SOUND DEFEASIBLE INFERENCE

Hector Geffner
Judea Pearl

September 1987
CSD-870058

TECHNICAL REPORT

CSD-8700XX

R-94

August 1987

SOUND DEFEASIBLE INFERENCE *

Hector Geffner & Judea Pearl

Cognitive Systems Laboratory

Computer Science Department

University of California, Los-Angeles, CA. 90024-1596

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

Sound Defeasible Inference

Hector Geffner Judea Pearl

August 13, 1987

Abstract

A new system of defeasible inference is presented having the following features:

- spurious extensions are prevented without forcing one to explicitly enumerate exceptions,
- the system has a sound, clear probabilistic semantics, guaranteeing that the consequences are highly probable whenever the premises are,
- the system is clean: proofs can be constructed very much like in natural deduction systems in logic.

Additionals implications of the framework proposed are precise, proof theoretic and semantic accounts of defaults, and a formalization of the notion of irrelevance in the context of non-monotonic reasoning.

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 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: every proposition is qualified by a degree of confidence 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 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[1] 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 a coarse-grained acceptance rule.

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 natural deduction systems in logic and, yet, is probabilistically sound. Among the implications of the framework proposed are new proof theoretic and semantic accounts of defaults, and a formalization of the notion of irrelevance in the context of non-monotonic reasoning.

The resulting system of defeasible inference is closely related to those systems proposed in the literature [Loui 86, Poole 85, Touretzky 84] which use the *structure of the arguments* to eliminate spurious extensions. In our approach however, the structure of arguments is not used for selecting an argument among many, but for preventing inferior arguments from ever being generated. The system's rules of inference occasionally examine the structure of the database, and extract from it a single meta-level predicate M which permits to infer only desirable conclusions.

The structure of the paper is as follows. In section 2 we define the object and meta-language, as well as the rules of inference which make up the the system of defeasible inference proposed. In section 3 we discuss its probabilistic semantics: we prove the system to be sound, and we conjecture it to be complete in a very interesting sense. In section 4 we go through a set of examples to show the applicability of the system proposed. In section 5 we discuss related work. We then investigate, in section 6, ways to enhance the expressiveness of the language to deal with defeasible defaults and reasoning about causality. Section 7 ends the paper with a brief summary.

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

2 A System of Defeasible Inference

2.1 Preliminary Definitions

The language comprises two types of formulas : logical formulas and defeasible rules of the form $P \rightarrow Q$, where P and Q are logical formulas. The intuition of a rule of that form, is that belief in the antecedent P, provides a reason to believe in Q. The precise probabilistic meaning of such rules will be given later.

A *context* is a pair $\langle L, D\rangle$ of logical formulas L and defeasible rules D. We will sometimes refer to the elements of L simply as formulas, and to the elements of D as defaults. In this subsection we will specify a set of conditions under which, a conclusion h obtained in a context $K = \langle L, D\rangle$, can still be preserved in the enhanced context $\langle L \cup E, D\rangle$, where E is an additional set of logical formulas. These preservation conditions define a meta-level predicate $M_K(h, E; L)$, which will later on be used in the inference rules.

We define an *argument* $\mathcal{A}^i(h; L, D)$ for formula h in context $\langle L, D\rangle$, as a sequence of formulas $F^i_{1,n} = \{F^i_1, ..., F^i_n\}$ such that:[2]

- $F^i_n = h$, and
- each $F^i_k, i = 1, ..., n$, is derived from the set of formulas $\mathcal{F}^i_k = L \cup F^i_{1,k-1}$ as:
    1. F^i_k is either a logical axiom, a logical theorem or a member of L,
    2. F^i_k logically follows from $\mathcal{F}^i_k$
    3. F^i_k is consistent with $\mathcal{F}^i_k$, and is the consequent of a default in D whose antecedent belongs to $\mathcal{F}^i_k$, or
    4. F^i_k is a disjunction of formulas G_l, consistent with $\mathcal{F}^i_k$, such that each G_l is the consequent of a default $H_l \rightarrow G_l$ in D, and the disjunction of the H_l's is in $\mathcal{F}^i_k$.

The existence of an argument $\mathcal{A}^i(h; L, D)$, does not necessarely sanction h as a legitimate conclusion in context $\langle L, D\rangle$, but it only indicates that h has a supporting reason. Note also, that we are restricting the formulas $\{F^i_1, ..., F^i_n\}$, in an argument $\mathcal{A}^i(h; L, D)$, to be consistent with the set of formulas L.

Since the notion of relevance will play an important role in the framework proposed, we will restrict the term 'argument', to those *which do not involve logical redundancies.*[3]

[2] The term argument is borrowed from [Loui 86]. His use of the term is very close to ours. Likewise, both are close to Touretzky's 'paths' [Touretzky 84] and Poole's 'theories' [Poole 85]; except for the irredundancy conditions introduced below.

[3] That is, if we let the *justification* of a formula F^i_k, $J(F^i_k)$, denote the subset of formulas in $\mathcal{F}^i_k$ used to derive it in $\mathcal{A}^i(h; L, D)$, then our irredundancy conditions amount to :

- Every formula $F^i_k, i = 1, ..., n-1$, takes part in the justification $J(F^i_l)$ of a formula F^i_l, $k \leq l \leq n$;

We say that h arguable in context $\langle L, D\rangle$, written as $L \mid\!\sim_{D} h$, if there exists an argument $\mathcal{A}^i(h; L, D)$ satisfying these conditions. The subset of L which takes part in such an argument will be called the *support* of the argument.

It will turn out to be useful to display the relationships embedded in a given context in the form of directed graphs, as the one depicted in fig. 1. Positive links ($\rightarrow$) connecting a proposition P to a proposition Q, will stand for either defaults $P \rightarrow Q$, or logical implications $P \supset Q$. Negative links ($\not\rightarrow$) will denote defaults of the form $P \rightarrow \neg Q$. Since we will usually only represent positive literals in the graphs, arguments will tend to appear as (hyper) paths composed of positive links, possibly ending with a negative link.

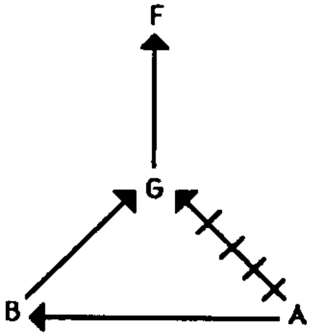

Figure 1: Arguments

For instance, we might take figure 1, as representing the context $\langle L, D\rangle$, with $L = \{\}$, and $D = \{A \rightarrow B, A \rightarrow \neg G, B \rightarrow G, G \rightarrow F\}$. In such a context, there exists an argument $\mathcal{A}(F; \{A\}, D)$, with support $\{A\}$ corresponding to the path $A \rightarrow B \rightarrow G \rightarrow F$ in the figure. There is also an argument $\mathcal{A}(F; \{A, B\}, D)$, with support $\{B\}$, which corresponds to the path $B \rightarrow G \rightarrow F$. Note however, that there is no argument $\mathcal{A}(F; \{A, B\}, D)$ with support involving A, since the truth of A does not add support to F, given the truth of B. Any such argument will involve logical redundancies, and will therefore be excluded by our irredundancy conditions. This is clear from the figure, in which the only path from A to F goes through B, which was also assumed to be part of the context $\langle\{A, B\}, D\rangle$. The presence of B in the context, renders A irrelevant to argue in favor of F. This is formalized in the next definition.

