A SOUND FRAMEWORK FOR REASONING WITH DEFAULTS

Hector Geffner
Judea Pearl

December 1987
CSD-870066

A condensed version of this report was submitted to CSCSI-88 Conference, Edmonton, AL. Canada.

TECHNICAL REPORT
CSD-8700XX
R-94-II
December 1987

A SOUND FRAMEWORK FOR REASONING WITH DEFAULTS *

Hector Geffner & Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los-Angeles, CA. 90024-1596

ABSTRACT

A new system of defeasible inference is presented. The system is made up of a body of six rules which allow proofs to be constructed very much like in natural deduction systems in logic. Multiple extensions do not arise. Five of the rules are shown to possess a sound and clear probabilistic semantics that guarantees the high probability of the conclusion given the high probability of the premises. The sixth rule appeals to a notion of irrelevance; we explain both its motivation and use. †

* This work was supported in part by the National Science Foundation Grant, DCR 83-13875, IRI 86-10155.

† This is a revised version of the report [Geffner et. al. 87]. The main departure here is the elimination of the notion of 'monotonicity in context' in favor of the more primitive notion of 'potential relevance' in section 2.3.

A Sound Framework for Reasoning with Defaults

Hector Geffner
hector@cs.ucla.edu

Judea Pearl
judea@cs.ucla.edu

Cognitive Systems Lab.
Dept. of Computer Science
UCLA

December 2, 1987

Abstract

A new system of defeasible inference is presented. The system is made up of a body of six rules which allow proofs to be constructed very much like in natural deduction systems in logic. Multiple extensions do not arise. Five of the rules are shown to possess a sound and clear probabilistic semantics that guarantees the high probability of the conclusion given the high probability of the premises. The sixth rule appeals to a notion of irrelevance; we explain both its motivation and use.[1]

1 Motivation

Belief commitment and belief revision are two distinctive characteristics of common sense reasoning. Classical logic as well as probability theory have been shown to be incapable of capturing these features by themselves. The former due to its inability to revise old beliefs in the light of new information; the latter due to its lack of commitment: propositions are believed only to a certain degree which dynamically changes with new information.

In recent years there has been an effort to enhance both formalisms in order to overcome these limitations. Those working within the probabilistic framework have tried to devise 'acceptance rules' to work on top of a body of probabilistic knowledge, as to create a body of believed, though defeasible, set of propositions [see Loui 85, Pearl 86b]. Those working within the logic framework have developed 'non-monotonic' inference systems [AI Journal 80] based on classical logic, in which old theorems can be defeated by new information.

The purpose of these extensions has been to produce an inference machinery capable of generating all conclusions that 'reasonably' follow from a given body of knowledge. It is in fact in this respect that the probabilistic approach has enjoyed a significant advantage over the logicist approach. A body of probabilistic knowledge together with an acceptance rule uniquely determines the conclusions that can be derived. Both the probabilistic knowledge base and the acceptance rule can be modified so as to capture those conclusions that seem reasonable. Non monotonic logics, on the other hand, have lacked such *clear semantics*. Not only it has been difficult to tune the set of defeasible rules so as to 'entail' the desired conclusions [see Hanks and

[1] This is a revised version of the report [Geffner *et. al.* 87]. The main departure here is the elimination of the notion of 'monotonicity in context' in favor of the more primitive notion of 'potential relevance' in section 2.3.

McDermott 86], but it has even been difficult to characterize what the desired conclusions are [see Touretzky et al. 87, "A clash of *intuitions* ..."].

While well understood, the probabilistic approach seems to be both too expensive and precise for the task at hand. Too many parameters are needed to fully specify a body of probabilistic knowledge[2] and, moreover, these parameters are sometimes very difficult to assess in a consistent way. For example, while we can estimate the probability of birds flying; it is much more difficult to estimate the probability of non-birds flying. Furthermore, the expense of computing with numerical parameters does not seem necessary for coarse-grained acceptance rules.

In this paper we show that it is possible to achieve the best of both worlds by presenting a system of defeasible inference which operates very much like natural deduction systems in logic and, yet, can be justified on probabilistic grounds. The resulting system is related on one hand to the logic of conditionals developed by Adams [Adams 66], in which conditionals of the form 'if P then Q' are interpreted as asserting that the conditional probability of Q given P is close to one. On the other hand, the appeal to a notion of relevance in our formulation bears a close relationship to those approaches in which the structure of arguments supporting contradictory conclusions are examined in order to eliminate the effects of spurious extensions [Loui 86, Poole 85, Touretzky 84].

The structure of the paper is as follows. In section 2 we define the language as well as the rules of inference that make up the system. In section 3 we illustrate its applicability by going through a set of examples. We discuss related work in section 4, and summarize the main contributions in section 5.

2 A System of Defeasible Inference

2.1 Rules of Inference 1–5

The language of the system is a straightforward extension of the language of FOL. Besides the logical formulas, it comprises defeasible rules of the form $P \rightarrow Q$, where both P and Q are closed logical formulas. A *context* is a pair $\langle L, D \rangle$ of logical formulas L and defeasible rules D. A default schema of the form $p(x) \rightarrow q(x)$ in D, where $p(x)$ and $q(x)$ are logical formulas with x as their only free variable, stands for the infinite collection of defeasible rules obtained by substituting x by each the ground terms comprised in the language. We will sometimes refer to the elements of L simply as formulas, and refer to the members of D as defaults.

A *theory* $T = (K, E)$ is composed of a *background context* $K = \langle L, D \rangle$ and an evidence set E of additional facts learned. We will sometimes find useful to refer to the global context $\langle L \cup E, D \rangle$ associated with the theory $T = (K, E)$ as E_K. The system of inference implicitly defines the set of conclusions h that follow from E_K. We will denote such a relation as $E \mid\!\!\sim_K h$, and say that h follows from the evidence set E in context K, or simply that h can be derived from E in K. The initial set of rules we are going to consider is given by:

Rule 1 (Defaults) If $E \rightarrow h \in D$ then $E \mid\!\!\sim_K h$

Rule 2 (Logic theorems) If $L \cup E \vdash h$ then $E \mid\!\!\sim_K h$

Rule 3 (Triangularity) If $E \mid\!\!\sim_K h$ and $E \mid\!\!\sim_K E'$ then $E, E' \mid\!\!\sim_K h$

Rule 4 (Bayes) If $E \mid\!\!\sim_K E'$ and $E, E' \mid\!\!\sim_K h$ then $E \mid\!\!\sim_K h$

Rule 5 (Disjunction) If $E \mid\!\!\sim_K h$ and $E' \mid\!\!\sim_K h$ then $E \vee E' \mid\!\!\sim_K h$

[2] Though not as many as is usually thought. See [Pearl 86a] for a discussion of structuring probabilistic knowledge.

