must hold in (every model of) the class. On the other hand, an exception that holds in a class, does not necessarily have an explanation in that class. Indeed, as hinted above, an exception δ can only be explained if there is a justification rule, explicit or implicit, whose consequent mentions δ. By definition, rules whose antecedents include the assumption $\neg\delta$ cannot take part of the set K_{AS} of rules needed to explain δ.

This definition of "explanation" introduces an important distinction between logically equivalent formulas, which can be illustrated by considering the rules $\neg\delta_1 \Rightarrow \delta_2$ and $\mathbf{true} \Rightarrow \delta_1 \vee \delta_2$, for two exceptions δ_1 and δ_2. The first rule permits an explanation for δ_2 in any class that validates the assumption $\neg\delta_1$. It does not permit however, an explanation of δ_1 in terms of $\neg\delta_2$; the rule $\neg\delta_1 \Rightarrow \delta_2$, with assumption $\neg\delta_1$, will belong to K_{AS} only if $\neg\delta_1$ belongs to AS. The second rule, on the other hand, does not involve any assumptions in its body and, as a result, permits explanations for δ_1 in terms of $\neg\delta_2$ and for δ_2 in terms of $\neg\delta_1$.

Recalling the example above, the reader can verify that the exception ab_1 is explained in C_1 and has an explanatory support $AS = \{\neg ab_2\}$. The exception ab_2, on the other hand, is *not* explained in the class C_2. Thus, C_1 is the only class with an empty *unexplained* gap, and it will therefore constitute the preferred class. We will make this notion more precise in the following definition.

Among two classes C and C' of a theory T, we say that C is *preferred* to C' when the unexplained gap of C is a strict subset of the unexplained gap of C'. In that case, we also say that the class C is *more coherent* than the class C'. If there is no class preferred to C, we say that C is a *preferred* class of T. Furthermore, we say that a proposition α is a consequence of T when α holds in all the preferred classes of T.

Note that it follows from these definitions that classes with smaller gaps are preferred to classes with larger gaps and, therefore, that preferred classes are always minimal. Furthermore, the preferred classes of a theory T can be determined by comparing minimal classes of T only.[2] Likewise, a class with an empty unexplained gap is always a preferred class. In these classes all exceptions are explained. We call these classes the *perfectly coherent* classes of T.

Example. Let us illustrate these definitions with the following example. We consider a theory T with a background context K comprising the following defaults (fig. 2):

1. $\forall x.\, A(x) \wedge \neg ab_1(x) \Rightarrow B(x)$
2. $\forall x.\, A(x) \wedge \neg ab_2(x) \Rightarrow C(x)$
3. $\forall x.\, B(x) \wedge \neg ab_3(x) \Rightarrow D(x)$
4. $\forall x.\, C(x) \wedge \neg ab_4(x) \Rightarrow \neg D(x)$

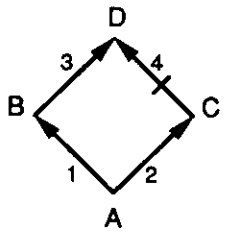

Figure 2: Simple diamond example

No undefeasible rules or explicit justifications are introduced, but the conflict between the last two defaults will result in the following implicit justification being added to K:

5. $\forall x.\, B(x) \wedge C(x) \Rightarrow ab_3(x) \vee ab_4(x)$.

Note that, as we said above, such a justification is already a deductive consequence of the formulas in K. Its role is not in affecting the models of T, but in permitting the constructions of explanations for exceptions $ab_3(x)$ and $ab_4(x)$, reflecting the conflicting expectations in which they participate.

Let us now consider in T a body of evidence $E = \{A(a)\}$. The goal is to determine the preferred classes of T. There are four minimal classes C_i in this context, with gaps $\{ab_i(a)\}$ for $i = 1, 2, 3, 4$ respectively. Furthermore, the exceptions $ab_1(a)$ and $ab_2(a)$ have no explanation in the classes C_1 and C_2, as there are no justifications for these exceptions in K. On the other hand, the exceptions $ab_3(a)$ and $ab_4(a)$ are explained in C_3 and C_4 respectively, by virtue of the justification encoded by (5). As a result, we end up with two preferred and indeed perfectly coherent classes C_3 and C_4, which sanction among other conclusions, $B(a)$

[2] This will no longer be true after admissibility constraints are introduced in section 2.2.

and C(a), and which suspend judgment regarding D(a). Note that this is indeed the behavior we would expect from the diamond structure encoded by the defaults 1–4 (fig. 2). A minimal model semantics, on the other hand, would propagate the uncertainty about D(a) to the propositions B(a) and C(a) as well.

2.2 Admissible Classes

Before proceeding with more interesting cases, we must address a problem that arises from a tradeoff between exceptions and explanations induced by the proposed preference ordering. We can illustrate this tradeoff by considering a theory T, with a background comprising a default (fig. 3):

1. $\mathtt{A} \wedge \neg \mathtt{ab}_1 \Rightarrow \mathtt{B}$

and explicit justifications:

2. $\mathtt{C} \wedge \neg \mathtt{ab}_2 \Rightarrow \mathtt{ab}_1$
3. $\mathtt{B} \Rightarrow \mathtt{ab}_3$,

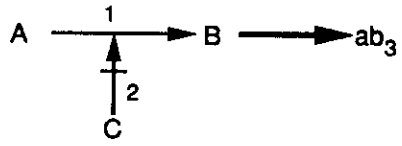

Figure 3: Spurious behavior and admissible classes

together with a body of evidence $E = \{\mathtt{A}, \mathtt{C}\}$. Such a theory gives rise to two minimal classes C_1 and C_2 with gaps $\{\mathtt{ab}_1\}$ and $\{\mathtt{ab}_2, \mathtt{ab}_3\}$, respectively. Furthermore, C_1 explains $\mathtt{ab}_1$ and C_2 explains $\mathtt{ab}_3$. The exception $\mathtt{ab}_2$, on the other hand, has no explanation in C_2. It follows then, that C_1, unlike C_2, has an empty unexplained gap and, therefore, that C_1 is the preferred class. This in turn can be interpreted as indicating, in agreement with intuition, that the proposition C defeats the default $\mathtt{A} \wedge \neg \mathtt{ab}_1 \Rightarrow \mathtt{B}$.