Definition: We say that a set of formulas R is *potentially relevant* to establish h in context $\langle L, D\rangle$, iff there is an argument $\mathcal{A}^i(h; R \cup L, D)$ with support S, such that $(S - L) \cap R \neq \emptyset$. Otherwise we say that R is *irrelevant* to establish h in $\langle L, D\rangle$, or that L *blocks* R from h.

- If a formula F_k^i can be derived according to case 1 above, $J(F_k^i) = \emptyset$.
- If a formula F_k^i logically follows from $J(F_k^i)$ (case 2 above), then it is not the case that F_k^i logically follows from a proper subset of $J(F_k^i)$.

That is, R is potentially relevant to h if it offers a *new* argument in favor of h. To test whether R is irrelevant to establish h in context $\langle L, D\rangle$, one needs to test whether L blocks all the paths from R to h. For instance, in figure 1 , we can easily infer that A is irrelevant to establish F in context $\langle\{B\}, D\rangle$, by noticing that B blocks the only path from A to F. Futhermore, since there are no paths to $\neg F$, A is also irrelevant for establishing the negation of F in such a context.

As we stated above, we are interested in finding a set of sufficient conditions that would allow to preserve a conclusion derived in a given context, when the context is enhanced with an additional set of propositions. We might be tempted to think, that irrelevance relative the negation of the conclusion would constitute such a set of sufficient conditions. A careful analysis of figure 1, however, will reveal that this condition is not sufficient. In the figure, F might be a reasonable conclusion when B is all we know. Furthermore, enhancing the context to include A does not produce counter-arguments in favor of its negation, $\neg F$. Still, since A is relevant to $\neg G$, and G is involved in the argument supporting F, A might potentially leave F without support. Thus, to guarantee that the belief in a given proposition h can be safely preserved upon learning a new set of facts E, we must go beyond the potential relevance of E to the negation of h, by also considering the impact that E might have on the arguments supporting h. This is the purpose of the following definition.

> **Definition:** We say that a set of formulas R *interferes* with an argument $\mathcal{A}^i(h; L, D)$, with formulas $F^i_{1,n}$, iff for some $1 \leq j \leq n$, there exists an argument $\mathcal{A}^k(\neg F^i_j; L \cup R, D)$, with formulas $F^k_{1,m}$ consistent with $\neg h$, i.e. $\mathcal{F}^k_{m+1} \not\vdash h$. Furthermore, if R is potentially relevant to $\neg F^i_j$, we say that R minimally interferes with the argument $\mathcal{A}^i(h; L, D)$.

In terms of graphs, a set of formulas R interfering with a given argument $\mathcal{A}^i(h; L, D)$ corresponds to the presence of a path from formulas in $L \cup R$, to the negation of some formula taking part in the argument. If such a path originates in R and it is not blocked by L, it means that R minimally interferes with such an argument. In such a case, if h was believed in context $\langle L, D\rangle$, and $\mathcal{A}^i(h; L, D)$ was its only supporting argument, enhancing the context to include R might potentially leave h without support, and therefore might lead to the retraction of our belief in h. That was in fact the case we discussed above, in which learning A in context $\langle\{B\}, D\rangle$ could lead to the retraction of the belief in F. In those cases, we will not only say that R interferes with an argument supporting h, but, as the next definition states, that R interferes with the formula h itself.

> **Definition:** We say that R *interferes with a formula* h in context $\langle L, D\rangle$ iff either R is potentially relevant to $\neg h$ in context $\langle L, D\rangle$, or R interferes with every argument $\mathcal{A}^i(h; L \cup R, D)$, and minimally interferes with, at least, one of those arguments.

Note that if R does not interfere with h in $\langle L, D\rangle$, it means either that there is an argument which is not interfered by R, or that R is irrelevant to establish the negation of any formula participating in arguments in favor of h.[4] In any case, if h was a reasonable conclusion in context $\langle L, D\rangle$, extending the context to include R will not provide a reason to retract our belief in h. This motivates the following definition.

> **Definition:** A formula h is said to be *monotonic on* R *in context* $\langle L, D\rangle$, iff R does not interfere with h in $\langle L, D\rangle$. We will denote this relation with the meta-level predicate $M_K(h, R)$, where $K = \langle L, D\rangle$. We will also write $M_K(h, R; L')$, when h is monotonic on R in the enhanced context $\langle L \cup L', D\rangle$. In this case we also say that L' *separates* R from h. If h is not monotonic on R in a given context, we say that R *undermines* h in that context.

diagrammatically, determining whether $M_K(h, R; L')$ holds in a context $K = \langle L, D\rangle$, amounts to testing either, that there is a path from $L \cup L'$ to h, not interfered by R, or, that the formulas in $L \cup L'$, block all the paths from R which interfere with formulas potentially relevant to h.[5].

2.1.1 Examples

Example 1. Let us consider the context $K = \langle L, D\rangle$, with :

$$\begin{aligned} L &= \{bird(x) \supset winged\_animal(x)\} \\ D &= \{winged\_animal(x) \rightarrow flies(x)\} \end{aligned}$$

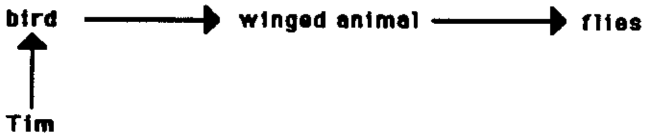

Figure 2: *winged_animal* separates *bird* from *flies*

The chain depicted in Fig. 2 displays the information conveyed by these formulas. We can easily see that there is an argument $\mathcal{A}(flies(Tim); L \cup \{bird(Tim)\}, D)$, associated with the path $bird(Tim) \rightarrow winged\_animal(Tim) \rightarrow flies(Tim)$. If $winged\_animal(Tim)$ is then learned, it would render $bird(Tim)$ irrelevant to $flies(Tim)$, since the former blocks the only path that connects the two. Additionally, since $bird(Tim)$ does not interfere with the path that connects $winged\_animal(Tim)$ to $flies(Tim)$, we have that

[4] And therefore, R does not interfere with any such a formula. Otherwise R would be minimally interfering with an argument for a formula which supports h, and therefore, minimally interfering with an argument for h.

[5] These diagrammatical considerations however, are no substitute of the formal definitions. They represent only intuitive guidelines, which will turn out to be sufficient for many of the examples we are going to deal with in this paper

$M_K(flies(Tim), bird(Tim); winged\_animal(Tim))$ holds. This almost trivial relation will provide our system with the capability to produce sound chains of inference.