Rule 1 permits us to conclude the consequent of a default when its antecedent is all that has been learned. Rule 2 states that theorems that logically follow from a set of formulas can be concluded in any theory containing those formulas. Rule 3 permits the incorporation of a set of established conclusions to the current evidence set, without affecting the status of any other derived conclusions. Rule 4 says that any conclusion that follows from a context whose evidence set was augmented with conclusions established in that context, also follows from the original context alone. Finally, rule 5 says that a conclusion that follows from either of two evidential sets, also follows from their disjunction.[3]

Rules 1–5 can be shown to share the inferential power of the system of rules proposed by Adams in [Adams 66] for deriving what he calls the probabilistic consequences of a given set of conditionals. Interestingly, rules 3 and 4 also appear in [Gabbay 85] as relations among wffs to be satisfied by any non-monotonic logic.

We proceed now to investigate some of the properties of the system of defeasible inference defined by rules 1–5. Later on, we shall discuss some of its limitations as we enhance the system with a sixth rule.

2.1.1 Some Meta-Theorems

Theorem 1 (Logical Closure 1) *If* $E \mid\!\!\sim_K h$ *and* $E, h \vdash h'$ *then* $E \mid\!\!\sim_K h'$.
It follows by sequentially applying rules 2 and 4.

Theorem 2 (Logical Closure 2) *If* $E \mid\!\!\sim_K h$, $E \mid\!\!\sim_K h'$, *and* $E, h, h' \vdash h''$, *then* $E \mid\!\!\sim_K h''$.
By rule 3, we obtain $E, h \mid\!\!\sim_K h'$. From rule 2, we get $E, h, h' \mid\!\!\sim_K h''$. Applying rule 4 twice, the theorem is proved.

Theorem 3 (Equivalent Contexts) *If* $E \equiv E'$ *and* $E \mid\!\!\sim_K h$, *then* $E' \mid\!\!\sim_K h$.
Since $E \vdash E'$, by applying rules 2 and 3 we get $E, E' \mid\!\!\sim_K h$; which together with $E' \vdash E$ and rules 2 and 4, leads to $E' \mid\!\!\sim_K h$.

Theorem 4 (Exceptions) *If* $E, E' \mid\!\!\sim_K h$ *and* $E \mid\!\!\sim_K \neg h$, *then* $E \mid\!\!\sim_K \neg E'$.
From $E, E' \mid\!\!\sim_K h$, we can obtain by theorem 1, $E, E' \mid\!\!\sim_K h \vee \neg E'$. On the other hand, from rule 2 we can conclude $E, \neg E' \mid\!\!\sim_K h \vee \neg E'$. Combining these two results by means of rule 5 and theorem 3, we get $E \mid\!\!\sim_K h \vee \neg E'$ and, therefore, $E \mid\!\!\sim_K \neg E'$ by virtue of theorem 2 and $E \mid\!\!\sim_K \neg h$.

Some non-theorems:

$E \vdash E'$ and $E' \mid\!\!\sim_K h$ does not necessarily imply $E \mid\!\!\sim_K h$
$E \mid\!\!\sim_K h$ and $E' \mid\!\!\sim_K h$ does not necessarily imply $E, E' \mid\!\!\sim_K h$

Note that the first non-theorem is clearly undesirable. If accepted, it would endow our system with monotonic characteristics of classical logic, precluding exceptions like non-flying birds, etc. Let us just say that neither of them is sound or, what amounts the same, that it is possible to find counter-examples which intuitively violate those rules.

As we shall see later, the system of rules 1–5 defines an extremely conservative non-monotonic logic. In fact, the inferences sanctioned by these rules do no involve any type of assumptions regarding information absent from the background context. To illustrate this fact, let $K = \langle L, D \rangle$ and $K' = \langle L', D' \rangle$ denote two background contexts, such that $K \leq K'$, i.e. $L \subseteq L'$ and $D \subseteq D'$. We have the following theorem:

Theorem 5 (K-monotonicity) *If* $E \mid\!\!\sim_K h$ *and* $K \leq K'$ *then* $E \mid\!\!\sim_{K'} h$.

This theorem follows easily by induction on the minimal length n of the derivation of $E \mid\!\!\sim_K h$. If $n = 1$, it means that h was derived from E in K either by rule 1 or by rule 2. In either case it is easy to show that

[3] Rule 5 can be shown to be equivalent, in the context of rules 1–4, to rule: 'If $E, E' \mid\!\!\sim_K h$ then $E \mid\!\!\sim_K \neg E' \vee h$ '. The latter was used, instead of rule 5, in the formulation reported in [Geffner *et. al.* 87].

h can be derived from E in K'. Let us assume now that h is derivable from E in K in n steps, $n > 1$, and that the theorem holds for all the proofs with length $m < n$. Clearly the last step in the derivation must involve one of the rules 3–5. In any case the antecedents of such a rule must be derivable in a number of steps smaller than n and, therefore, by the inductive assumption, they are also derivable in K', from which it follows that, using the same rule, h is also derivable from E in K'.

Finally, we can show rules 1–5 to be probabilistically sound. For that purpose we define probabilistic models in the following way. A probability distribution $P_K(\cdot)$ is a probabilistic model of the background context K, iff:

$$\begin{aligned} P_K(L|E) &= 1 \quad \text{for any body of evidence } E, \text{ and} \\ P_K(a|b) &\approx 1 \quad \text{for every default } a \rightarrow b \in D. \end{aligned}$$

Theorem 6 (Probabilistic Soundness of Rules 1–5) Let the expressions $E_i \mid\!\!\sim_K h_i$, $i = 1, ..., n_j$, denote each of the antecedents of rule j above, $j = 1, ..., 5$; and let $E \mid\!\!\sim_K h$ stand for the consequent of such a rule. Then for any probabilistic model of K, $P_K(\cdot)$, such that $P_K(h_i|E_i) \approx 1$, for $i = 1, ..., n_j$; then $P_K(h|E) \approx 1$.

A proof can be found in the appendix. This theorem guarantees the high probability of conclusions derived by means of rules 1–5 from a set of highly likely premises.[4] Adams [Adams 66] additionally provides a completeness result for a system of inference with equivalent expressive power to the one defined by rules 1–5.

We now turn our attention to an example that shows how the body of rules introduced so far can account for simple patterns of non-monotonic reasoning.

2.2 Example

Example 1. Let us consider the theories $T_1 = (K, E_1)$[5] and $T_2 = (K, E_2)$, with background context $K = \langle L, D \rangle$, and $L = \{\forall x. penguin(x) \supset bird(x)\}$, $D = \{penguin(x) \rightarrow \neg flies(x), bird(x) \rightarrow flies(x)\}$, $E_1 = \{penguin(Tim)\}$, and $E_2 = \{penguin(Tim), bird(Tim)\}$.

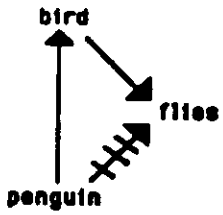

Figure 1: The penguin triangle

Concluding that 'Tim does not fly' in context K knowing that 'Tim is a penguin' amounts to proving $penguin(Tim) \mid\!\!\sim_K \neg flies(Tim)$. The proof gets reduced to a single application of rule 1, since $penguin(Tim) \rightarrow \neg flies(Tim) \in D$.