Consider now the case in which the exception $\mathtt{ab}_3$ is incorporated into the current evidence pool, so that the total evidence becomes $E = \{\mathtt{A}, \mathtt{C}, \mathtt{ab}_3\}$. In such a context, again, two minimal classes C_1' and C_2' arise; the former with a gap $\{\mathtt{ab}_1, \mathtt{ab}_3\}$, and the latter with a gap $\{\mathtt{ab}_2, \mathtt{ab}_3\}$. As before, $\mathtt{ab}_1$ is explained in C_1' and $\mathtt{ab}_3$ is explained in C_2'. Nonetheless, in the current context, neither class turns out to be preferred over the other. As a result, unexpectedly, the introduction of the exception $\mathtt{ab}_3$ has the effect of reinstating the default encoded by 1, which is no longer defeated.

This spurious effect can be explained in terms of the abductive bias embedded in the preference ordering, by which classes capable of explaining their exceptions are rewarded. In this case, the reinstatement of the assumption $\neg \mathtt{ab}_1$ permits the construction of an explanation for the exception $\mathtt{ab}_3$, but comes at the price of introducing the *unexplained* exception $\mathtt{ab}_2$. This tradeoff can be shown indeed to underlie this and other forms of abnormal behavior arising from the proposed preference ordering. In what follows a restriction on classes will be defined which will rule out such type of situations. Classes to be considered will have to be *admissible* in the sense defined below.

First, let us say that a class C of T with gap G *supersedes* a class C' of T with gap G', when the set $G - G'$ is not empty and only contains exceptions explained in C, while the set $G' - G$ contains exceptions unexplained in C'.

Thus, the gap of a class C' superseded by a class C can be constructed by eliminating some explained exceptions from C's gap, and by adding new exceptions, not all of them explained in C'. In terms of the example above, it can be verified that the class C_1' with gap $\{\mathtt{ab}_1, \mathtt{ab}_3\}$ supersedes the class C_2' with gap $\{\mathtt{ab}_2, \mathtt{ab}_3\}$. The latter gap can indeed be obtained from the former by removing the explained exception $\mathtt{ab}_1$ and by adding the unexplained exception $\mathtt{ab}_2$.

Finally, a class is *admissible* when it is not superseded by any other class. Hereafter, preferred classes will be selected by considering admissible classes only.

3 Applications

In the previous section we have laid out a semantic framework for the characterization of default theories. Our goal in this section is to illustrate how such a framework applies to a variety of domains ranging from problems in temporal reasoning, to problems in inheritance hierarchies, logic programming and abductive reasoning. Special emphasis will be placed on the type of behavior legitimized by the proposed account. Recall that our main goal is to arrive at an interpretation of the theories of interest which better approximates the intended interpretation.

3.1 Reasoning about Action

Our appeal to coherence considerations in pruning the set models of a given theory makes the proposed framework closely related to the proposals of Lifschitz [87], Haugh [87] and Morgenstern and Stein [88] for formalizing reasoning about action. In these proposals, clippings (persistence exceptions) can only originate from acting causes. Lifschitz and Haugh then minimize then over these causes, subject to explaining the clippings. Morgenstern and Stein take a slightly different view and select those models in which the actions are causally 'motivated' by the available evidence.

In our proposal, we do not require a cause behind every clipping, but 'reward' those classes of models in which this is the case and, therefore, those classes in which clippings are explained. We thus avoid some undesirable features of these approaches (Lifschitz's and Haugh's ontologies and Morgenstern's and Stein's lim-

itation to accommodate defeasible causal rules), while obtaining an additional degree of flexibility.

Consider the following version of the 'Yale Shooting Problem' raised by Hanks and McDermott [86]. We have a theory T with a background context K given by the following expressions:[3]

1. $\forall t.\, \mathtt{LD}(t) \Rightarrow \mathtt{LDD}(t+1)$
2. $\forall t.\, \mathtt{LDD}(t) \wedge \neg \mathtt{ab}_1(t) \Rightarrow \mathtt{LDD}(t+1)$
3. $\forall t.\, \mathtt{ALV}(t) \wedge \neg \mathtt{ab}_2(t) \Rightarrow \mathtt{ALV}(t+1)$
4. $\forall t.\, \mathtt{SHT}(t) \wedge \mathtt{LDD}(t) \wedge \neg \mathtt{ab}_3(t) \Rightarrow \neg \mathtt{ALV}(t+1)$
5. $\forall t.\, \mathtt{SHT}(t) \wedge \mathtt{LDD}(t) \Rightarrow \mathtt{ab}_2(t)$

Thus, we have that a loading event makes the gun loaded, that loaded guns remain loaded and that alive 'animals' remain alive unless shot with a loaded gun. Furthermore, due to the conflict between the persistence of 'alive' (3) and the shooting rule (4), an implicit justification of the following form is added to K:

6. $\forall t.\, \mathtt{SHT}(t) \wedge \mathtt{LDD}(t) \wedge \mathtt{ALV}(t) \Rightarrow \mathtt{ab}_3(t) \vee \mathtt{ab}_2(t)$

The evidence indicates that a turkey called Fred was alive at time $t = 1$, that a loading event took place at time $t = 2$, and that a shooting event directed at Fred took place at $t = 3$. Intuitively, it appears that Fred should no longer be alive as a result of the shooting. However, as Hanks and McDermott noted, several minimal classes pop up, in some of which Fred survives the shooting. In our formulation, these are classes in which the gun is mysteriously unloaded or in which the shooting, for some reason, misses its target. The collection of minimal classes of T thus corresponds to classes with a single exception among $\mathtt{ab}_1(2)$ ('mysterious unloading'), $\mathtt{ab}_2(1)$, $\mathtt{ab}_2(2)$ ('mysterious death'), $\mathtt{ab}_2(3)$ ('death by shooting'), and $\mathtt{ab}_3(3)$ ('target missed'). It is not difficult to show that the class corresponding to $\mathtt{ab}_2(3)$ is the only perfectly coherent class, and is thus the single preferred class of T. This is due to the fact that the persistence exception $\mathtt{ab}_2(3)$ can be explained in terms of the explicit justification encoded by (5). None of the other exceptions, on the other hand, can be given an explanation.

An undesirable feature of the above formulation which is shared by Hanks' and McDermott's, is the need to explicate by means of an explicit justification (5), that the shooting rule is supposed to prevail over the persistence of 'alive' (3). This is done by declaring the latter persistence to be abnormal in the context of a shooting. In a more realistic setting though, where a number of events leading to different type of changes are to coexist, the number of intended 'clippings' which must be enumerated could be overwhelming. Moreover, these explicit exceptions, as we show below, make theories less modular.

Consider for instance the possibility that Fred was wearing a metal vest at the time of the shooting. We could describe the effect of wearing a metal vest by asserting that the shooting rule is not applicable to somebody wearing a metal vest:

7. $\forall t.\, \mathtt{VEST}(t) \Rightarrow \mathtt{ab}_3(t)$

Such rule, however, would fail to achieve its intended effect. While the death of Fred would no longer follow, the justification encoded by (5) would still prevent proving Fred alive after the shooting. This suggests that a more flexible means of specifying the intended priority of rules about change is needed.