Example 2. Let us consider the context $K = \langle L, D\rangle$, depicted in Fig. 3, with:

$$\begin{aligned} L &= \{adult(Tom)\} \\ D &= \{u\_student(x) \rightarrow adult(x), adult(x) \rightarrow work(x), \\ &\quad u\_student(x) \rightarrow \neg work(x), adult(x) \wedge under\_22(x) \rightarrow u\_student(x)\}. \end{aligned}$$

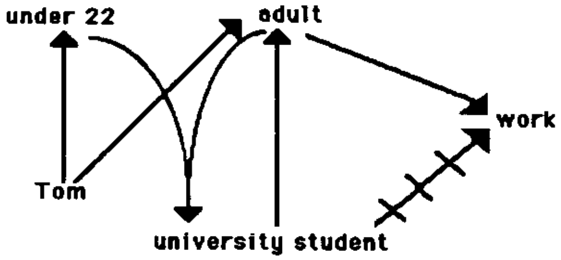

Figure 3: *u_student* and *adult* separate *under_22* from both *work* and $\neg$*work*

Clearly $work(Tom)$ is not monotonic on $under\_22(Tom)$ in context K, since $under\_22(Tom)$ provides a new argument in favor of its negation $\neg work(Tom)$, which is not blocked by L. On other hand, $work(Tom)$ becomes monotonic on $under\_22(Tom)$, when $u\_student(Tom)$ is learned, since the argument in favor of $\neg work(Tom)$ is blocked by the presence of $u\_student(Tom)$ in the context.

2.2 The Rules of Inference

In this subsection we present a system of defeasible inference made up of six rules of inference together with the meta-level predicate M defined earlier. In section 3 we provide its probabilistic semantics and discuss its soundness and completeness properties.

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. The system of inference implicitly defines the set of conclusions h that follow from the enhanced context $\langle L \cup E, D\rangle$. We will denote such a relation as $E \mathrel{|\!\!\approx_K} h$, and say that h follows from the evidence set E in context K, or simply that h can be derived from E in K. E and E' represent any sets of logical formulas. The rules are:

Rule 1 (Defaults)
If $E \rightarrow h \in D$ then $E \mathrel{|\!\!\approx_K} h$

Rule 2 (Logic theorems)
If $L \cup E \vdash h$ then $E \mathrel{|\!\!\approx_K} h$

Rule 3 (Frame axiom)
If $E \mathrel{|\!\!\approx}_K h$ and $M_K(h, E'; E)$ then $E, E' \mathrel{|\!\!\approx}_K h$

Rule 4 (Triangularity)
If $E \mathrel{|\!\!\approx}_K h$ and $E \mathrel{|\!\!\approx}_K E'$ then $E, E' \mathrel{|\!\!\approx}_K h$

Rule 5 (Bayes)
If $E \mathrel{|\!\!\approx}_K E'$ and $E, E' \mathrel{|\!\!\approx}_K h$ then $E \mathrel{|\!\!\approx}_K h$

Rule 6 (Deduction)
If $E, E' \mathrel{|\!\!\approx}_K h$ then $E \mathrel{|\!\!\approx}_K \neg E' \vee h$

Rule 1 says that if the background context includes a defeasible rule whose antecedent is all that has been learned, then its consequent can be concluded. Rule 2 states that theorems that logically follow from a set of formulas can be concluded in any theory containing those formulas. Rule 3 establishes that a derived proposition remains so, when an additional set of facts is learned which does not undermine the conclusion in the current context. Rule 4 states that the incorporation of a set of established conclusions to the current context, does not affect the status of any other derived conclusions. Rule 5 says that any conclusion that follows from the current context augmented by a set of conclusions established in that context, also follows from the current context alone. Rule 6 says that if a conclusion follows from a context augmented by a set of formulas, then either the proposition or the negation of (the conjunction of the formulas in) the set follow from the context.

2.2.1 Some Meta-Theorems

Theorem 1 (Logical Closure 1): If $E \mathrel{|\!\!\approx}_K h$ and $E, h \vdash h'$ then $E \mathrel{|\!\!\approx}_K h'$.
It follows by sequentially applying rules 2 and 5

Theorem 2 (Logical Closure 2): If $E \mathrel{|\!\!\approx}_K h$, $E \mathrel{|\!\!\approx}_K h'$, and $E, h, h' \vdash h''$, then $E \mathrel{|\!\!\approx}_K h''$.
By rule 4, we obtain that $E, h \mathrel{|\!\!\approx}_K h'$. From rule 2, we get $E, h, h' \mathrel{|\!\!\approx}_K h''$. Applying then rule 5 twice, the theorem is proved.

Theorem 3 (Weak Transitivity): If $E \mathrel{|\!\!\approx}_K E'$, $E' \mathrel{|\!\!\approx}_K h$ and $M_K(h, E; E')$, then $E \mathrel{|\!\!\approx}_K h$.
It follows by sequentially applying rules 3 and 5.

Theorem 4 (Equivalent contexts) : If $E \equiv E'$ and $E \mathrel{|\!\!\approx}_K h$, then $E' \mathrel{|\!\!\approx}_K h$.
Since $E \vdash E'$, by applying rules 2 and 4 we get $E, E' \mathrel{|\!\!\approx}_K h$; which together with $E' \vdash E$ and rules 2 and 5, leads to $E' \mathrel{|\!\!\approx}_K h$.

Theorem 5 (Disjunction) : If $E \mathrel{|\!\!\approx}_K h$ and $E' \mathrel{|\!\!\approx}_K h$, then $E \vee E' \mathrel{|\!\!\approx}_K h$.
By theorem 1 and rule 4 we have that $E', E \vee E' \mathrel{|\!\!\approx}_K h$, and therefore $E \vee E' \mathrel{|\!\!\approx}_K \neg E' \vee h$. Using the same arguments we obtain $E \vee E' \mathrel{|\!\!\approx}_K \neg E \vee h$. The conclusion then follows

from theorem 2, and $E \vee E', \neg E \vee h, \neg E' \vee h \vdash h$.

Some non-theorems :

$E \vdash E'$ and $E' \mathrel{|\!\sim}_K h$ do not necessarely imply $E \mathrel{|\!\sim}_K h$.

$E \mathrel{|\!\sim}_K h$ and $E' \mathrel{|\!\sim}_K h$ do not necessarely imply $E, E' \mathrel{|\!\sim}_K h$.

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

3 Examples

Example 3. Let us consider the theories $T_1 = (K, E_1)$ and $T_2 = (K, E_2)$, with background context $K = \langle L, D \rangle$,

$$\begin{aligned} L &= \{penguin(x) \supset bird(x)\} , \\ D &= \{penguin(x) \rightarrow \neg flies(x), bird(x) \rightarrow flies(x)\}, \end{aligned}$$

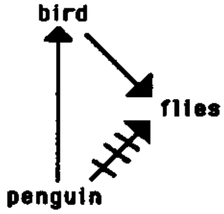

Figure 4: Penguins are birds which usually do not fly

$E_1 = \{penguin(Tim)\}$, and $E_2 = \{penguin(Tim), bird(Tim)\}$. Concluding that 'Tim does not fly' in context K knowing that 'Tim is a penguin' amounts to proving $penguin(Tim) \mathrel{|\!\sim}_K \neg flies(Tim)$. The proof gets reduced to a single application of rule 1, since $penguin(Tim) \rightarrow \neg flies(Tim) \in D$.