[4] J. Pearl [Pearl 87] has also recently advocated the use of probability theory to fill the 'semantic gap' that have characterized algorithms dealing with inheritance hierarchies with exceptions. He proposes an ϵ-semantics, which implicitly defines, in terms of probability theory, the set of conclusions which ought to follow from a given default hierarchy. While we also appeal to probability theory to define the semantics of the system proposed, its soundness will follow directly directly from the soundness of its rules of inference.

[5] As it is usual, we display the relationships embedded in a given theory by means of graphs. Positive defaults $P \rightarrow Q$ and implications $P \supset Q$ are depicted by positive links ($\rightarrow$). Negative defaults $P \rightarrow \neg Q$ and implications $P \supset \neg Q$ on the other hand, are represented by negative links ($\not\rightarrow$).

Proving $E_2 \mid\!\!\sim_R \neg flies(Tim)$ is slightly different since a new fact, $bird(Tim)$, needs to be assimilated. The proof goes as follows:[6]

1.	$penguin(Tim) \mid\!\!\sim_R \neg flies(Tim)$	rule 1
2.	$penguin(Tim) \mid\!\!\sim_R bird(Tim)$	rule 2
3.	$penguin(Tim), bird(Tim) \mid\!\!\sim_R \neg flies(Tim)$	rule 3; 1, 2.

Note that the new piece of information available in T_2, $bird(Tim)$, does not alter the consequences that followed from the older theory T_1 since, as reflected by rule 3, the new information learned was itself one of the consequences of T_1. It is interesting to note that the system proposed here, in contrast with other systems of defeasible reasoning reported in the literature, has different proofs for the proposition $\neg flies(Tim)$ in theories $T_1 = (K, E_1)$ and $T_2 = (K, E_2)$. In fact, in the first theory, the resulting proof qualifies as a single shot proof : it was not even necessary to consider the impact which the consequences of being a penguin (its birdness) could have on its (in)ability to fly.

To better illustrate this difference, let us consider a new theory $T_1' = (K', E_1')$, with $K = (L', D')$, $L' = \{\}$, $D' = \{penguin(x) \rightarrow \neg flies(x), bird(x) \rightarrow flies(x)\}$, and $E_1' = \{\forall x.\, penguin(x) \supset bird(x), penguin(Tim)\}$; in other words, T_1' is identical to T_1 except for the fact that the class inclusion $penguin(x) \supset bird(x)$ is now treated as a learned fact, rather than as part of the background context. We find that, although both theories share the same set of defaults D and the same set of logical formulas, $L' \cup E_1' = L \cup E_1$, the conclusion $\neg flies(Tim)$, shown to be derivable in theory T_1, is not derivable from T_1', i.e., $E_1' \not\mid\!\!\sim_R \neg flies(Tim)$. The reason for this unusual, but desirable behavior, is that the system now takes the relation 'penguins are birds' as a new piece of knowledge, independent of the background knowledge used to assume that most penguins do not fly, and which happens to support the opposite conclusion. [7]

In our framework, the preference for the conclusion that penguins do not fly in spite of beings birds, is not to be explained in terms of class specificity alone, but in terms of the knowledge that went into defining the default rules. If the system cannot ensure that the default stating that 'most penguins do not fly' already took into account the facts that penguins are birds, and that birds usually do fly, it cannot guarantee, upon learning the former, that it should not revise its conclusion.

This shows that formulas cannot be freely moved between the background context and the evidence set without altering the meaning of the theory they define. Propositions in a background context K represent knowledge *shared* by all the defaults in K. Unlike formulas in the evidence set, they do not represent pieces of evidence that need to be assimilated in order to reach a conclusion. That is the proof theoretic significance of rule 1.

2.3 Relevance

The common interpretation of defaults of the type $a \rightarrow b$ is in the form of a disposition to believe b when a is believed and no reason for not doing so is apparent. This reading has two implications we shall be concerned with: one which requires conclusions to be retractable in the light of new refuting evidence; the second which requires conclusions to persist in the light of new but irrelevant evidence. Rules 1–5 excel at the first requirement: their soundness prevents preserving a conclusion in a context in which its high probability cannot be guaranteed. In example 1 we have shown, for instance, that while birds can be assumed to fly, birds known to be penguins cannot. On the other hand, it is easy to discover that the same body of rules fail miserably in the second aspect. To illustrate these limitations, let us consider the background context $K = (L, D)$ with $L = \{\}$ and $D = \{a \rightarrow b\}$. Rule 2 allows then to conclude $a \mid\!\!\sim_R b$. However, if a new piece of information e, that bears no relation to b is discovered, rules 1–5 fail to prove $a, e \mid\!\!\sim_R b$ and, therefore, to maintain the belief in b in the new context.

[6] Proofs appear as a sequence of lines. Each formula in a proof has associated both a number and a justification. The latter indicates the rule used in deriving the formula, as well as the conditions that make the rule applicable.

[7] If this behavior does not seem convincing, consider for instance the case in which you have been assuming your neighbors to be respectable people and you suddenly come to know that they were found suspect of drug dealing. Surely learning the latter might lead some people to doubt, at least, about the previous assumption.

This 'conservatism' arises as no surprise from a set of rules insisting on probabilistic soundness: while there is no reason to believe that the presence of e in the context $\{a\}_K$ could render b less likely, such a situation would be perfectly consistent and, since a sound conclusion must hold in every probabilistic model of K, $a, e \not\vdash_K b$ is not sound and, therefore, not provable from rules 1–5.[8] Furthermore, closer inspection of the rules 1–5 reveals that the only type of evidence that can be assimilated without affecting the status of a derived conclusion is evidence which is subsumed by older information (like in example 1, in which we 'learn' that Tim is a bird after knowing he is a penguin).

It is clear that if we want the system to exhibit reasonable inferences, like the one illustrated by the example above, we need to restrict the family of probabilistic models relative to which a given conclusion must be checked for soundness. We want those models to embed the common sense assumption that no conclusion should be retracted when there is no explicit reason for doing so. With that purpose in mind, we now focus on the characterization of those situations in which a default $a \rightarrow b$ justifies concluding b from a, even in the presence of an additional piece of evidence e. We will refer to those situations in which e precludes such a conclusion as saying that e defeats $a \rightarrow b$.[9]

The problem of specifying the conditions under which a given conclusion can be preserved upon acquiring new information does not appear as such in other formalisms of default reasoning. Reiter's default logic [Reiter 80], for instance, lacks an independent notion of provability. Provability in default logic is defined indirectly, by appealing to (logical) provability in the extensions of the given theory. A proposition is provable in a theory only if it holds in every extension of it. So, if a proposition b was believed in the theory and a new piece of information e, irrelevant to b, is learned, we would expect that no extensions will be generated in which b does not hold. Since there is no notion of proof for b in the theory, but only proofs for b in each one of its extensions, there is no no need to assess the effect of the new piece of evidence e upon the status of b in the whole theory; 'all' that is needed is to recompute the status of b in each one of the new set of extensions.