Our proposal is a simple one, consisting of two parts. First, we allow the user to lexically distinguish the assumptions associated with rules about change from the assumptions associated with rules about persistence.[4] We do so by replacing the generic type of normality assumption $\neg \mathtt{ab}_i(\cdot)$ with two different types of assumptions: $\neg \mathtt{cp}(\cdot)$, read 'not clipped', which is used for assumptions about persistence; and $\neg \mathtt{pv}(\cdot)$, read 'not prevented', which is used for assumptions about the result of actions. The formulation above would then be translated into the more concise description:

1'. $\forall t.\, \mathtt{LD}(t) \Rightarrow \mathtt{LDD}(t+1)$
2'. $\forall t.\, \mathtt{LDD}(t) \wedge \neg \mathtt{cp}_1(t) \Rightarrow \mathtt{LDD}(t+1)$
3'. $\forall t.\, \mathtt{ALV}(t) \wedge \neg \mathtt{cp}_2(t) \Rightarrow \mathtt{ALV}(t+1)$
4'. $\forall t.\, \mathtt{SHT}(t) \wedge \mathtt{LDD}(t) \wedge \neg \mathtt{pv}_3(t) \Rightarrow \neg \mathtt{ALV}(t+1)$

where the priority of the shooting rule over the aliveness persistence rule is not explicated.

As usual, the conflict between the last pair of default rules results in the addition of a corresponding implicit justification to K. Recall that these implicit justifications permit us to explain the failure of certain expectation in terms of the success of an alternative, incompatible expectation. Now, however, we have *two different types of expectations:* we have expectations of change on the one hand, and expectations of persistence on the other. The two expectations, however, are not intended to be treated symmetrically. While it is assumed that a successful change explains a corresponding clipping, it is also assumed that a failed

[3] Note that, unlike Hanks and McDermott, we have expressed the shooting rule (5) as a default rule. Independently of whether this is a more appropriate encoding, such a choice is motivated by the assumption embedded in the definition of 'explanations', by which expectations are assumed to be encoded by defaults. A slight extension would be needed to accommodate, for instance, undefeasible causal rules. We shall not pursue that extension in this paper. Let us just point out, however, that the shooting rule could be made undefeasible by simply declaring the proposition $\neg \exists t.\, \mathtt{ab}_3(t)$ as part of the evidence.

[4] If we were using a reified temporal notation in the style of [Shoham, 88], a single persistence rule would suffice. Nonetheless, in order to simplify the description of the example, we have found a non-reified notation more convenient and, therefore, a collection of persistence rules is needed.

action is not to be explained in terms of the persistence it fails to clip. We incorporate this asymmetry into our account by defining the implicit justifications associated with the conflict of a rule about change and a rule about persistence in a different manner. Thus, from the conflict between defaults (3') and (4') above, rather than eliciting the implicit justification:

$$5''.\ \forall \mathtt{t}.\ \mathtt{SHT(t)} \wedge \mathtt{LDD(t)} \wedge \mathtt{ALV(t)} \Rightarrow \mathtt{pv_3(t)} \vee \mathtt{cp_2(t)}$$

we assert the logically equivalent, but asymmetric justification:

$$5'.\ \forall \mathtt{t}.\ \mathtt{SHT(t)} \wedge \mathtt{LDD(t)} \wedge \mathtt{ALV(t)} \wedge \neg \mathtt{pv_3(t)} \Rightarrow \mathtt{cp_2(t)}$$

Thus, we allow the clipping of 'alive', $\mathtt{cp_2}(\cdot)$, to be explained in terms of a 'successful' shooting, but preclude the 'successful' persistence of 'alive' from explaining an 'unsuccessful' shooting, $\mathtt{pv_3}(\cdot)$. The reader can verify that from a theory with the background context defined by the formulas 1'–5', and given the same evidence as above, the same conclusion about Fred follows. The difference now is that certain preferences are handled implicitly, and that the resulting formulation is more flexible. The metal vest variation, for instance, would work without modification in this setting.

3.2 Inheritance Hierarchies

Another area in which minimal model semantics falls short of delivering the intended models of a set of defaults is in the context of inheritance hierarchies. Inheritance hierarchies are convenient devices for organizing knowledge about prototypical classes of individuals. Rather than explicitly stating the attributes of each possible individual, individuals are assumed to implicitly inherit a certain set of attributes by virtue of the place they occupy in the hierarchy. The key problem to address in these structures arises when an object belongs to classes with incompatible attributes. The typical example goes like this: Tweety is a penguin and, therefore, a bird. Typically birds fly and typically penguins do not fly. What should be concluded about the flying abilities of Tweety?

It is commonly accepted that there is an *implicit* preference among the defaults represented in these networks. Such preference appears to establish a priority for defaults rooted in more specific information [Touretzky, 86]. In terms of the example above, such preference would favor for instance the belief that Tweety is likely not to fly, on the basis that penguins are a subclass of birds. For more complex cases, the preferences are not always so clear, though significant progress has been made in recent years, both in the context of inheritance hierarchies [Horty *et al.*,87] and in more general settings (e.g. [Loui, 87; Delgrande, 87; Geffner and Pearl, 87]).

An important insight into the nature of the intended preferences among conflicting defaults that has emerged from these proposals is that a default 'if A then B' constitutes a license to infer B when A represents *all* the available evidence. In other words, a default antecedent provides a safe context on which the truth of the default consequence can be asserted. We refer to this aspect of defaults as the *context sensitivity* property of defaults.

In the context of the framework we have been developing, accounting for the context sensitivity of a default 'if A then B' would amount to making B true in all the preferred classes of A.[5] With that goal in mind, we shall impose a further restriction on the classes to be considered when dealing with theories, such as inheritance hierarchies, where there is an implicit preference to be uncovered among defaults.

Let $\mathrm{A} \wedge \neg\delta \Rightarrow \mathrm{B}$ be the ground instance of a default in K such that $K, \mathrm{A} \not\vdash \delta$. We say that a set of assumptions $AS = \{\neg\delta_1, \ldots, \neg\delta_n\}$ is in *conflict* with the assumption $\neg\delta$, if the assumptions in AS compete with $\neg\delta$ upon learning A, i.e. if $K, \mathrm{A} \vdash \delta \vee \delta_1 \vee \cdots \vee \delta_n$. Since $\neg\delta$ is the intended assumption in such context, it is reasonable to assume that the user intends to reject some of the assumptions in AS. We say then, that the set of assumptions AS is *dominated* by A. A *context-admissible* class C is then defined simply as an admissible class which does not validate any assumption set dominated by propositions that hold in C.[6]

For inheritance theories, only context-admissible classes will be considered. Note that in order to test context-admissibility, it is sufficient to examine *minimal* dominated assumption sets only.