Proving $E_2 \mathrel{|\!\sim}_K \neg flies(Tim)$ is slightly different since a new fact, $bird(Tim)$, needs to be assimilated. The proof goes as follows:

1. $penguin(Tim) \mathrel{|\!\sim}_K \neg flies(Tim)$ by rule 1
2. $penguin(Tim) \mathrel{|\!\sim}_K bird(Tim)$ by rule 2
3. $penguin(Tim), bird(Tim) \mathrel{|\!\sim}_K \neg flies(Tim)$ by rule 4 on lines 1 and 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 4, 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 $T(K, E_1)$ and $T(K, E_2)$. 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 the new theory $T_1' = (K', E_1')$, defined in terms of T_1, with $K' = \langle L', D \rangle$, $L' = \emptyset$ and $E_1' = L \cup E_1$. T_1' appears 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 $\neg flies(Tim)$ cannot be established from T_1', i.e., $E_1' \not\vdash_{K'} \neg flies(Tim)$. 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. Being an independent piece of evidence, which supports the opposite conclusion, the implication learned cannot be assimilated by the system to preserve the conclusion that penguins usually do not fly.

What this shows is that logical 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.[6]

Notice that if a system of defeasible inference were to allow the derivation of $\neg flies(Tim)$ in background context K' from the evidence set $E_p = \{penguin(Tim), penguin(Tim) \supset bird(Tim)\}$; by symmetry reasons it should also allow the derivation of the opposite conclusion $flies(Tim)$, in the same context K', from the evidence set $E_b = \{bird(Tim), bird(Tim) \supset penguin(Tim)\}$, yet both E_p and E_b are logically equivalent to $\{penguin(Tim), bird(Tim)\}$. This also illustrates that the preference for the conclusion that penguins do not fly, in spite of beings birds, is not to be explained in terms of logical relations, 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 about the ability of penguins to fly.

It is interesting to note, that while it has long been acknowledged that the 'meaning' of defaults depends on the 'theory' in which they appear, 'theories' were normally taken as composed only of a set of logical formulas and a set of defeasible rules. From this perspec-

[6] The semantics of defaults will be treated below, in section 4.

tive, theories T_1 and T_1', should be equivalent, and they could not differ in the conclusions they entail. As we have seen, in our framework these theories do differ since they have different **background contexts**. Moreover, from both the proof theoretic and semantic accounts of defaults in the proposed framework, the 'meaning' of defaults that emerges does not depend on the whole theory, but just on the background context in which they are defined.

Example 4. Let us consider the theory $T = (K, E)$, with $K = \langle L, D\rangle$, and

$$\begin{aligned} L &= \{\}, \\ D &= \{u\_student(x) \rightarrow adult(x), adult(x) \rightarrow work(x), u\_student(x) \rightarrow \neg work(x), \\ &\quad adult(x) \wedge under\_22(x) \rightarrow u\_student(x)\}, \\ E &= \{adult(Tom), under\_22(Tom)\}. \end{aligned}$$

Before proceeding, we will briefly describe a proof strategy common to most of the examples we are going to analyze (see Fig. 5). Roughly, the strategy consists of three main (recursive) steps :

1. Select a set E' of formulas which separates E from h in $\langle L, D\rangle$,
2. Partition E' into two subsets E_1' and E_2', and prove $E_1' \vdash_K h$ and $E_1' \vdash_K E_2'$;
3. Prove $E \vdash_K E'$.

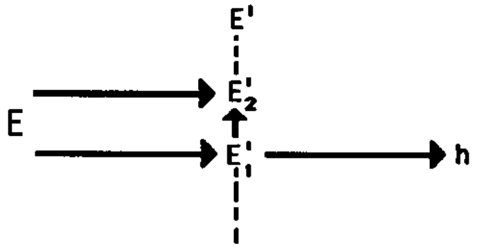

Figure 5: A common proof strategy to show $E \vdash_K h$. Dark arrows represent proofs.

A proof for $E \vdash_K h$ can then be built by noticing that rule 4 allows to conclude $E' \vdash_K h$ from step 2, which by virtue of theorem 4 (weak transitivity), together with the results of steps 1 and 3, yields the desired conclusion $E \vdash_K h$. The proof strategy is displayed in Figure 5, where dark arrows stand for proofs, e.g. $E \vdash_K E_1'$. The vertical bar indicates that $E' = E_1' \cup E_2'$, separates E from h in context K, i.e., that h is monotonic on E in context $\langle L \cup E', D\rangle$.

For this example, depicted in Figure 6, we are want to show that $adult(Tom), under\_22(Tom) \vdash_K \neg work(Tom)$. The set $E' = \{adult(Tom), student(Tom)\}$, does in fact separate E from $\neg work(Tom)$. Showing then $E' \vdash_K \neg work(Tom)$ is simple, and

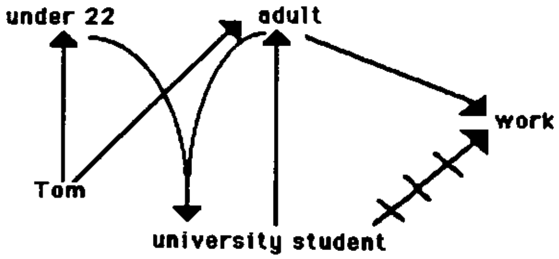

Figure 6: Adults under 22 usually do not work

follows from rule 4, with $student(Tom) \mathrel{|\!\sim}_K \neg work(Tom)$ and $student(Tom) \mathrel{|\!\sim}_K adult(Tom)$. It only remains to show $E \mathrel{|\!\sim}_K E'$, which follows from rules 1 and theorem 1.

It is interesting to note, that we can also derive $adult(x) \mathrel{|\!\sim}_K \neg student(x)$. Letting a stand for an arbitrary constant, we can show as above that $adult(a), student(a) \mathrel{|\!\sim}_K \neg work(a)$. Therefore, by rule 6, we have that $adult(a) \mathrel{|\!\sim}_K \neg student(a) \vee \neg work(a)$, which together with $adult(a) \mathrel{|\!\sim}_K work(a)$, and theorem 2, yields the desired conclusion.

Example 5.[Sandewal 86, Touretzky *et. al.* 87]. Let $T = (K, E)$, $K = \langle L, D\rangle$ and :

$$\begin{aligned} L &= \{royal\_elephant(x) \supset elephant(x), african\_elephant(x) \supset elephant(x)\}, \\ D &= \{elephant(x) \rightarrow gray(x), royal\_elephant(x) \rightarrow \neg gray(x)\}, \\ E &= \{royal\_elephant(clyde), african\_elephant(clyde)\}. \end{aligned}$$

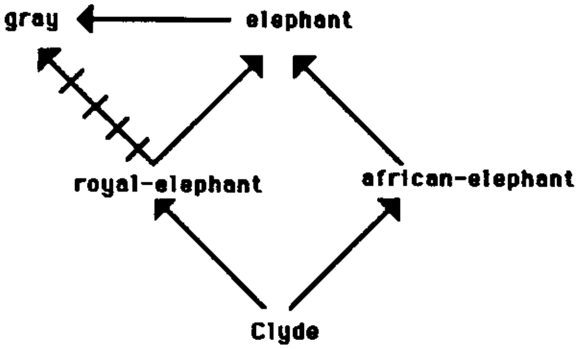

Figure 7: Clyde is not gray

The same proof strategy applies to show that $E \mathrel{|\!\sim}_K \neg gray(clyde)$, once we choose the separating set $E' = \{royal\_elephant(clyde), elephant(clyde)\}$.