The characterization of the conditions under which the acquisition of new information does not affect the status of a given derived proposition is closely related to the problem of characterizing the conditions for argument defeat investigated by Loui. Loui [Loui 86] provides a set of rules which specify the conditions under which an argument supporting a given conclusion is defeated by a counter-argument supporting its negation. Rather than appealing directly to a notion of provability, these rules provide the means for comparing pair of arguments solely in terms of their structure.

Arguments in his system represent reasons for belief in a proposition given a set of logical formulas and defaults[10]. Informally, in our terms, the background context $K = \langle L, D \rangle$ and the evidence set E provide an argument in support of formula h, in the case where either h logically follows from $L \cup E$, or there is an argument supporting a formula h' and a default $h' \rightarrow h$ in D. For instance, we might take figure 2 as representing the context $\{r\}_K$, with $K = \langle L, D \rangle$, $L = \{\}$ and $D = \{r \rightarrow f, a \rightarrow f, r \rightarrow \neg g, f \rightarrow g\}$. In such a context, there is an argument supporting $\neg g$, referred as $\mathcal{A}_K(\neg g, \{r\})$, which corresponds to the path $r \not\rightarrow g$ in the figure. There is also a counter-argument, $\mathcal{A}_K(g, \{r\})$, which corresponds to the path $r \rightarrow f \rightarrow g$.

An argument $\mathcal{A}_K(h, E)$ can be regarded as a directed acyclic graph, with h as its only sink, tautologies and formulas in $L \cup E$ as its only sources, and every formula, except the sources, properly justified by its parents. Formulas are justified in three different ways:

- a set of formulas J_f justifies a formula f, if J_f, but no proper set of it, logically entails f,
- a formula f justifies g, if there is a default $f \rightarrow g$ in D,
- a formula $f \vee h$ justifies $g \vee h$, if there is a default $f \rightarrow g$ in D.

[8] Another way of looking at this example is by considering the background context $K' = \langle L', D' \rangle > K = \langle L, D \rangle$, with $L' = \{\}$ and $D' = \{a \rightarrow b, a \wedge e \rightarrow \neg b\}$. Clearly K' does not permit the conclusion b from a and e. However, if K sanctioned such a conclusion, so should K', in light of the K-monotonicity of the rules (Theorem 5).

[9] The reason for choosing defaults as the objects of defeat, rather than the consequents of those defaults, will be explained below.

[10] Very much like Poole's theories [Poole 85] in default logic and Touretzky's paths in inheritance hierarchies [Touretzky 84]

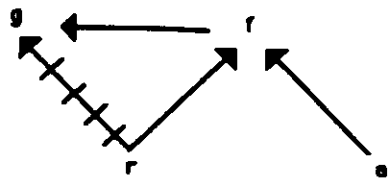

Figure 2: Paths as Arguments

Additionally, we require arguments to be both consistent and non redundant. An argument $\mathcal{A}_K(h, E)$ is consistent if the set which comprises the formulas in $\mathcal{A}_K(h, E)$ together with the formulas in $E \cup L$ is consistent whenever the latter set is. For an argument to be redundant, let f denote a formula in $\mathcal{A}_K(h, E)$, with parents $J_f = \{f^1, ..., f^i, ..., f^n\}$ and ancestors J_f^*, and let f^i denote a non-source formula. Then we say that $\mathcal{A}_K(h, E)$ is redundant if there is a formula f in $\mathcal{A}_K(h, E)$ which is logically entailed by the formulas in the set $E \cup L \cup J_f^* - \{f^i\}$. In such a case we also say that the justification of f in such an argument is redundant.

For instance, the path $a \rightarrow f \rightarrow g$ in fig. 2, does not qualify as an argument for g in a context having $E = \{a, f\}$, since as f belongs to $L \cup E$ but does not appear as source, it makes the justification of g redundant. Note however that in the same context, the path $f \rightarrow g$ does qualify as an argument for g.

The *support* of an argument refers to the set of formulas in $L \cup E$ which actually take part in it, i.e., to the formulas which appear as sources in the argument graph. For instance, both the arguments $\mathcal{A}_K(g, \{r\})$ and $\mathcal{A}_K(\neg g, \{r\})$ above, have support $\{r\}$.

Rather than defining defeat by comparison of pairs of arguments as Loui does[11], we appeal to the notion of *irrelevance*. The defeasibility criterion we shall propose asserts essentially, that a formula e would defeat a default $a \rightarrow b$ only if it is relevant to $\neg b$. As a first step in formalizing such an idea, we introduce the following definition.

We say that a formula e is *potentially relevant* to formula h in the context E_K, where $K = (L, D)$, if and only if e does not logically follow from $L \cup E$ and there is an argument $\mathcal{A}_K(h, L \cup \{e\})$, with support S, such that $e \in S$. When e represents a formula not potentially relevant (p.r.) to h in E_K we will write $I_K(h, e; E)$. Sometimes we will use the notation $I_K(h, e)$ as an abbreviation for $I_K(h, e; \emptyset)$.

For instance, from the argument that corresponds to the path $a \rightarrow f \rightarrow g$, it follows that a is p.r. to g in the context $\{r\}_K$. This is no longer true however in the enhanced context $\{r, f\}_K$, because the presence of f in the context renders a no longer p.r. to g. We refer to those situations as saying that f *blocks* a from g.

Note that the definition of potential relevance can be easily extended to deal with set of formulas, i.e. if we let $\phi(E')$ denote the formula obtained by conjoining the formulas in E', we say that E' is potentially relevant to h in context E_K, if and only if $\phi(E')$ is. As an illustration we might consider fig. 2 again. Clearly both formulas f and $a \wedge f$ are potentially relevant to g in context $\{r\}_K$. Notice however that $\neg a \wedge f$ is not.

As argued above, defaults of the form $a \rightarrow b$ are usually understood to state that 'a is a reason for believing b, as long as there is no reason for believing $\neg b$'. In our terms, that amounts to say that if there is a default $a \rightarrow b$ in D, and a is all that has been learned, the belief in the conclusion b should persist upon acquiring a new piece of evidence e, as long as e is not potentially relevant to $\neg b$. Expressed as a new rule, we have that, given a theory $T = (K, E)$, with $K = (L, D)$:

Rule 6' (Explicit Irrelevance) If $a \rightarrow b \in D$ and $I_K(\neg b, e; \{a\})$, then $a, e \not\vdash_R b$.

Rule 6' expresses a condition under which a new piece of evidence e can be safely assumed to be *irrelevant* to a given proposition in a given context. Together with rules 1–5 it indeed succeeds in producing the desired

[11] That would turn out to be too involved for our purposes and sometimes would render results slightly different than those that follow from our definition.

conclusion in the example discussed at the beginning of this subsection.

We are now in a position to provide a justification for choosing defaults, rather than the consequents of defaults, as the objects of defeat. Let us consider the context $K = (L, D)$, with $L = \{\}$ and $D = \{ d_1 : a \rightarrow b,\ d_2 : a \wedge c \rightarrow \neg b,\ d_3 : a \wedge c \wedge d \rightarrow b \}$. Clearly, from d_1 we can prove $a \mid\!\!\sim_K b$. Furthermore, if we consider the new piece of evidence $e = c \wedge d$, we can still show $a, e \mid\!\!\sim_K b$. Notice however that latter conclusion does not follow from the presence of d_1 in D, but from the presence of d_3; the evidence e defeats d_1, though not its consequent b. On the other hand, had d_1 not been defeated by e, rule 6' guarantees that e would not have defeated its consequent either. Defeat of defaults appears in our framework as a finer grained notion than defeat of formulas.