Example. This example illustrates the type of specificity preferences entailed by the context-admissibility restriction. Let T be a theory with a background context K given by the following defaults (fig. 4):

$$\forall \mathtt{x}.\ \mathtt{A(x)} \wedge \neg \mathtt{ab_1(x)} \Rightarrow \mathtt{B(x)}$$
$$\forall \mathtt{x}.\ \mathtt{B(x)} \wedge \neg \mathtt{ab_2(x)} \Rightarrow \mathtt{C(x)}$$
$$\forall \mathtt{x}.\ \mathtt{B(x)} \wedge \neg \mathtt{ab_3(x)} \Rightarrow \mathtt{D(x)}$$
$$\forall \mathtt{x}.\ \mathtt{C(x)} \wedge \neg \mathtt{ab_4(x)} \Rightarrow \neg \mathtt{D(x)}$$
$$\forall \mathtt{x}.\ \mathtt{F(x)} \wedge \neg \mathtt{ab_5(x)} \Rightarrow \mathtt{C(x)}$$

Due to the conflict between the defaults associated with the assumptions $\mathtt{ab_3}(x)$ and $\mathtt{ab_4}(x)$ (fig. 4), the following implicit justification will also be part of K:

$$\forall \mathtt{x}.\ \mathtt{B(x)} \wedge \mathtt{C(x)} \Rightarrow \mathtt{ab_3(x)} \vee \mathtt{ab_4(x)}$$

We consider a body of evidence $E = \{\mathtt{A(a)}, \mathtt{F(a)}\}$.

[5] Kraus *et al.* [88] interpret defaults in a similar manner. Selman and Kautz [88], on the other hand, account for specificity preferences by intepreting defaults as imposing an ordering over pairs of models.

[6] Let us point out that in a pathological net with default instances $\mathrm{A} \wedge \neg\delta \Rightarrow B$ and $\mathrm{A} \wedge \neg\delta' \Rightarrow \neg B$, a context-admissible class would be forced to reject both assumptions $\neg\delta$ and $\neg\delta'$ upon learning A. With good reason, such networks are inconsistent in the frameworks of Horty *et al.* [87], Delgrande [87] and Geffner and Pearl [87].

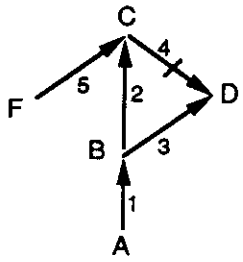

Figure 4: A simple inheritance hierarchy

Intuitively, we would expect the preferred classes of T to sanction the propositions $\mathtt{B(a)}$, $\mathtt{C(a)}$ and $\mathtt{D(a)}$. There are however four minimal classes of T, among which are two perfectly coherent classes C_3 and C_4, with explained gaps $\{\mathtt{ab_3(a)}\}$ and $\{\mathtt{ab_4(a)}\}$ respectively. C_4 represents the intended class of models, while C_3, which sanctions $\neg\mathtt{D(a)}$, fails to embed the right specificity preferences. We show below that C_3 is not a context-admissible class.

Consider the default instance $\mathtt{B(a)} \wedge \neg\mathtt{ab_3(a)} \Rightarrow \mathtt{D(a)}$. It follows from the body of defaults in K that the assumption set $AS = \{\neg\mathtt{ab_2(a)}, \neg\mathtt{ab_4(a)}\}$ is in conflict with the assumption $\neg\mathtt{ab_3(a)}$, and therefore, that the set AS is dominated by the proposition $\mathtt{B(a)}$. The class C_3, however, validates both $\mathtt{B(a)}$ and AS, and thus is not context-admissible. This leaves C_4, the intended class, as the single preferred class of T.

Let us remark that even if we add the following explicit justification to T:

$$\forall \mathtt{x}.\, \mathtt{B(x)} \Rightarrow \mathtt{ab_4(x)}$$

a simple minimization of abnormality would still not yield the expected conclusions in this case. Among the minimal models of T there would still be models validating both the abnormality $\mathtt{ab_1(a)}$ and the proposition $\neg\mathtt{B(a)}$.

3.3 Closed World Reasoning

The reader may have noticed that the preference ordering introduced above does not involve a minimization of the *extension* of the (abnormality) predicates in AB, but rather, a minimization of the set of *truths*. In other words, the gap of a class is not defined in terms of the exceptional individuals in the domain, but in terms of a set of ground exceptions. As a result, and in contrast to minimal model semantics, a default like 'birds fly' allow us to conclude that 'Tweety flies' upon learning that 'Tweety is a bird', without ever being committed to the conclusion that 'all birds fly'. A model in which only certain unnamed birds do not fly is as preferred as a model in which all birds fly, and both models would be indeed part of the same preferred class.

The same choice also permits jumping to conclusions without the need of unique name axioms [Reiter, 80b]. If Tim is a penguin, we do not need to prove that Tweety is different from Tim in order to jump to the conclusion that Tweety flies. This again contrasts with the form the same inference would be certified by a minimal model semantics, in which such inequality would be required.

Both of these features suggest that, in terms of jumping to conclusions, the proposed framework bears a closer similarity to Reiter's default logic than to McCarthy's circumscription. Furthermore, as we shall see, the framework inherits a difficulty of Reiter's logic in handling conclusions regarding unbounded sets of individuals, pointed out by McCarthy in [McCarthy, 80].

Let us say, for instance, that we want to capture the type of default behavior found in some relational database. We might have in the database a collection of tuples of the form $\mathtt{P(a)}, \mathtt{P(b)}$ and $\mathtt{Q(a)}, \mathtt{Q(b)}, \mathtt{Q(c)}$. From such database, conclusions like as that "a and b are the only P's," or that "all the P's are Q's" would follow. These are conclusions that non-monotonically depend on the state of the database, and that can potentially be defeated by the acquisition of new tuples (e.g. $\mathtt{P(d)}$).