Example 6. [Touretzky *et. al.* 87]. Let us consider now the theory $T = (K, E)$, with $K = \langle L, D\rangle$, and :

$$\begin{aligned} L &= \{\}, \\ D &= \{A \rightarrow B, A \rightarrow \neg G, B \rightarrow G, B \rightarrow C, C \rightarrow F, G \rightarrow \neg F\} \\ E &= \{A\}. \end{aligned}$$

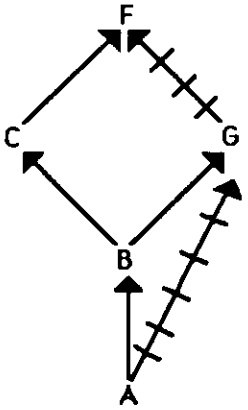

Figure 8: A's are F's

The goal is to prove that proposition F is entailed by A. For that purpose, we can show that $E' = \{C, \neg G\}$ separates the evidence A from the target proposition F. This follows from the fact, that $\neg G$ rules out any argument which involves proposition G. It is also possible to prove that $C, \neg G \mathrel{|\!\!\sim}_{K} F$, since $\neg G$ is irrelevant to $\neg F$. Then, since both $A \mathrel{|\!\!\sim}_{K} C$, and $A \mathrel{|\!\!\sim}_{K} \neg G$ can be shown to hold, we get $A \mathrel{|\!\!\sim}_{K} C \wedge \neg G$, and, therefore, the target proposition F.

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

$$\begin{aligned} L &= \{\} \\ D &= \{quaker(x) \rightarrow pacifist(x), republican(x) \rightarrow \neg pacifist(x)\} \\ E &= \{quaker(Nixon), republican(Nixon)\}. \end{aligned}$$

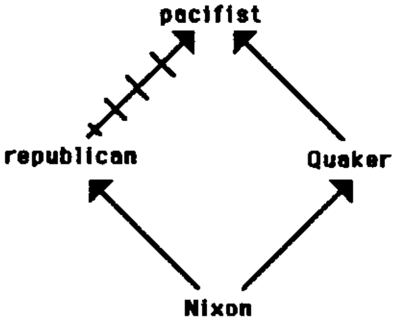

Figure 9: No conclusion can be drawn regarding Nixon's pacifism

In this theory, no conclusion regarding Nixon's pacifism can be drawn from E. 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 insufficient to integrate the available pieces of evidence in order to arrive to a conclusion. Enhancing the background context to include another default like that quakers who also are republicans are still pacifists, would solve the ambiguity without introducing any inconsistencies.

Example 8: (M. Ginsberg) Let us $T = (K, E)$, $K = \langle L, D \rangle$,

$$\begin{aligned} L &= \{\} \\ D &= \{quaker(x) \rightarrow dove(x), republican(x) \rightarrow hawk(x), dove(x) \rightarrow \neg hawk(x), \\ &\quad hawk(x) \rightarrow \neg dove(x), dove(x) \rightarrow p\_motivated(x), hawk(x) \rightarrow p\_motivated(x)\} \\ E &= \{quaker(Nixon), republican(Nixon)\}. \end{aligned}$$

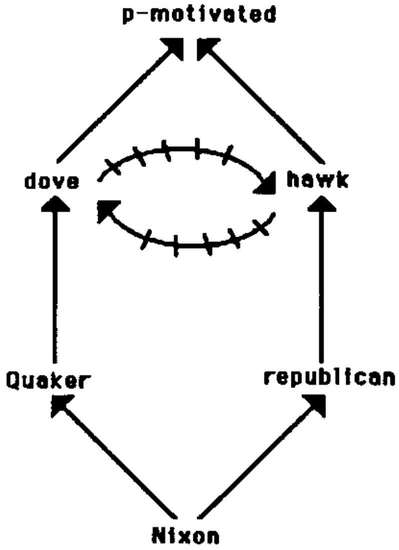

Figure 10: Nixon is politically motivated

We want to show that Nixon is politically motivated. If we let $E' = \{dove(Nixon) \vee hawk(Nixon)\}$, it is easy to see that $M_K(p\_motivated(Nixon), E; E')$ holds. Moreover since, by rules 1 and theorem 5, we can obtain that $E' \mathrel{|\!\approx}_K p\_motivated(Nixon)$, we can then infer by rule 3 that $E, E' \mathrel{|\!\approx}_K p\_motivated(Nixon)$. It can also be shown that $M_K(dove(Nixon) \vee hawk(Nixon), quaker(Nixon); republican(Nixon))$ holds, and, therefore, we can obtain $E \mathrel{|\!\approx}_K E'$. Finally, by rule 5, the target conclusion $E \mathrel{|\!\approx}_K p\_motivated(Nixon)$ is proved.

Example 9.[Horty *et al.* 87]. Let us consider the theory $T = (K, E)$, with $K = \langle L, D \rangle$, $L = \{\}$ and

$$\begin{aligned} D &= \{quaker(x) \rightarrow pacifist(x), republican(x) \rightarrow \neg pacifist(x), \\ &\quad republican(x) \rightarrow football\_fan(x), pacifist(x) \rightarrow anti\_military(x), \\ &\quad football\_fan(x) \rightarrow \neg anti\_military(x)\} \\ E &= \{quaker(Nixon), republican(Nixon)\}. \end{aligned}$$

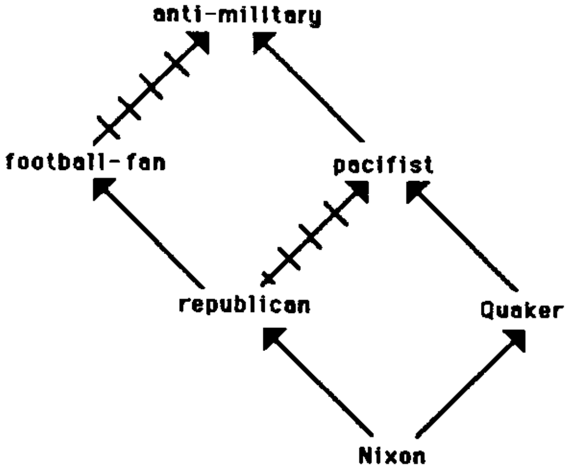

Figure 11: Is Nixon anti-military ?

In this example we can conclude $football\_fan(Nixon)$, but correctly fail to conclude anything regarding Nixon's pacifism or antimilitarism. This is in contrast with Horty's skeptical inheritance algorithm [Horty *et al.* 87], in which the ambiguity regarding Nixon's pacifism permits the conclusion $\neg anti\_military(Nixon)$. The uncommitment in our framework seems however justified. As figure 12 shows, equivalent topologies can be constructed in which the opposite conclusion seems more reasonable.

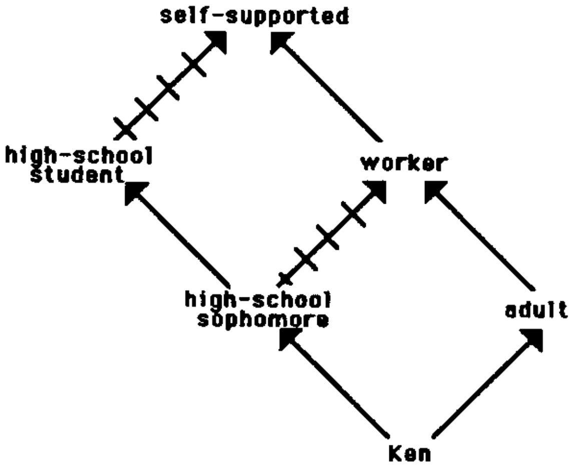

Figure 12: Is Ken self-supported ?