The question that we shall address now, is whether there are other formulas which can be reasonably assumed to be irrelevant to a given proposition, even when they are potentially relevant to it. In the penguin example, for instance, the fact that Tim is a circus bird would be p.r. to flying, via the argument that corresponds to the path $Tim \rightarrow circus\text{-}bird \rightarrow bird \rightarrow fly$. Thus, if we know that Tim is a penguin, further discovering that he is also a circus-bird would lead us to retract the conclusion that Tim does not fly. Note that such retraction should be prevented because, as we argued earlier, the default 'penguins typically do not fly' placed in a background context together with 'penguins are birds', and 'birds typically do fly', already presumes that Tim, the penguin, is also a bird. This suggests that for a new piece of evidence e to cast doubt upon Tim's flying handicap, e must support Tim's flying *on grounds different* than birdness. Learning that Tim is a circus bird should not lead to such a retraction, unless the background context contained information suggesting that circus birds have exceptional flying abilities[12].

These considerations show that some propositions should not affect the status of the default consequent, even if they can appear to be potentially relevant to its negation. The identity of these propositions can be uncovered by means of rule 3 (triangularity). As a special case, the triangularity rule states that, if $a \rightarrow b$ is a default in K, then any piece of evidence e', which can be explained in terms of a, i.e. $a \mid\!\!\sim_K e'$, will not defeat $a \rightarrow b$. This suggests that any other piece of evidence e, which is potentially relevant to $\neg b$ only on the grounds of e', should not defeat $a \rightarrow b$ either. Expressed as a new rule, we obtain:

Rule 6 (Implicit Irrelevance)
For any default $a \rightarrow b$ in D, if there exists a formula s such that
$a \mid\!\!\sim_K s$, $a, e \mid\!\!\sim_K s$ and $I_K(\neg b, e; \{a, s\})$, then $a, e \mid\!\!\sim_K b$.

Clearly, rule 6' is a special case of rule 6 in which s appears restricted to $s = true$. That is the reason we refer to rule 6' as capturing 'explicit' irrelevance relationships, while to rule 6 as capturing irrelevance relationships only implicit in the structure of the background context.

Note that for any non-tautological formula s, rule 6 imposes the requirement that $a, e \mid\!\!\sim_K s$ must hold. This is necessary since, while we have presented the reasons for preserving the conclusion b upon learning e in the context $\{a, s\}_K$, s is not known with certainty and, therefore, we have to make sure that s is not defeated by e. The antecedent of rule 6 precludes such a possibility.

3 Examples

In this section we shall illustrate the inferential power of the system of default inference proposed by analyzing several examples. To simplify notation, we will associate with each default schema of the form $p(x) \rightarrow q(x)$ a name d_i. When no confusion arises, we will use that name, d_i, to refer to the particular defeasible

[12] That would be the case for instance, if an additional default stating that 'typically circus birds fly' would be added to a context which already contains the default that 'typically birds fly'. Rather than redundant information, the first default would be taken to imply that for some bird instances, coming to know that they are also circus birds can make a difference at the time we want to predict their flying abilities. See [Pearl 87] for a brief discussion of the probabilistic semantics associated with these assumptions.

rule of interest, e.g. $p(a) \rightarrow q(a)$. Moreover, for such a rule we will sometimes abbreviate the predicate $I_K(\neg q(a), r(a); \{p(a)\})$ by the simpler expression $I_K(\mathtt{d}_i, r)$, in which the ground term a is left implicit. With the same purpose, we will often appeal to the following proposition which trivially follows from rules 1 and 3:

Proposition 1. If $a \rightarrow b$ and $a \rightarrow c$ then $a, b \mid\!\!\sim_K c$.

Example 2. Let us consider the theory $T = (K, E)$, with $K = (L, D)$, $L = \{\}$ and

$$\begin{aligned} D \;=\; & \{\mathtt{d}_1 : u\_student(x) \rightarrow adult(x),\ \mathtt{d}_2 : adult(x) \rightarrow work(x), \\ & \mathtt{d}_3 : u\_student(x) \rightarrow \neg work(x),\ \mathtt{d}_4 : adult(x) \wedge under\_22(x) \rightarrow u\_student(x)\}. \end{aligned}$$

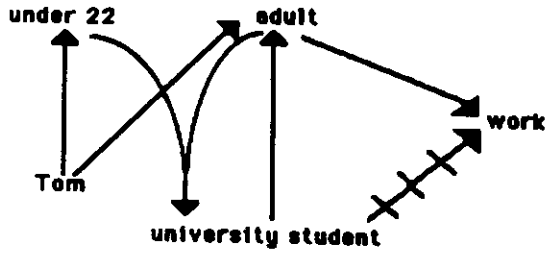

Figure 3: Adults under 22 usually do not work

We want to show that if all that we know is that Tom is an adult under 22 years old, then, with high likelihood, Tom does not work. The proof proceeds as follows:[13]

1.	$adult(Tom), under\_22(Tom) \mid\!\!\sim_K u\_student(Tom)$	rule 1; $\mathtt{d}_4$
2.	$u\_student(Tom), under\_22(Tom) \mid\!\!\sim_K \neg work(Tom)$	rule 6'; $\mathtt{d}_3$, $I_K(\mathtt{d}_3, under\_22)$
3.	$u\_student(Tom), under\_22(Tom) \mid\!\!\sim_K adult(Tom)$	rule 6'; $\mathtt{d}_1$, $I_K(\mathtt{d}_1, under\_22)$
4.	$u\_student(Tom), under\_22(Tom), adult(Tom) \mid\!\!\sim_K \neg work(Tom)$	rule 3; 3, 2
5.	$adult(Tom), under\_22(Tom) \mid\!\!\sim_K \neg work(Tom)$	rule 4; 4, 1.

It is interesting to note that from the same background knowledge, we can also derive that most adults are not university students. For that purpose let a stand for an arbitrary constant, then we obtain:

1.	$u\_student(a), adult(a) \mid\!\!\sim_K \neg work(a)$	prop 1; $\mathtt{d}_1$, $\mathtt{d}_3$
2.	$adult(a) \mid\!\!\sim_K work(a)$	rule 1; $\mathtt{d}_2$
3.	$adult(a) \mid\!\!\sim_K \neg u\_student(a)$	theorem 4; 3, 4.