A minimal model semantics would have no difficulty in accounting for such behavior. A simple minimization of the extensions of P and Q, together with the appropriate unique-name axioms will do. In our framework however, the straightforward approach of declaring the predicates P and Q as 'abnormality' predicates, members of AB, would not quite work. From such a declaration, we could derive conclusions such as $\neg\mathtt{P}(x)$, for any x different than a and b, but not *universals* such as $\forall \mathtt{x}.\, \mathtt{P(x)} \Leftrightarrow \mathtt{x} = \mathtt{a} \vee \mathtt{x} = \mathtt{b}$, which involve a commitment regarding unnamed individuals in the relevant models. In the remainder of this subsection we show that it is possible to capture this type of *closed world reasoning* in the present setting. The key, as hinted in [McDermott, 82], consists of incorporating *sets* into the universe of discourse. We shall not elaborate here on the details of how such an extension can be defined; suffice it to say that any weak set theory will do.[7]

In order to illustrate how the behavior of the database described can be captured in terms of defaults involving reference to sets of individuals, we shall introduce the following two abbreviations:

$$\begin{array}{ll} \mathtt{P[S]} & : \quad \forall \mathtt{x}.\, \mathtt{x} \in \mathtt{S} \Rightarrow \mathtt{P(x)} \\ \hat{\mathtt{P}}\mathtt{[S]} & : \quad \forall \mathtt{x}.\, \mathtt{x} \in \mathtt{S} \Leftrightarrow \mathtt{P(x)} \end{array}$$

where S stands for an arbitrary set of individuals. Thus, $\hat{\mathtt{P}}$ represents the definition, or as we shall we say, the 'closed' version of P. Having these abbreviations available, we can capture the database behavior by a theory with background:

[7] The interested reader might want to consult [Perlis, 88] for a relevant discussion.

$$\forall S. \quad P[S] \wedge \neg ab_1(S) \Rightarrow \hat{P}[S]$$
$$\forall S, S'. \quad P[S] \wedge S \supset S' \Rightarrow ab_1(S')$$

That is, if the members of a set S' are all instances of P, then it is assumed that S' contains *all* the instances of P *unless* there is a larger set S whose members are also instances of P. We shall also need a unique name hypothesis in order to distinguish different sets:

$$\forall x, y. \neg ab_2(x, y) \Rightarrow x \neq y$$

Notice that this default introduces a set of unexplained exceptions of the form $ab_2(x, x)$ in every class. We refer below to these exceptions as the common exceptions.

We can now analyze different states of the database as well as the conclusions which are sanctioned in each case.

Case 1. $E = \emptyset$. Without any tuples in the database, the preferred class includes no exception in addition to the common exceptions mentioned above. Thus, $\neg ab_1(\emptyset)$ forms part of the support of the preferred class and, therefore, $\hat{P}[\emptyset]$ and $\neg\exists x. P(x)$ follow.

Case 2. $E = \{P(a)\}$. $ab_1(\emptyset)$ becomes an explained exception, part of the gap of the preferred class. Still, the assumption $\neg ab_1(\{a\})$ holds, and the conclusion $\hat{P}[a] : \forall x. P(x) \Leftrightarrow x = a$ follows.

Case 3. $E = \{P(a), P(b)\}$. Now the gap of the preferred class is enhanced by two new explained exceptions $ab_1(\{a\})$ and $ab_1(\{b\})$. The assumption $\neg ab_1(\{a, b\})$ still holds, so that $\hat{P}[a, b] : \forall x. P(x) \Leftrightarrow x = a \vee x = b$ follows.

Case 4. $E = \{P(a) \vee P(b)\}$. In this context, the preferred class includes only the explained exception $ab_1(\emptyset)$ in addition to the common exceptions. Its models can be divided into two sets: those in which $P(a) \wedge \neg P(b)$ holds, and those in which $P(b) \wedge \neg P(a)$ holds. In the first class of models, $\hat{P}[a]$ holds as well, and in the second set of models, $\hat{P}[b]$ does. As a result, the disjunction $\hat{P}[a] \vee \hat{P}[b]$, which abbreviates the expression $[\forall x. P(x) \Leftrightarrow x = a] \vee [\forall x. P(x) \Leftrightarrow x = b]$, holds in the preferred class.

These results illustrate that it is possible to capture in the present framework the form of closed world reasoning found in databases. Circumscription will sanction the same conclusions in each of these cases [Lifschitz, 85]. In other cases, however, the results might differ. One such case, for instance, would correspond to $E = \{\exists x.P(x)\}$. Given such a context, circumscription would conclude that there is a single instance of P, i.e. $\exists x.\forall y.P(y) \Leftrightarrow x = y$. However, such a conclusion does not follow from the account presented.

3.4 Logic Programming

The semantic framework proposed can also be applied to logic programs with negation (see [Shepherson, 88] for a review). For logic programs, the set AB of predicates whose truth sets are expected to be minimal is identical to the set of all predicates of interest. A logic program is thus a collection of what we have been calling explicit justifications. In what follows, for ease of comparison with other proposals, we consider only *Herbrand* models.

The first result we show relates the proposed semantics to the stable semantics for logic programs proposed by Gelfond and Lifschitz [88] and, independently, by Kit Fine [88]. Elsewhere [Van Gelder *et al.*, 88], it has been proved that the stable model of stratified logic programs is unique, and that it coincides with the canonical model of Apt *et al.* [88] and the perfect model of Przymusinski [88].

Following Gelfond and Lifschitz we assume that each rule in the program P of interest has been replaced by all its ground instances. A Herbrand model M of P is then defined as stable if and only if M is the minimal model of the program P_M^+. P_M^+ is the positive program obtained by removing from P all the rules whose bodies contain assumptions $\neg\delta$, with $\delta \in M$, and by deleting the assumptions (i.e. negative literals) from the remaining rules.

If M is a model of a program P, we will denote by C_M the class of models of P with a gap equal to M. Thus M, as well as models of P smaller than M, will belong to C_M. The following theorem then holds:[8]

> **Theorem 1.** *M is stable if and only if the class C_M is perfectly coherent.*

In words, the theorem says that M is stable if each atom of M has an explanation in terms of the assumptions validated by M. Note that since a stable model is always minimal, the class C_M will contain a single model, namely M.

Still, there are programs which have no stable models. These programs might nonetheless have a well defined set of preferred classes. One typical example is the program P, composed of the single clause $p \Leftarrow \neg p$. The preferred class of P has the single *unexplained* atom p in its gap.

The correspondence between stable models and perfectly coherent classes suggests that the criterion of stability which is used in defining the stable semantics of logic programs embeds an abductive bias by which models capable of explaining their atoms are rewarded. This feature becomes apparent when we consider the following two programs:[9]

$$P_1: \quad q \Leftarrow \neg r \qquad P_2: \quad q \Leftarrow \neg r$$
$$r \Leftarrow \neg q \qquad\qquad p \Leftarrow \neg q$$
$$p \Leftarrow \neg p \qquad\qquad p \Leftarrow \neg p$$
$$p \Leftarrow \neg r$$

In both programs, the clause $p \Leftarrow \neg p$ introduces, but does not explain, the atom p. This leads the stable semantics to produce results in both cases which differ

[8] Proofs are omitted due to lack of space. They can be obtained by writing to the author.