Example 10. Let us consider now the theory $T = (K, E)$, $K = \langle L, D\rangle$, with

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

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 $E \mid\!\approx_K \neg happy(John)$. 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 11. Let us consider $T = (K, E)$, $K = \langle L, D\rangle$, $L = \{\}$, $D = \{a \rightarrow c, a \rightarrow b, a \wedge c \rightarrow \neg b\}$ 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.

4 Probabilistic Semantics

4.1 Soundness

In this section we will be concerned with probabilistic soundness of the system proposed.[7] In order to prove the system sound, we will first 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.[8] For that purpose we are going to map statements of the form $E \mid\!\approx_K h$ in the meta-language 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 *admissible* probability distribution with respect to context K. That is, $P_K(\cdot)$ stands for any probability distribution over the formulas of the language, such that, if $K = \langle L, D\rangle$

[7] J. Pearl [Pearl 87b], 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 follows directly from the soundness of its rules of inference.

[8] It was recently brought to our attention that a similar paradigm was pursued by E. Adams [Adams 66], who devised a more restricted set of inference rules. In particular his formulation does not involve the 'frame assumption' embedded in our rule 3.

then $P_K(\cdot)$ satisfies the following conditions:

$$\begin{array}{rcl} P_K(L|E) & = & 1 \quad \text{for any body of evidence } E, \text{and} \\ P_K(a|b) & \approx & 1 \quad \text{for every default } a \rightarrow b \in D \\ P_K(h|R, L') & \approx & 1 \quad \text{if } P_K(h|L') \approx 1 \text{ , whenever } M_K(h, R; L') \text{ holds.} \end{array}$$

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 $P_K(h|E) \geq P_K(h|E, L)\ P_K(L|E) = 1$.

The soundness of rule 3 follows from the third constraint imposed on $P_K(\cdot)$. Such a constraint imposes a reasonable 'frame assumption' on any admissible probabilistic model, which states that a belief in a proposition does not change, unless there is a 'reason' to believe so. What constitutes such a reason was the subject of subsection 2.1. In probabilistic terms, it amounts both to make the probability of propositions without supporting arguments very low, and to ensure that the addition of supporting arguments would not render a proposition less likely.

For proving rule 4, 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 4, 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 it must be the case that $P_K(h|E, E') \approx 1$.

Rule 5 is a straightforward consequence of axiom P-4. To show the soundness of rule 6, note that from axiom P-4 :

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

If $E, E' \mathrel{|\!\!\approx}_K h$, we must have, from axiom P-3, that $P_K(\neg h|E, E') \approx 0$. Therefore from axioms P-3 and P-4 we can obtain that $P_K(\neg(\neg h \wedge E')|E) \approx 1$ which, combined with axiom P-2, leads to $P_K(h \vee \neg E'|E) \approx 1$ and, therefore, to the soundness of $E \mathrel{|\!\!\approx}_K h \vee \neg E'$.

4.2 Completeness Conjecture

The question arises whether the set of theorems $Th(K, E) = \{\alpha | E \mathrel{|\!\!\approx}_K \alpha\}$, coincides with the set of conclusions dictated by probabilistic considerations. We have the following conjecture :

Completeness Conjecture. Let $T = (K, E)$ be a theory, with an associated background context $K = \langle L, D \rangle$. Let $P^*_K(\cdot)$ stand for any probability distribution which satisfies the following conditions:[9]

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

Then, if for some proposition α it follows[10] from axioms P-1 to P-4 that $P^*_K(\alpha|E) \approx 1$, then $\alpha \in Th(K, E)$.

The idea is that an admissible probability distribution $P^*_K(\cdot)$, partially specified over a set of proposition with statements of the form $P^*_K(\cdot) \approx 1$, can only give rise to new $P^*_K(\cdot)$ entries either of the form $P^*_K(\cdot) \approx 1$ or of the form $P^*_K(\cdot) \approx 0$. We conjecture that rules 1,2,4,5 & 6, capture all these inferences.

It is clear that rules 1 & 2 capture all the original $P^*_K(\cdot)$ entries. An inductive argument showing that the five rules capture any single inference from the axioms should then suffice. Rule 2 captures all the entries produced by axiom P-2 alone. Rules 5 and 6, together with theorem 1 seem to capture the inferences based on axiom P-4 alone. Axiom P-3 seems to be more problematic. However combined with axiom P-2, it only leads to obvious inferences. Its power lies when combined with axiom P-4. In fact, axiom P-3 can be replaced by an equivalent axiom in the same context of axioms P-1, P-2 and P-4:

3′. $P(RQ|e) + P(R\neg Q|e) = P(R|e)$.

We show in the appendix how all the inferences that follow from this expression seem to be captured by the proposed set of rules.

5 Related Work

As noted in [Reiter *et. al.* 81], the logic for default reasoning proposed by Reiter in [Reiter 80] requires to explicitly state the exceptions of defaults, 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

[9] Note that $P^*_K(L|E) \approx 1$ obviously follows from $P^*_K(L|E) = 1$.

[10] We are assuming here, that statements of the form $x \approx 1$, for algebraic purposes, are equivalent to statements of the form $x = 1$. The distinction in the formulation, however, needs to be preserved, in order to avoid ruling as inconsistent the probability distributions induced by most of the interesting theories.

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 are going to 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. For instance, if we consider the context $K = \langle L, D \rangle$ (see Fig. 13), 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 $\neg B$, is now 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 $\neg B$, can 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.

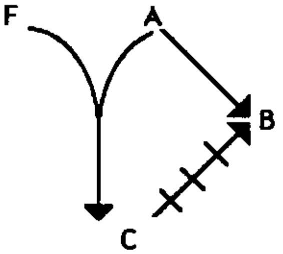

Figure 13: Loui's system would conclude $\neg B$, when given F and A.

The system reported by Touretzky in [Touretzky 84,86] was motivated by the goal of providing a semantics for inheritance hierarchies with exceptions. He argues that there exists 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 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.[11] It is interesting to note that both rule 4 (triangularity), and the proof strategy summarized in figure 5, seem to convey

[11] For instance, examples 3 and 8 above involve cycles.

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 4 (fig. 6), none of the arguments supporting the conclusion that Tom works, or that Tom does not work, can be determined to be more specific, if we ignore the default that states that most students are adults, which does not take part in the competing arguments. 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.

It is interesting to examine how our system gives rise to a new 'meaning' of defaults. Defaults have traditionally been taken to be very much like 'heuristic' rules that could be 'applied' to a given belief state, to get an extended belief state, whenever such application would not lead to inconsistencies. The 'meaning' of defaults that emerges from our framework is quite different. A default $P \rightarrow Q$, in a given background context, represents a clear cut constraint among beliefs. It states that if P is all that has been learned, then Q can be inferred. The non-monotonicity exhibited by the system is not the result of 'soft' default constraints, in contrast with 'hard' logical constraints, but the result of the context dependence of the former, absent from the latter. We have argued that these 'constraints' have a logic of their own, very much like the logic which governs classical connectives. For example, if 'most birds fly' and 'penguins are birds, but they do not fly', then it *must* be the case that 'most birds are not penguins'. Yet, we are not aware of any AI system of default reasoning which will draw such a conclusion.