Example 3.[Sandewal 86, Touretzky *et. al.* 87]. Let $T = (K, E)$, $K = (L, D)$ and :

$$\begin{aligned} L \;&=\; \{\forall x.\, royal\_elephant(x) \supset elephant(x),\ \forall x.\, african\_elephant(x) \supset elephant(x)\}, \\ D \;&=\; \{\mathtt{d}_1 : elephant(x) \rightarrow gray(x),\ \mathtt{d}_2 : royal\_elephant(x) \rightarrow \neg gray(x)\}, \\ E \;&=\; \{royal\_elephant(clyde), african\_elephant(clyde)\}. \end{aligned}$$

The proof for $\neg gray(clyde)$ in E_K proceeds as follows:

1.	$royal\_elephant(clyde) \mid\!\!\sim_K elephant(clyde)$	rule 2
2.	$royal\_elephant(clyde), african\_elephant(clyde) \mid\!\!\sim_K elephant(clyde)$	rule 2
3.	$royal\_elephant(clyde), african\_elephant(clyde) \mid\!\!\sim_K \neg gray(clyde)$	rule 6; $\mathtt{d}_2$, 1, 2, $I_K(\cdot)$

[13] We implicitly use the results of theorems 1–3 within proofs to freely change the order of conjuncts both to the left and to to right of the provability symbol.

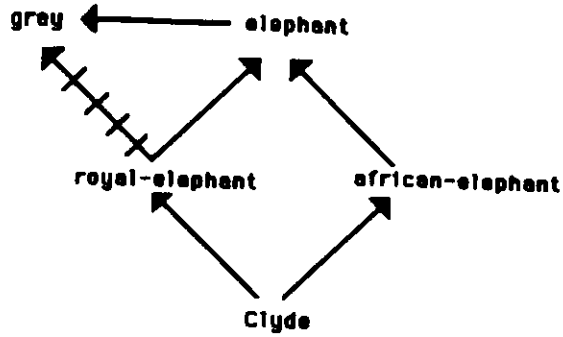

Figure 4: Clyde is not gray

The last step uses the fact $I_K(\neg gray, african; \{royal, elephant\})$, which can be understood as carrying the implicit assumption that the default 'most royal elephant are gray', also holds among african elephants. We assume that if this were not the case, the default set D in the background context would be modified accordingly, either by explicitely asserting that most african elephant are gray, or by qualifying the default that states the most royal elephants are not gray[14]. In either case, the conclusion we have derived in this context would be blocked.

Example 4. [Touretzky *et. al.* 87]. Let us consider now the theory $T = (K, E)$, with $K = \langle L, D \rangle$, $L = \{\}$ and $D = \{ \mathtt{d}_1 : A \rightarrow B, \mathtt{d}_2 : A \rightarrow \neg G, \mathtt{d}_3 : B \rightarrow G, \mathtt{d}_4 : B \rightarrow C, \mathtt{d}_5 : C \rightarrow F, \mathtt{d}_6 : G \rightarrow \neg F \}$.

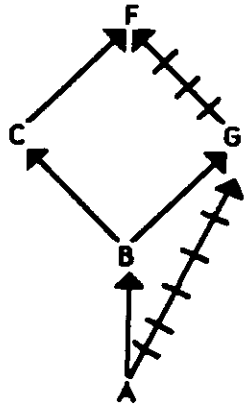

Figure 5: A's are F's

The goal is to derive proposition F from A. The intuition is to show that both C and $\neg G$ follow from A, and that the latter blocks A from $\neg F$. The proof proceeds as follows:

1.	$C, \neg G, A \mid\!\!\sim_K F$	rule 6'; $\mathtt{d}_5$, $I_K(\mathtt{d}_5, \neg G, A)$
2.	$A \mid\!\!\sim_K \neg G$	rule 1; $\mathtt{d}_2$
3.	$A \mid\!\!\sim_K B$	rule 1; $\mathtt{d}_1$
4.	$B, A \mid\!\!\sim_K C$	rule 6'; $\mathtt{d}_4$, $I_K(\mathtt{d}_4, A)$
5.	$A \mid\!\!\sim_K C$	rule 4; 4, 3
6.	$A \mid\!\!\sim_K C \wedge \neg G$	theorem 2; 2, 5
7.	$A \mid\!\!\sim_K F$	rule 4; 1, 5.

[14] We have not shown however how to accommodate defaults with exceptions in our framework. The modifications for such an enhancement are minor, and require only to modify rule 6, as to allow a new piece of evidence e to defeat a default $a \rightarrow b$ with exception c, not only when e is relevant to $\neg b$ but also when it is relevant to c.

Note that $I_K(\mathsf{d}_5, \neg G, A)$ holds due to the fact that $\neg G$ renders the path $A \rightarrow B \rightarrow G \not\rightarrow F$ inconsistent, preventing it to qualify as an argument.

Example 5. Let us consider now the theory $T = (K, E)$, $K = \langle L, D \rangle$, with

$$\begin{aligned} L &= \{\forall x.miserable(x) \equiv \neg happy(x)\} \\ D &= \{works\_at(x, university) \rightarrow happy(x), works\_at(x, office) \rightarrow happy(x)\}, \\ &\quad works\_at(x, office) \wedge works\_at(x, university) \rightarrow miserable(x)\} \\ E &= \{works\_at(John, university), works\_at(John, office)\}, \end{aligned}$$

i.e. working either at the university or at the office makes everybody happy. However, working simultaneously at both, creates a conflict that makes everybody unhappy. Rule 1 together with theorem 1 leads to $E \mid\!\!\sim_K \neg happy(John)$. If E were reduced to either $works\_at(John, university)$ or $work\_at(John, office)$, or even the disjunction of both, the opposite conclusion would be obtained. No inconsistencies appear.

Example 6. The Nixon diamond is encoded by the theory $T = (K, E)$, with $K = \langle L, D \rangle$, $L = \{\}$, and $D = \{quaker(x) \rightarrow pacifist(x), republican(x) \rightarrow \neg pacifist(x)\}$.

In this theory, no conclusion regarding Nixon's pacifism can be drawn knowing that Nixon is both a quaker and a republican. In our opinion, drawing no conclusion is, in this case, preferred to drawing two conflicting extensions, as in normal default theories. It clearly indicates that the knowledge embedded in K is not sufficient to integrate the available pieces of evidence to arrive at a conclusion. Enhancing the background context to include another default, like quakers who also are republicans are still pacifists, would solve the ambiguity without introducing any inconsistencies.

Example 7: (M. Ginsberg) Let us consider the $T = (K, E)$, $K = \langle L, D \rangle$,

$$\begin{aligned} L &= \{\} \\ D &= \{\mathsf{d}_1 : quaker(x) \rightarrow dove(x), \mathsf{d}_2 : republican(x) \rightarrow hawk(x), \mathsf{d}_3 : dove(x) \rightarrow \neg hawk(x), \\ &\quad \mathsf{d}_4 : hawk(x) \rightarrow \neg dove(x), \mathsf{d}_5 : dove(x) \rightarrow p\_motivated(x), \mathsf{d}_6 : hawk(x) \rightarrow p\_motivated(x)\} \\ E &= \{quaker(Nixon), republican(Nixon)\}. \end{aligned}$$

The conclusion that Nixon is politically motivated would follow if we could derive that he is either a hawk or a dove. However the latter does not follow from rules 1–6, since D does not provide sufficient reasons for believing either that quakers who are also republicans are still likely to be doves, or that republicans who are also quakers are still likely to be hawks.[15] [16]

Example 8. Let us finally consider the theory $T = (K, E)$, $K = \langle L, D \rangle$, $L = \{\}$, $D = \{a \rightarrow c, a \rightarrow b, a \wedge c \rightarrow \neg b\}$ and $E = \{a\}$. The theory turns out to be inconsistent: both b and $\neg b$ can be concluded, and then by theorem 1, any other proposition. Note that most default logics will not regard this knowledge base as inconsistent. Yet, a theory comprising the sets $L' = \{\}$, $D' = \{a \rightarrow b\}$ and $E' = \{a, \neg b\}$ would be perfectly consistent.