[9] P_1 is taken from [Van Gelder *et al.*, 88].

from those which would be obtained if this clause were replaced by the simpler clause $\mathtt{p} \Leftarrow$. The problem is that our preference ordering, as well as the stable semantics for logic programs, rewards those classes in which p *gets* an explanation. Thus in P_1, both semantics favors the model $M_1 = \{\mathtt{q}, \mathtt{p}\}$ over the apparently equally meritorious model $M_1' = \{\mathtt{r}, \mathtt{p}\}$, while in P_2, the apparently superior model $M_2 = \{\mathtt{q}, \mathtt{p}\}$ fails to receive a better ranking than the model $M_2' = \{\mathtt{r}, \mathtt{p}\}$.

These examples appear to suggest that a more intuitive preference criterion for selecting the intended models of general logic programs should have this abductive bias removed. We discussed the effects of such a bias when the admissibility restriction was introduced in section 2.2. Recall that an admissible class is a class not superseded by any other class. Likewise, a class C' with gap G' is superseded by a class C with gap G when the set $G - G'$ is non-empty and only contains exceptions explained in C, while the set $G' - G$ contains exceptions unexplained in C',

In the examples above, the *minimal admissible classes* turn out to be in precise correspondence with the more intuitive models M_1, M_1' and M_2. The class with gap M_2', on the other hand, is superseded by the class with gap M_2, and is therefore not admissible.

Interestingly enough, the perfectly coherent class of a stratified program is also its unique minimal admissible class. That is, there is a correspondence between the minimal admissible class of a stratified program P and the canonical model of P, as defined by Apt *et al.*, Przymusinski and others. This correspondence is summarized in the following theorem:

> **Theorem 2.** For a stratified program P, there is a unique minimal admissible class, whose gap is the canonical model of P.

Thus we have two alternative semantics for general logic programs: one based on the preference ordering formerly introduced, the other which simply selects the minimal admissible classes. Both semantics coincide for the family of stratified programs, but diverge outside that family. The examples above suggest that a semantics based on minimal admissible classes is free from the abductive bias exhibited by the stable semantics and the preferential semantics here proposed and, therefore, that it might constitute a more appropriate basis for identifying the intended model(s) of general logic programs.

As a final illustration, we will consider a program P in which none of the minimal classes is admissible and in which, therefore, the preferred class is non-minimal. P is given by the following rules:

$$\mathtt{p} \Leftarrow \neg\mathtt{q}$$
$$\mathtt{q} \Leftarrow \neg\mathtt{r}$$
$$\mathtt{r} \Leftarrow \neg\mathtt{p}$$

The minimal Herbrand models of P are $M_1 = \{\mathtt{p}, \mathtt{r}\}$, $M_2 = \{\mathtt{q}, \mathtt{p}\}$ and $M_3 = \{\mathtt{q}, \mathtt{r}\}$, while the minimal classes are C_{M_1}, C_{M_2} and C_{M_3}. It can be shown that none of these classes is admissible: C_{M_2} supersedes C_{M_1}, C_{M_3} supersedes C_{M_2}, and C_{M_1} supersedes C_{M_3}. Thus, the minimal admissible class of P which is also the admissible class favored by preference ordering turns out to be the non-minimal class C_M, with gap $M = \{\mathtt{p}, \mathtt{q}, \mathtt{r}\}$. C_M stands in this case for a *collection of models;* indeed, it represents the set of subsets of M which are models of P, i.e. $C_M = \{M, M_1, M_2, M_3\}$.[10] It follows then, that none of p, q, r, or any of their negations are sanctioned as consequences of P.

3.5 Abductive Reasoning

Work in non-monotonic reasoning has been inspired by the goal of providing a formal account of some of the pervasive patterns of inference found in commonsense reasoning. Most of this work to date has been focused on the characterization of what has been called default inference, a form of reasoning akin to deductive inference, in which certain assumptions are adopted in the absence of contrary evidence. Nonetheless, other forms of non-monotonic inference, qualitatively different from default reasoning, also appear to play an important role in commonsense inference. One such form, analyzed in some detail in [Harman, 86], is what has been variously referred to as "inference to the best explanation," "abductive reasoning" or "conjectural reasoning." This is a form of inference which attempts to make sense of the evidence by increasing the coherence of a given set of beliefs. The characterization of these patterns of inference involves both the determination of sources of incoherence in a given belief state and the identification of hypotheses capable of explaining such incoherence away. In this subsection, we shall attempt to show that the framework we have so far developed lends itself to a characterization of this sort.

We assume that the unexplained gaps associated with the preferred classes of a given context provide a useful measure of the coherence of such context; indeed, they point out 'what needs to be explained.' For instance, in a inheritance hierarchy about animals, a context which mentions a bird Tweety that does not fly would be slightly incoherent. In such an incoherent state, it might make sense to jump to conclusions which could explain away the source of incoherence. We could hypothesize for instance, that Tweety is sick, or that he is penguin and so on. We shall refer to those propositions as *conjectures*. More precisely, a ground

[10] Recall from section 2, that for a theory T, a class C with gap G stands for the non-empty collection of models of T which validate all the assumptions $\neg\delta$, for $\delta \notin G$. C_M thus represents the collection of models of P which validate all the literals $\neg\alpha$, for $\alpha \notin M$, i.e. all the Herbrand models of P included in M.

atomic proposition γ would be regarded as a conjecture in a context T if its adoption yields a new context $T \cup \{\gamma\}$ *more coherent* than the original context T.

In section 2.1, we defined the conditions that make one class more coherent than another. Now we must define a similar order between *contexts*. For this purpose, we will associate with every context T a coherence descriptor $H[T]$, given by the vector of unexplained gaps in its preferred classes. A context T with a coherence descriptor $[G_1, \ldots, G_n]$ would then be said to be *as coherent* as a context T' with a coherence descriptor $[G'_1, \ldots, G'_m]$, if each G_i, $1 \leq i \leq n$, is included in G'_j, for some j, $1 \leq j \leq m$. Furthermore, if T is as coherent as T' but T' is not as coherent as T, we say that T is *more coherent* than T'.