6 Extending the expressiveness of the language

We have shown how most of the examples reported in the literature admit a solution within the framework proposed. Yet, it seems that there are many types of relationships which can not be reasonably coded neither as logical assertions or as defeasible rules. In this section we discuss possible benefits that could be gained by enhancing the expressivity of the language.

6.1 Defeasible Defaults

An advantage of the framework proposed here over other default reasoning systems (e.g. [Reiter 80]) is the absence of the multiple extension problem. The inference rules ensure that derived conclusions lie in the single 'preferred' extension. In default logic, the approach taken to filter spurious extensions, was the use of non-normal defaults in which exceptions are stated explicitly [Reiter *et. al.* 81, Etherington *et. al.* 83]. We have shown that in many cases these exceptions do not need to be explicated to achieve the desired behavior. Still, there are many cases in which exceptions might be needed. For instance, we might want to express the facts that 'adults usually work, unless they are students', meaning that being an adult is a reason to conclude that s/he works, except when s/he is believed to be a student. The difficulty to code this type of exceptions in our framework arises, because these exceptions, rather than providing a counter argument to the consequent of the default, only invalidate arguments based on it. That is, we do not want to imply that students do not work, but, only, that adults known to be students do not necessarely work.

In section 2, we discussed the conditions under which a set of formulas R interferes with a given proposition h in a given context K. The idea was that if R interferes with h in K, extending the context K to include R might result in the retraction of the belief in h. For that purpose we appealed to the concept of a set of formulas R interfering with a given argument in a given context. We said that R interferes with an argument $\mathcal{A}^i(h; L, D)$, when there is a counter-argument $\mathcal{A}^k(\neg F^i_j; L \cup R, D)$, for one of its formulas F^i_j. The natural way to incorporate defeasible defaults in our framework, is to extend this definition, to allow exceptions to interfere with arguments that appeal to defaults that they preclude. In particular, if d is a default with exception x_d , then any set containing x_d will interfere with any argument which involves default d. We shall also need to restrict arguments, so that if $\mathcal{A}^i(h; L, D)$ is an argument which uses default d with exception x_d, and $F^i_{1,n}$, are the formulas in the in the argument, then $L \cup F^i_{1,n} \not\vdash x_d$.

This modification of the definition of 'interference' makes the monotonicity in context predicate M more restrictive, thus restricting the application of rule 3. A background context will include now, not only a set of logical formulas L and a set of defaults D, but also a set of default exceptions X. Each default exception will be a pair $\langle x_d, d\rangle$, meaning that x_d is an exception of default d.

Example 12. Let us consider the theory $T = (K, E)$, with $K = \langle L, D, X\rangle$ and

$$\begin{array}{rcl} L & = & \{TA(x) \supset work(x)\}, \\ D & = & \{student(x) \rightarrow adult(x), adult(x) \rightarrow work(x)\}, \\ X & = & \{\langle student(x), adult(x) \rightarrow work(x)\rangle\} \\ E & = & \{student(Peter)\} \end{array}$$

Clearly $\neg M_K(work(Peter), student(Peter); adult(Peter))$, since $student(Peter)$ interferes now with the only argument for $work(Peter)$. So nothing can be concluded regarding whether Peter works or not. If in the present context, $TA(Peter)$ is learned, then $work(Peter)$ would follow.

6.2 Reasoning about causality

In their AAAI-86 paper, Hanks and McDermott addressed an issue which served as the main motivation for this work: how can knowledge be expressed as a set of logical formulas and defaults, so as to allow the derivation of all, and nothing but, the 'reasonable' conclusions that follow. They looked at a simple example from the domain of temporal reasoning, and showed how Reiter's default logic and McCarthy's circumscription failed to derive a conclusion which seemed to be implicit in the given set of axioms.

A simplified version of the problem (without quantification) would be :

l_0: gun loaded at time t_0
l_1: gun loaded at time $t_1 > t_0$
A_0: Ringo is alive at time t_0
A_1: Ringo is alive at time t_1
S_1: gun shot at Ringo at time t_1

with $T = (K, E)$, $K = \langle L, D\rangle$ and:

$L = \{\}$
$D = \{l_0 \rightarrow l_1, A_0 \rightarrow A_1, l_1 \wedge S_1 \rightarrow \neg A_1\}$
$E = \{l_0, A_0, S_1\}$

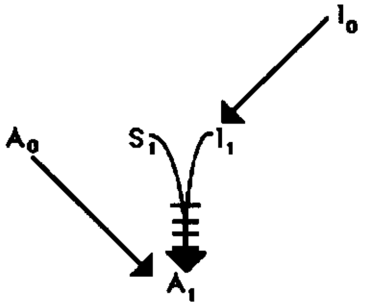

Figure 14: Shooting puzzle

In this formulation of the problem, it is possible to derive l_1, and still fail to derive what seems to be the reasonable conclusion $\neg A_1$. Note however that concluding $\neg A_1$ from such a theory will not be sound — just change the interpretation of A_0 to 'Ringo alive at time t_0, wearing a metal vest'. Clearly in such a case, deriving $\neg A_1$ will not be as reasonable.

In a way, the resulting ambiguity resembles the ambiguity found in the 'Nixon diamond' (see Example 7), in which it was not possible to integrate in a single conclusion the pieces of evidence supporting and denying the pacifism of Nixon. In the current example though, the resulting ambiguity appears counter-intuitive. It seems as if there is some additional semantic information for the reader of the example which allows her/him to derive the desired conclusion, which is missing from the formal formulation of the problem. [12]

The ambiguity exhibited in the example, lies in the fact that it is not always the case that 'the pattern of influences' among sets of propositions contains all the information necessary to reason about their overall combined effect. As important as the structure of the relation among propositions, is the *nature* of those relations. For the example above, the fact that a person was alive, does not affect the expectation that if shot with a loaded gun would stop living. In probabilistic terms this amounts to $P_K(\neg A_1|S_1, l_1, A_0) \approx P_K(\neg A_1|S_1, l_1)$, and therefore, if we were able to conclude $S_1, l_1 \mid\!\approx_K \neg A_1$, we should be able to derive $S_1, l_1, A_0 \mid\!\approx_K \neg A_1$. In our framework however, the conclusion follows in the first context but not in the second. The problem being that $M_K(\neg A_1, A_0; S_1, l_1)$ does not hold, while it should hold, according to the intuitions that led to its definition.

It seems that the natural way to overcome the *syntactic myopia* of M, would consist of enhancing the object level language to include *explicit independence assertions among propositions*, and extending the inference machinery to take these independence assertions into account. A first step in this direction would be the addition of another rule of inference, rule 3', which will take into account those weak independences that escape the syntactic machinery of M :

Rule 3' (weak semantic independences)
If $E \mid\!\approx_K h$ and $I_K(h, E'; E)$ then $E, E' \mid\!\approx_K h$,

where $I_K(h, E'; E)$, denotes the fact that h is weakly independent of E' given E in context K; in probabilistic terms :

$$P_K(h|E, E') \approx 1 \quad \text{if } P_K(h|E) \approx 1 \,.$$

In the 'shooting' example above, we would extend the background context to $K = \langle L, D, I \rangle$, where $I = \{I_K(\neg A_1, A_0; S_1 \wedge l_1)\}$, stands for a set of weak independences assertions, whose only member states that 'the conclusion that a person shot with a loaded gun would die, would not be affected by learning that the person was alive before'. Rule 3' would then allow to maintain the conclusion $S_1, l_0 \mid\!\approx_K \neg A_1$, when we also consider the evidence A_0, i.e. we get the desired result $S_1, l_0, A_0 \mid\!\approx_K \neg A_1$.