[15] If this lack of commitment seems counter-intuitive it is because the information contained in the fact that 'typically republicans are politically motivated' (independently of whether they are hawks or doves) has not been codified in the background context. In fact, if we replace 'politically motivated' by 'having an extreme position in defense issues', not drawing a conclusion seems to be the most reasonable choice.

[16] The system reported earlier in [Geffner *et. al.* 87, '*Sound* ... Inference'], mistakenly permitted to conclude that Nixon is politically motivated. This was due a consequence of a wrong definition of the monotonicity-in-context predicate M, which is no longer needed in the present formulation.

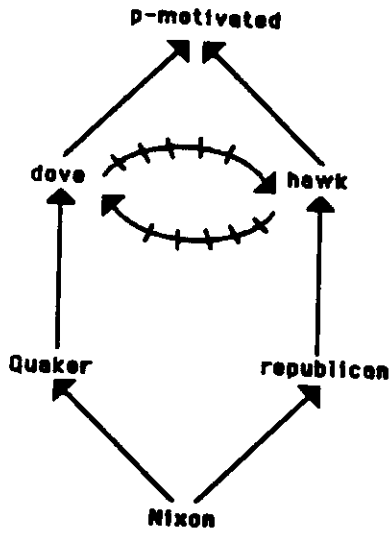

Figure 6: Is Nixon politically motivated ?

4 Related Work

As noted in [Reiter *et. al.* 81], the logic for default reasoning proposed in [Reiter 80] requires exceptions to be explicitly stated in order to prevent the multiplicity of spurious extensions. Recently, several novel systems of defeasible inference have been proposed, motivated by the intuition that it should be possible to filter the effect of spurious extensions, without the need to make exceptions explicit. Among them, the system closest in spirit to the scheme proposed in this paper is the system of defeasible inference proposed by Loui.

Loui's system [Loui 86] is made up of a set of rules to evaluate arguments. He defines a set of (syntactic) argument attributes (like 'has more evidence', 'is more specific', etc.), and a set of rules, which allow the comparison, evaluation, and selection of arguments. This set of rules seems to implicitly embed most of the inference rules that define our system, and can be mostly justified in terms of them. Still, it is possible to find some differences. One such difference is that Loui's system is not (logically) closed. It is possible to believe propositions A and B, and still fail to believe $A \wedge B$ [Loui 86]. In our scheme, the closure of the propositions believed follows from theorems 1 and 2. In particular, if the arguments for A and B in a given theory are completely symmetric, and $A \wedge B$ does not follow for some reason (like conflicting evidence), then neither A nor B will follow.

Another difference arises due to the absolute preference given by his system to arguments based on 'more evidence'. As the following example shows, this criterion might lead to counter-intuitive results. Consider the context $K = (L, D)$ with $L = \{\}$ and $D = \{A \rightarrow B, C \rightarrow \neg B, A \wedge F \rightarrow C\}$; Loui's system would conclude $\neg B$, given the evidence $E = \{A, F\}$, merely because the evidence supporting the argument $A \rightarrow B$, constitutes a proper subset of the evidence supporting the competing argument $A \wedge F \rightarrow C \rightarrow \neg B$. Yet, if proposition C, whose truth was presumed in the argument supporting $\neg B$, were learned, Loui's system would retract its belief in $\neg B$, since C renders both F and A irrelevant to $\neg B$ and, therefore, neither the argument which supports B, nor the argument that supports $\neg B$, could be said to be based on 'more evidence' than the other. Our system, as expected, will draw no conclusion in either case, since the joint influence of both A and C on B (or $\neg B$) cannot be derived from the given context.

The system reported by Touretzky in [Touretzky 84] was motivated by the goal of providing a semantics for inheritance hierarchies with exceptions. He argues that there is a natural ordering of defaults in inheritance hierarchies that can be used to filter spurious extensions. In this way, his system succeeds in capturing

inferences that seem to be reasonable, but which escape unaided, fixed-point semantic systems like Reiter's. Still, Toureztky's system can be regarded more as a refinement of Reiter's logic than as a departure from it (see [Etherington 87]). As such, it still requires to test, outside the 'logic', whether a given proposition holds in every (remaining) extension. Moreover, requirements of acyclicity are at the heart of the definition of the inferential distance principle, restricting therefore its range of applicability. It is interesting to note that rules 3 and 6 seem to convey ideas very similar to Touretzky's inferential distance. Still, while the inferential distance principle is used to discard 'inadmissible' arguments, the rules presented in section 2 are used to to prevent them from ever evolving to a ratified conclusion.

In [Poole 85], Poole has proposed another mechanism for dealing with the problem of multiple, spurious, answers that arises in Reiter's default logic. This mechanism consists of comparing the 'specificity' of the knowledge embedded in the arguments supporting contradictory conclusions. An argument shown to be strictly 'more general' than another argument, can be discarded. This criterion seems in fact very close to Touretzky's inferential distance. Still, they seem to differ in an important aspect. Unlike Touretzky, Poole compares the specificity of the arguments *isolated* from the rest of the knowledge base. It seems that this might lead to undesirable results. For instance, in example 2 (fig. 3), none of the arguments supporting the conclusion that Tom works, or that Tom does not work, can be determined to be more specific if the default that states that most students are adults— which does not take part in the competing arguments— is ignored. Like Reiter's and Toureztky's, Poole's system seems to also require to test, outside the 'logic', whether a proposition holds in every (remaining) extension in order for the proposition to be accepted.

5 Summary

The main contribution of the proposed framework for defeasible inference is the emergence of a more precise, proof theoretic and semantic account of defaults. A default $P \rightarrow Q$, in a background context K, represents a clear cut constraint on states of affairs, stating that if P is *all* that has been learned, then Q can be concluded. We appealed to probability theory to uncover the logic that governs this type of 'context dependent' implications when other facts besides P are learned. We have then shown that the inferences permitted by our system are authorized in light of the probabilistic interpretation.

Additionally we have introduced a notion of irrelevance as a set of sufficient conditions under which belief in the consequent of a given default can be preserved upon acquiring new information. This notion is used very much like a frame axiom: beliefs are assumed to persist upon acquiring a new piece of evidence e, unless e provides a 'reason' for not doing so.

The scheme proposed here avoids the problem of multiple, spurious extensions that normally arises in default logics. Moreover, we do not need to explicitly consider all the extensions in order to prove that a given proposition follows from a given theory. Proofs in our system proceed 'inside the logic', and look very much like proofs constructed in natural deduction systems in logic.