Recalling, the example above, we obtain then that the proposition 'Tweety is a penguin' would qualify as a conjecture in the above context, since its adoption would lead to a context whose preferred classes are perfectly coherent. Nonetheless, a proposition such as 'Tweety is a brown arctic penguin' would also qualify as a conjecture. In order to rule out unnecessarily specific conjectures, we must restrict the space of admissible conjectures. Let us say that a conjecture γ' is less specific than a conjecture γ in a context T if γ' follows from $T \cup \{\gamma\}$, but γ does not follow from $T \cup \{\gamma'\}$. Then, we say that a conjecture γ is *admissible* if there are no less specific conjectures leading to contexts as coherent as those resulting from the adoption of γ. Thus, while 'Tweety is a penguin' and 'Tweety is sick' would represent admissible conjectures in the above context, the proposition 'Tweety is a brown arctic penguin' would not.

Note that there is an important distinction between the set of default conclusions that follow from a given theory and the set of admissible conjectures legitimized by it. The set of admissible conjectures, unlike the set of default consequences, is not deductively closed. Indeed, while it is reasonable to conjecture that Tweety does not fly because it is sick, or because it is a penguin, it is not so reasonable to conjecture that Tweety does not fly because it is a sick penguin. Conjectures, unlike defaults, represent *alternative* belief changes.

While this account of conjectural reasoning does not limit the space of admissible conjectures a priori,[11] it may nonetheless be useful to provide the user with the facility of expressing conditions under which certain conjectures would be preferred over others. For instance, following Pearl [87], we might want to express things like 'if you are not aware of an explanation for the grass being wet, then conjecture that it rained'. Thus, if the grass is observed to be wet, even in the presence of other admissible conjectures, an explanatory conjecture stating that it rained would be adopted. However, if an alternative explanation is learned, say that the sprinkler was on, the reason for having postulated 'rain' would vanish, and so the 'rain' conjecture. Pearl refers to these defaults as *evidential* defaults, and convincingly argues that they require a treatment different from the one given to normal (causal) defaults. In our framework, evidential defaults turn out to be essentially context guided conjectures. The details on how they can be accommodated as part of the language are elaborated in [Geffner, 89].

[11] See [Poole, 87] for a different view.

4 Discussion

We have presented a framework for characterizing defeasible inference based on a preference ordering among classes of models. The ordering favors classes with a minimal unexplained set of exceptions. We have shown how such a framework permits to unify ideas stemming from work in default reasoning, logic programming and abductive inference. We have also illustrated how the proposed account eliminates the spurious models that arise in minimal model semantics, permitting a behavior in closer correspondence with intuition.

The account presented here is unorthodox in several ways. First, a preferential ordering is described which does not apply directly to models, but to classes of models. The motivation for such a choice originates from viewing default reasoning in the 'abnormality' setting as a labeling problem, in which the set of legitimate assumptions in a given context needs to be identified. Each class of models thus represents a choice of assumptions, and these choices are evaluated according to the preference ordering.[12]

The distinction between explained and unexplained abnormality plays a central role in such ordering. We have argued that the value of a class is not in inverse proportion to its abnormalities, but rather to its *unexplained* abnormalities. No penalty, for example, we have maintained should be associated with a class in which a bird does not fly, if the bird is, say, a penguin. In that situation, being an 'abnormal bird with respect to flying' is the normal, expected condition. Abnormalities are unlikely in certain contexts but likely in others, and a reasonable preference ordering should be able to make this distinction.

Our reliance on justifications which are syntactically extracted from the database and used to construct ex-

[12] This view also suggests an alternative, stronger definition of default entailment which we have not pursued in this paper. Rather than defining α to be a default consequence of T when α holds in all the preferred classes of T (section 2.1), we could require the existence of a set of assumptions AS validated in all the preferred classes of T, such that $T, AS \vdash \alpha$. This stronger definition appears to bear some resemblance with those semantic accounts based on partial models (e.g. [Van Gelder *et al.*, 88]), which we have not yet investigated.

planations is potentially more controversial. We have assumed that abnormalities represent expectation failures, and as such, could be either explained by explicit exceptions which assert that a default is not applicable in a given circumstance, or by competing expectations. This choice, however, is not unique and, quite possibly is not the best. It is, however, relatively simple and intuitive, and as we have illustrated, it can 'reasonably' account for 'reasonable' examples. Further refinements may still be necessary.

The framework for defeasible inference proposed here shares several features with the system **L** proposed in [Geffner, 88]. Both systems represent defaults in the same way and they appeal to the same distinction between background and evidence. Both regard the antecedent of a default as providing a safe context in which the truth of the consequent can be asserted, and both attempt to capture the distinction between defaults and their contrapositives in a similar way.

There are, nonetheless, significant differences between the two frameworks. First, **L** has the form of a natural deduction system whose rules originate from a probabilistic interpretation of defaults.[13] This set of rules is supplemented by an additional, more ad-hoc rule, which attempts to supply the probabilistic rules with appropriate assumptions about conditional independence. Such an "irrelevance rule," as it is called, permits us to infer for instance conclusions like 'a red bird flies,' given that 'birds fly'. These conclusions would otherwise escape the probabilistic machinery.

The irrelevance rule in **L**, and an analogous construction in the conditional logic of Delgrande [87], plays a central role in endowing these systems with a reasonable inferential power. There has been, however, a difficulty in justifying and making precise the form this rule should take. This difficulty has been a primary motivation behind the work reported in this paper. The framework we have elaborated here provides a rationale for identifying the set of assumptions to adopt in a given context.

Nonetheless, the proposed semantics does not validate the probabilistic rules of **L**; indeed, unlike **L**, the semantics is not *cumulative*.[14] In other words, even if H defeasible follows from T, the contexts T and $T \cup \{H\}$ are not guaranteed to yield the same conclusions.[15]

In this regard, two important questions remain to be answered. The first has to do with whether 'cumulativity' is a reasonable property to have in a defeasible logic, and if so, whether it is possible to embed a cumulative logic within a semantics capable of drawing sensible assumptions about conditional independence. A discussion of some of the issues involved in these questions can be found [Geffner, 89].

Acknowledgments

I would like to thank Kit Fine and Judea Pearl for insightful discussions. Special thanks to Bill Dolan for his invaluable help with the presentation.

References

[Adams, 66] E.Adams. Probability and the logic of conditionals. In *Aspects of Inductive Logic*, J. Hintikka and P. Suppes (Eds), North Holland Publishing Company, Amsterdam, 1966.

[Apt *et al.*, 88] K. Apt, H. Blair and A. Walker. Towards a theory of declarative knowledge. In *Foundations of Deductive Databases and Logic Programming*, J. Minker (Ed), Morgan Kaufmann, Los Altos, 1988, 89-148.

[Delgrande, 87] J. Delgrande. An approach to default reasoning based on a first-order conditional logic. *Proceedings AAAI-87*, Seattle, 1987, 340-345.