Note that $I(\cdot)$ is the semantic counterpart of $M(\cdot)$. The monotonicity in context predicate attempts to extract all those weak independences which follow from the (syntactic) structural relations among the propositions in the theory. $I(\cdot)$ would play an analogous role to $M(\cdot)$ in proof theoretic terms; but it would reflect (weak) **semantic** independences, which can only be specified by the user.

[12] Note that including A_0 in L would lead to the desired conclusion. It does not seem however, that this 'solution' would be general enough.

6.3 Extensions: Discussion.

Interestingly, the probabilistic semantics of the framework we propose, suggests possible ways in which the expressiveness of the language can be enhanced. As it follows from the discussion in section 4, we can think of defining a background context by providing a set of logical formulas L, a set of defeasible rules D, a set of default's exceptions X, and a set of weak independences I; as a way to **partially specify** a probability distribution $P_K(\cdot)$, which implicitly sanctions the set of admissible conclusions.

If we take into account the extensions we have just discussed, then for a background context $K = \langle L, D, X, I\rangle$, $P_K(\cdot)$ would be partially specified by statements of the form :

- $P_K(L|E) = 1$ for any body of evidence E,
- $P_K(a|b) \approx 1$, for any default $a \rightarrow b$ in D,
- $P_K(a|b) \approx 1$, but not necessarely $P_K(a|b, c) \approx 1$, for any default $a \rightarrow b$ with exception c in X,
- $P_K(a|b, c) \approx 1$ if $P_K(a|b) \approx 1$, for weak independence assertions of the form $I_K(a, c; b)$ in I.

This view suggests other ways in which the expressivity of the language could be further enhanced. An interesting direction to investigate, would be to allow strong independence assertions in the language, with semantics :

$$P_K(h|E) = P_K(h|E, E') \, .$$

These probabilistic statements are known to possess a logic of their own [Pearl *et. al.* 86c], and might turn out to be important for reasoning about causality.

Let us also add, that as important as providing the language with the desired expressive power, is the design of a set of primitives in which relevant pieces of knowledge could be easily coded. In this respect, it is also worth looking for rich semantic primitives (perhaps like 'predicts', 'causes', 'suggests', etc), from which the semantic independence assertions could automatically be extracted, rather than explicitly asserted by the user.

7 Summary

The main contribution of the proposed framework for defeasible inference is the emergence of a 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 *must* 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 all the inferences permitted by our system are authorized in light of the probabilistic interpretation. Moreover, we have also conjectured, that this set of inferences is identical to the set of inferences allowed from probabilistic considerations.

Additionally we have defined a meta-level predicate M, which embodies a set of sufficient conditions under which a belief in a proposition can be reasonably preserved when an additional set of facts is learned. Predicate M is used, in fact, very much like a frame axiom: we assume that the belief in a proposition does not change, unless there is a 'reason' to believe so. Subsection 2.1 specified what constitutes such a reason.

The scheme proposed 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 derive the meta-level predicate M. But, in contrast to most non-monotonic logics, there is no circularity in its definition. M is derived in terms of arguments, while it is used to build proofs.

We have also briefly discussed possible ways to enhance the expressive power of the language, by accomodating both defeasible defaults and explicit independence assertions. We have argued that the latter might turn out to be important for reasoning about causality.

Acknowledgment.

Reading [Loui 86] prompted us to realize that it should be possible to embed a notion similar to probabilistic independence in a predicate M, computable by purely syntactic considerations.

We want to thank Michelle Pearl, for having drawn all 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.

[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.

[Horty *et. al.* 87] Horty J.F, Thomason R.H., and Touretzky D.S., 'A Skeptical Theory of Inheritance in Non-monotonic Semantic Nets', *Proceedings of the AAAI-87*, 1987, pp. 358-363.

[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.

[McCarthy 1984] McCarthy J., 'Applications of Circumscription to Formalizing Common Sense Knowledge', *Proceedings of the AAAI Workshop on Non-Monotonic Reasoning*, 1984, pp 295-324.

[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 Belief Commitement in Multi-Hypothesis Interpretation', 2nd. AAAI Workshop on Uncertainty in AI, 1986, Philadelphia, PA., also in *AI Journal*, 33, No 2, Oct. 87.

[Pearl 86c] Pearl J. and Verma T., 'The Logic of Representing Dependencies by Directed Graphs', *Proceedings of the AAAI-87*, Seattle, WA, July 1987, pp 374-379.

[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 86] Touretzky D.W., *The Mathematics of Inheritance Systems*, Morgan Kaufmann, Los Altos, California, 1986.

[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 Completeness Conjecture (cont'd)

We want to show that all the new entries that follow from the partial specification of $P^*_K(\cdot)$, given in section 4.2, according to the equation :

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

are captured by the rules of inference proposed. We exhaustively analyze all the cases. Probabilistic statements of the form $P^*_K(S|R) \approx 1$, are translated to $R \mathrel{|\!\approx_K} S$, while statements of the form $P^*_K(S|R) \approx 0$, are translated to $R \mathrel{|\!\approx_K} \neg S$.

1. If $E \mathrel{|\!\approx_K} h$ and $E \mathrel{|\!\approx_K} E'$, then $E, E' \mathrel{|\!\approx_K} h$.
   This is in fact rule 4.

2. If $E \mathrel{|\!\approx_K} h$ and $E, E' \mathrel{|\!\approx_K} \neg h$, then $E \mathrel{|\!\approx_K} \neg E'$ and $E, \neg E' \mathrel{|\!\approx_K} h$.
   We have already seen in example 4, the proof for the first consequent, which follows from rule 6 and theorem 2. The second follows from the first antecedent, together with the first consequent and rule 4.

3. If $E \mathrel{|\!\approx_K} \neg h$ and $E \mathrel{|\!\approx_K} E'$, then $E, E' \mathrel{|\!\approx_K} \neg h$.
   Again this is simply the triangle rule 4.

4. If $E \mathrel{|\!\approx_K} \neg h$ and $E, E' \mathrel{|\!\approx_K} h$, then $E \mathrel{|\!\approx_K} \neg E'$ and $E, \neg E' \mathrel{|\!\approx_K} \neg h$.
   The proof is the same as in line 2, with h substituted by $\neg h$.

5. If $E, E' \mathrel{|\!\approx_K} h$ and $E \mathrel{|\!\approx_K} E'$, then $E \mathrel{|\!\approx_K} h$.
   This is rule 5.

6. If $E, E' \mathrel{|\!\approx_K} h$ and $E, \neg E' \mathrel{|\!\approx_K} h$, then $E \mathrel{|\!\approx_K} h$.
   This follows from theorems 4 and 5.