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

ON THE LOGIC OF DEFAULTS

Hector Geffner

July 1988
CSD-880058

Submitted to Proceedings AAAI-88, St. Paul, Minnesota, August 1988

On the Logic of Defaults

Technical Report
R-110
June, 1988

Hector Geffner
Cognitive Systems Lab.
Dept. of Computer Science, UCLA
LA, CA 90024

Abstract

We present an alternative interpretation of defaults which draws on probability theory and notions of relevance. The result is a system made up of a body of six rules which appears to overcome some of the weaknesses of other non-monotonic logics proposed in AI. We also analyze several examples and discuss some of the issues that require further research.

1 Introduction

A main feature exhibited by commonsense reasoning is the ability to jump to conclusions which additional information might later defeat. The limitation of classical logic to handle this kind of reasoning, has in recent years prompted the development of non-monotonic logics: logics in which the addition of new axioms might render old theorems invalid (see [Ginsberg, 87]).

The usual approach for defining these logics has been to extend classical first order logic by appealing to notions such as consistency [McDermott and Doyle, 80; Reiter, 80] or minimal models [McCarthy, 80;86]. More recently however, these logics have become subject of closer scrutiny and some of their weaknesses have become more apparent (see e.g. [Reiter and Criscuolo 81; Hanks and McDermott, 86; Morris, 87]). These analyses have revealed that the interpretation of defaults provided by these formalisms is weaker than what appears to be the intended interpretation. Conclusions that appear to be implicit in a given set of defaults fail to be sanctioned and, furthermore, as no semantic account of defaults themselves is provided, it is usually not clear where the source of the difficulties lie.

We argue here that there is more to default reasoning than non-monotonicity. We say that defaults represent hard, context-dependent constraints among beliefs and, as such, obey certain laws. Our approach is to uncover such laws and incorporate them into the logic. For that purpose, and following [Geffner and Pearl, 87], we advocate an interpretation of defaults which draws on probability theory and notions of relevance. We show that not only does the resulting system of defeasible inference usually exhibits the intended preferences when dealing with interacting defaults, but that it also provides a perspective from which such preferences can be understood.

The proposed scheme is presented in section 2.1 In the rest of section 2 we analyze several examples and introduce some refinements. In section 3 we discuss some issues that require further research.

2 A Logic of Defeasible Inference

2.1 Preliminary Definitions

Conventions. We use roman capital letters A, B, ... as syntactic variables standing for first order wffs, and capital italic letters Γ, K, E, ... for sets of closed wffs or sentences. Object level formulas are typed in typewriter style, e.g. $\exists \mathtt{x}.\mathtt{block}(\mathtt{x})$. Tuples of variables are represented by $\mathbf{x}$, $\mathbf{y}$, ... while $\mathbf{a}$, $\mathbf{b}$, ... stand for tuples of ground terms. The symbols '$\vdash$' and '$\nvdash$' stand for provability and non-provability in first order logic with equality, respectively. Material implication is represented by the symbol '$\Rightarrow$.' For a set S of formulas, we use $\phi(S)$ to refer to the formula obtained by conjoining the formulas in S. When no confusion arises, we omit the $\phi(\cdot)$ operator and write, for instance, $\vdash \neg S$, as a shorthand for $\vdash \neg\phi(S)$.

The logic we shall present will be referred as **L** and will be characterized by a body of six rules of inference. The goal of **L** is to sanction as theorems the highly likely consequences that follow from a given context. A context E_K is built from two sets of wffs: a set K of sentences presumed to be true in every conceivable situation, called the *background context*, and a set of E of facts which characterize a particular situation and referred here as the *evidential set*.

Defaults are represented in K by sentences of the form $\forall \mathbf{x}.\mathrm{A}(\mathbf{x}) \wedge \neg \mathbf{ab_i}(\mathbf{x}) \Rightarrow \mathrm{B}(\mathbf{x})$, where A and B are wffs with free variables among those of $\mathbf{x} = \{x_1, \ldots, x_n\}$, and with $\mathbf{ab_i}$ playing the role of McCarthy's abnormal predicate. As we assume different defaults to involve different abnormality predicates, we shall sometimes abbreviate such defaults as $\Delta_{\mathbf{i}}(\mathbf{x})$. For a particular tuple $\mathbf{a}$ of ground terms, the formula $\mathrm{A}(\mathbf{a}) \wedge \neg \mathbf{ab_i}(\mathbf{a}) \Rightarrow \mathrm{B}(\mathbf{a})$ represents a particular default instance, sometimes abbreviated as $\Delta_{\mathbf{i}}(\mathbf{a})$.