[Etherington, 88] D. Etherington. *Reasoning with Incomplete Information.* Pitman, London, 1988.

[Fine, 88] K. Fine. The justification of negation as failure. Dept. of Philosophy, UCLA, 1988. To appear in *Proceedings of 8th International Congress of Logic Methodology and Philosophy of Science*, North Holland, 1989.

[Geffner and Pearl, 87] H. Geffner and J. Pearl. A framework for reasoning with defaults. TR-94b, Cognitive Systems Lab., October 1987, UCLA. Also in *1988 Proceedings of the Society for Exact Philosophy*, to appear.

[Geffner, 88] H. Geffner. On the logic of defaults. *Proceedings of the AAAI-88*, St. Paul, Minnesota, 449-454.

[Geffner, 89] H. Geffner. *Explorations in non-monotonic reasoning.* Forthcoming dissertation, Computer Science Dept., UCLA.

[Gelfond and Lifschitz, 88] M. Gelfond and V. Lifschitz. The stable model semantics for logic programming. *Proceedings 1988 Symposium on Logic Programming*, MIT Press, Cambridge, Mass., 1988, 1070-1080.

[Hanks and McDermott, 86] S. Hanks and D. McDermott. Default reasoning, non-monotonic logics, and the frame problem. *Proceedings AAAI-86*, Philadelphia, 1986, 328-333.

[13] Indeed, **L** comprises a set of rules which define a sound and complete logic of high probability (see [Adams, 66] and [Pearl and Geffner, 87]). This probabilistic interpretation, however, is not essential; Kraus *et al.* [88] have developed a system with equivalent power within a preferential semantics setting.

[14] The term "cumulativity" has apparently been coined by Makinson [89]. See also Kraus *et al.* [88].

[15] Just consider a theory with defaults 'if A then B', 'if A and B then C' and 'if C then ¬B'. It is possible to verify in such a theory that both C and B follow from A. B, on the other, does not follow from A and C.

[Harman, 86] G. Harman. *Change in View.* MIT Press, Cambridge, MA, 1986

[Haugh, 87] B. Haugh. Simple causal minimizations for temporal persistence and projection. *Proceedings of the AAAI-87*, Seattle, Washington, 218-223.

[Haugh, 88] B. Haugh. Tractable theories of multiple defeasible inheritance in ordinary non-monotonic logics. *Proceedings AAAI-88*, St. Paul, Minnesota, 421-426.

[Horty *et al.*, 87] J. Horty, R. Thomason and D. Touretzky. A skeptical theory of inheritance. *Proceedings AAAI-87*, Seattle, Washington, 358-363.

[Kraus *et al.*, 88] S. Kraus, D. Lehmann and M. Magidor. Preferential models and cumulative logics. Dept. of Computer Science, Hebrew University, Jerusalem 91904, Israel, August 1988.

[Lifschitz, 85] V. Lifschitz. Computing circumscription. *Proceedings IJCAI-85*, Los Angeles, California, 121-127.

[Lifschitz, 87] V. Lifschitz. Formal theories of action. *Proceedings of the 1987 Workshop on the Frame Problem in AI*, Kansas, 1987, pp. 35-57.

[Loui, 87b] R. Loui. Defeat among arguments: a system of defeasible inference. *Computational Intelligence*, 3, 1987.

[Makinson, 89] D. Makinson. General theory of cumulative inference. *Proceedings of the Second International Workshop on Non-Monotonic Reasoning*, Springer Lecture Notes on Computer Science, January 1989.

[McCarthy, 80] J. McCarthy. Circumscription–A form of non-monotonic reasoning. *Artificial Intelligence*, 13, 1980, 27-39.

[McCarthy, 86] J. McCarhty. Applications of circumscription to formalizing commonsense knowledge. *Artificial Intelligence*, 28, 1986, 89-116.

[McDermott, 82] D. McDermott. Non-monotonic logic II: non-monotonic modal theories. *JACM*, 29, 1982, 33-57.

[Morgenstern and Stein, 88] L. Morgenstern and L. Stein. Why things go wrong: a formal theory of causal reasoning. *Proceedings AAAI-88*, St. Paul, Minnesota, 1988, 518-523.

[Pearl, 88] J. Pearl. Embracing causality in default reasoning. *Artificial Intelligence*, 35, 1988, 259-271.

[Pearl and Geffner, 88] J. Pearl and H. Geffner. Probabilistic semantics for a subset of default reasoning. TR-93-III, Cognitive Systems Lab., UCLA, March 1988. Also in J. Pearl. *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, Los Altos, 1988, Ch. 10.

[Perlis, 88] D. Perlis. Commonsense set theory. In *Meta-Level Architectures and Reflection*, P. Maes and D. Nardi (Eds), Elsevier Science Publishers, 1988.

[Poole, 87] D. Poole. Defaults and conjectures: hypothetical reasoning for explanation and prediction. Report CS-87-4, University Waterloo, October 1987.

[Przymusinski, 88] T. Przymusinski. On the declarative semantics of stratified deductive databases and logic programs. In *Foundations of Deductive Databases and Logic Programming*, J. Minker (Ed), Morgan Kaufmann, Los Altos, 1988, 193-216.

[Reiter, 80] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13, 1980, 81-132.

[Reiter, 80b] R. Reiter. Equality and domain closure in first order databases. *JACM*, 27, 1980, 235-249.

[Reiter, 87] R. Reiter. Non-monotonic reasoning. *Annual Review of Computer Science*, 2, 1987, 147-186.

[Sandewal, 88] E. Sandewal. The semantics of non-monotonic entailment defined using partial interpretations. Technical Report LiTH-IDA-R-88-31, Dept. of Computer and Information Science, Linköping University, Sweden.

[Selman and Kautz, 88] B. Selman and H. Kautz. The complexity of model-preference default theories. Proceedings CSCSI-88, Edmonton, Alberta, June 1988, 102-109.

[Shepherson, 88] J. Shepherson. Negation in logic programming. In *Foundations of Deductive Databases and Logic Programming*, J. Minker (Ed), Morgan Kaufmann, Los Altos, 1988, 19-88.

[Shoham, 88] Y. Shoham. *Reasoning about Change.* MIT Press, Cambridge, Mass., 1988.

[Touretzky *et al.*, 86] D. S. Touretzky. *The Mathematics of Inheritance Systems.* Morgan Kaufmann Publishers, Los Altos, CA, 1986.

[Van Gelder *et al.*, 88] A. Van Gelder, K. Ross and J. S. Schlipf. Unfounded sets and well-founded semantics for general logic programs. *Proceedings Seventh Symp. on Principles of Database Systems*, 1988, 221-230.