The system is also clean: the only appeal to 'provability' in the inferential machinery is to determine when a proposition can be safely assumed to be irrelevant to another proposition in a given context. But, in contrast to most non-monotonic logics, the definition of non-monotonic provability is not circular. The irrelevance predicate used for constructing proofs can be inferred syntactically in terms of arguments only.

Acknowledgment

We wish to thank E. Adams, F. Bacchus, M. Ginsberg and P. Hayes for comments on an earlier version of this paper. We also wish to thank Michelle Pearl for drawing the figures.

References

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

[AI Journal 80] Special Issue on Non-Monotonic Logics, *AI Journal*, No 13, 1980.

[Cox 46] Cox R., Probability, Frequency and Reasonable Expectation, *American Journal of Physics 14*, 1, pp 1-13.

[Etherington *et al.* 1983] Etherington D.W., and Reiter R., 'On Inheritance Hierarchies with Exceptions', *Proceedings of the AAAI-83*, 1983, pp 104-108.

[Etherington 87] Etherington D.W., 'More on Inheritance Hierarchies with Exceptions. Default Theories and Inferential Distance', *Proceedings of the AAAI-87*, 1987, Seattle, Washington, pp 352-357.

[Gabbay 85] Gabbay D.M.,'Theoretical Foundations for Non-Monotonic Reasoning in Expert Systems', in *Logics and Models of Concurrent Systems*, Edited by K. R. Apt, Springer-Berlag, Heilderberg, 1985.

[Geffner *et. al.* 87] Geffner H. and Pearl J., 'Sound Defeasible Inference', *TR-94*, August 1987, Cognitive Systems Lab., UCLA.

[Hanks *et. al.* 86] Hanks S. and McDermott D., 'Default Reasoning, Non-Monotonic Logics, and the Frame Problem', *Proceedings of the AAAI-86*, Philadelphia, PA, 1986, pp 328-333.

[Loui 85] Loui R.P., 'Real Rules of Inference', unpublished draft, 1985.

[Loui 86] Loui R.P.,'Defeat Among Arguments: A System of Defeasible Inference', Dept. of Computer Science, TR-190, Dec. 1986, University of Rochester.

[Pearl 86a] Pearl J., 'Fusion, Propagation, and Structuring in Belief Networks', *AI Journal*, Vol. 29, No 3., 1986, pp 241-288.

[Pearl 86b] Pearl J., 'Distributed Revision of Composite Beliefs', *AI Journal*, 33, No 2, Oct. 87.

[Pearl 87] Pearl J., 'Probabilistic Semantics for Inheritance Hierarchies with Exceptions', *TR-93*, July 1987, Cognitive Systems Lab., UCLA.

[Poole 85] Poole D. 'On the Comparison of Theories: Preferring the Most Specific Explanation', *Proceedings of the IJCAI-85*, Los Angeles, 1985.

[Reiter 80] Reiter. R., 'A Logic for Default Reasoning' *AI Journal*, No 13, 1980, pp 81-132.

[Reiter *et. al.* 81] Reiter R. and Criscuolo G., 'On Interacting Defaults', *Proceedings of the IJCAI-81*, pp 270-276.

[Sandewal 86] Sandewal E., 'Non-monotonic Inference Rules for Multiple Inheritance with Exceptions', *Proceedings of the IEEE*, vol. 74, 1986, pp 1345-1353.

[Touretzky 84] Touretzky D.W., 'Implicit Ordering of Defaults in Inheritance Systems', *Proceedings of the AAAI-84*, Austin, Texas, 1984, pp 322-325.

[Touretzky *et. al.* 87] Touretzky D.W., Horty J.F., Thomason R.H., 'A Clash of Intuitions: The Current State of Non-monotonic Multiple Inheritance Systems', *Proceedings of the IJCAI-87*, Milano, Italy, 1987.

A Probabilistic Soundness

In order to prove the system sound, we will enumerate the standard axioms of probability [Cox 46]; they are:

P-1. $0 \leq P(Q|e) \leq 1$

P-2. $P(\mathbf{true}\,|e) = 1$

P-3. $P(Q|e) + P(\neg Q|e) = 1$

P-4. $P(QR|e) = P(Q|R,e)P(R|e) = P(R|Q,e)P(Q|e)$.

A sound inference rule would be one that, given highly likely premises, only derives highly likely conclusions. For that purpose, statements of the form $E \mid\!\!\sim_K h$ will be mapped to probabilistic statements of the form $P_K(h|E) \approx 1$; meaning that h is an almost certain conclusion of E in the background context K. $P_K(\cdot)$ denotes any probabilistic model of K. That is, $P_K(\cdot)$ stands for any probability distribution over the formulas of the language, such that, if $K = \langle L, D\rangle$ then $P_K(\cdot)$ satisfies the following conditions:

$$\begin{aligned} P_K(L|E) &= 1 \quad \text{for any body of evidence } E, \text{ and} \\ P_K(a|b) &\approx 1 \quad \text{for every default } a \rightarrow b \in D \end{aligned}$$

To prove an inference rule sound, we show that for any such probability distribution, the probability of its consequent is close to one when the probability of its antecedent is close to one.

Rule 1 is clearly sound from the definition of $P_K(\cdot)$. To show the soundness of rule 2, we need to show that if $P_K(h|E,L) = 1$, then $P_K(h|E) \approx 1$. This follows by noticing that $P_K(h|E) \geq P_K(h|E,L)\ P_K(L|E) = 1$.

To prove rule 3 sound, we have from axioms P-3 and P-4 that :

$$P_K(h|E) = P_K(h|E,E')\ P_K(E'|E)\ + P_K(h|E,\neg E')\ P_K(\neg E'|E)\,,$$

so that if, as in rule 3, we have that $P_K(h|E) \approx 1$ and $P_K(E'|E) \approx 1$ (and therefore $P_K(\neg E'|E) \approx 0$), then we must also have $P_K(h|E,E') \approx 1$.

Rule 4 is a straightforward consequence of axiom P-4. To show the soundness of rule 5 we use rules 1–4 shown already to be sound. First notice that $E \mid\!\!\sim_K E \vee E'$ holds by rule 2, thus, knowing that $E \mid\!\!\sim_K h$, we can derive $E, E \vee E' \mid\!\!\sim_K h$. Analogously, from $E' \mid\!\!\sim_K h$, we can derive $E', E \vee E' \mid\!\!\sim_K h$. Now from P1–P4 the following equation holds:

$$P_K(\neg h \wedge E_2|E_1) = P_K(\neg h|E_1,E_2)\ P_K(E_2|E_1)\,,$$

from which we can show that if $E_1, E_2 \mid\!\!\sim_K h$ holds, so must $E_1 \mid\!\!\sim_K E_2 \supset h$. We can show then that if $E \mid\!\!\sim_K h$ and $E' \mid\!\!\sim_K h$, we must have both $E \vee E' \mid\!\!\sim_K E_1 \supset h$ and $E \vee E' \mid\!\!\sim_K E_2 \supset h$, from which the conclusion of rule 5 follows by virtue of theorem 2.