Abnormality predicates $\mathbf{ab_i}$ receive a special treatment in **L**. For a tuple of ground terms $\mathbf{a}$, sentences of the form $\neg \mathbf{ab_i}(\mathbf{a})$ are regarded as *candidate assumptions*, i.e. they may be assumed to hold in certain contexts. When the assumption $\neg \mathbf{ab_i}(\mathbf{a})$ holds, we also say that the default instance $\Delta_{\mathbf{i}}(\mathbf{a})$ holds, and viceversa. A *candidate assumption set* simply refers to a finite set of candidate assumptions.

We say that a candidate assumption set AS is consistent in context Γ, if $\Gamma \nvdash \neg AS$. A formula H derivable from a context Γ augmented by a consistent candidate assumption set AS, will be said to be *arguable* in such a context, and we shall refer to such a derivation as an *argument* for H in Γ, and to AS as the *support* of the argument.

L defines an irrelevance predicate $I(\cdot)$, which is used to certify whether it is legitimate to jump to a defeasible

conclusion in a given context. Roughly, the idea is that if H represents an assumption believed in context Γ, and E' represents an additional body of evidence, then belief in H is authorized to persist as long as E' does not provide *additional* support for H's negation, or, as we shall say, when E' is *irrelevant* to $\neg$H in context Γ. This is captured by the following definition:

> **Definition.** A set of sentences E' is said to be irrelevant to a sentence H in context Γ, written $I(\mathrm{H}; E'|\Gamma)$, iff for any candidate assumption set AS, such that $E', \Gamma \not\vdash \neg AS$ and $E', \Gamma, AS \vdash \mathrm{H}$, we also have that $\Gamma, AS \vdash \mathrm{H}$.

This definition of irrelevance possesses a convenient graphical interpretation we shall often exploit. For instance, fig. 1, depicts a background context K with formulas:

(1) $\forall \mathbf{x}.\mathtt{B}(\mathbf{x}) \wedge \neg \mathtt{ab}_1(\mathbf{x}) \Rightarrow \mathtt{F}(\mathbf{x})$

(2) $\forall \mathbf{x}.\mathtt{P}(\mathbf{x}) \wedge \neg \mathtt{ab}_2(\mathbf{x}) \Rightarrow \neg \mathtt{F}(\mathbf{x})$

(3) $\forall \mathbf{x}.\mathtt{P}(\mathbf{x}) \Rightarrow \mathtt{B}(\mathbf{x})$

(4) $\forall \mathbf{x}.\mathtt{CB}(\mathbf{x}) \Rightarrow \mathtt{B}(\mathbf{x})$

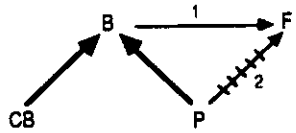

Figure 1: B separates CB from F, i.e. $I(\mathtt{F}(\mathbf{t}); \mathtt{CB}(\mathbf{t})|K, \mathtt{B}(\mathbf{t}))$

Paths in this type of graphs[1] correspond to arguments and, irrelevance, to a form of graph separation.[2] For instance, the path CB $\rightarrow$ B $\rightarrow$ F suggests that for any particular individual t, F(t) is derivable from CB(t), K, and any support including the assumption $\neg \mathtt{ab}_1(\mathbf{t})$. Notice that provided any such support, it is easy to verify that F(t) can be also derived from B(t) and K, what amounts to say, considering that there are no more paths from CB to F, that CB(t) is irrelevant to F(t) in context $\{\mathtt{B}(\mathbf{t})\}_K$, i.e. $I(\mathtt{F}(\mathbf{t}); \mathtt{CB}(\mathbf{t})|K, \mathtt{B}(\mathbf{t}))$.

Usually we will show a set of sentences E' to be irrelevant to a sentence H in a context E_K, by showing that in the corresponding graph, all the relevant paths that connect nodes corresponding to formulas in E' to the node that corresponds to H, are mediated by E. Clearly in such situations, if from a given support, H is not derivable from E and K, H will be certainly not derivable from E, K and E'. We should keep in mind, however, that links 'contrapose'. So, a path from P to $\neg$F not only represents an argument for $\neg$F(t) given P(t), but also an argument for $\neg$P(t) given F(t). The reader might verify for instance, in the example above, that, by virtue of the different 'signs' of the links converging to F, B(t) is relevant to $\neg$P(t) in K.

As for the most part the background context will remain fixed, we will find useful to abbreviate $I(\mathrm{H}; E'|\Gamma)$, with $\Gamma = E_K$, as $I_K(\mathrm{H}; E'|E)$. We also say that E' is relevant to H in context Γ, whenever $I(\mathrm{H}; E'|\Gamma)$ does not hold.

2.2 The Rules of Inference

The core of L is given by two sets of inference rules. We write $\Gamma \vdash \mathrm{H}$ to denote that sentence H is derivable from context Γ. Likewise, $\Gamma, E' \vdash \mathrm{H}$ states that H is derivable from the context that results from augmenting Γ with E'. Notice that the provability relation associated with the symbol '$\vdash$' is not monotonic: $\Gamma, E' \vdash \mathrm{H}$ does not always follow from $\Gamma \vdash \mathrm{H}$.

The first set of rules is given by [Geffner and Pearl, 87]:

Rule 1 (Logic Theorems)
If $\Gamma \vdash \mathrm{H}$ then $\Gamma \vdash \mathrm{H}$

Rule 2 (Triangularity)
If $\Gamma \vdash \mathrm{H}'$ and $\Gamma \vdash \mathrm{H}$ then $\Gamma, \mathrm{H}' \vdash \mathrm{H}$

Rule 3 (Bayes)
If $\Gamma \vdash \mathrm{H}'$ and $\Gamma, \mathrm{H}' \vdash \mathrm{H}$ then $\Gamma \vdash \mathrm{H}$

Rule 4 (Disjunction)
If $\Gamma, \mathrm{H}' \vdash \mathrm{H}$ and $\Gamma, \mathrm{H}'' \vdash \mathrm{H}$ then $\Gamma, \mathrm{H}' \vee \mathrm{H}'' \vdash \mathrm{H}$

It can be shown [Pearl and Geffner, 88] that the consequences of each rule are guaranteed to be highly likely whenever its premises are. Similar rules were proposed by Adams in his logic of conditionals [Adams, 66].

Hereafter, considering that the background context remains fixed for the most part, we will find useful to abbreviate $K, E \vdash \mathrm{H}$ as $E \vdash_K \mathrm{H}$.

Rules 1–4 express how conclusions that hold in one context can be carried to a slightly different context provided certain conditions are satisfied. They do not specify however, the contexts under which candidate assumptions in K, i.e. defaults, can be assumed to hold. In particular, they do not authorize to infer that Tweety flies for instance, given that Tweety is a bird and that typically birds fly. This issue is addressed by another pair of rules, the first of which, specifies the initial context in which a given candidate assumption might be assumed to hold, while the second one uses such assumption to 'jump' to conclusions not refuted by the evidence.

Clearly, if $\forall \mathbf{x}.\mathrm{A}(\mathbf{x}) \wedge \neg \mathtt{ab}_\mathtt{i}(\mathbf{x}) \Rightarrow \mathrm{B}(\mathbf{x})$ is a default in K, then for a tuple of ground terms a, it is reasonable to assume $\neg \mathtt{ab}_\mathtt{i}(\mathbf{a})$ to hold when A(a) is *all* that is believed. Each default, however, is a belief in itself, not formed in vacuum, but on top of other relevant and irrelevant beliefs. Here we assume K to partially model such set of beliefs for every default in it, thus, authorizing for a default $\forall \mathbf{x}.\mathrm{A}(\mathbf{x}) \wedge \neg \mathtt{ab}_\mathtt{i}(\mathbf{x}) \Rightarrow \mathrm{B}(\mathbf{x})$, the following inference

[1] In these graphs, we usually label the link that corresponds to default $\Delta_\mathtt{i}(\mathbf{x})$ with the index i, in order to facilitate reference.

[2] A similar correspondence between graph separation and conditional independence has been extensively exploited by Judea Pearl in the context of probabilistic networks (see for instance [Pearl and Verma, 87]) We borrow here some of his terminology.

rule:[3][4]

Rule 5 (Assumptions)
If $A(\mathbf{a}), K \not\vdash \mathbf{ab_i(a)}$ then $A(\mathbf{a}), K \vdash \neg\mathbf{ab_i(a)}$

As we shall see, such assumption turns out to be quite reasonable provided we restrict K to contain only statements whose truth does not depend on the particular context (e.g. "penguins are birds"), leaving in E, the context dependent information available (e.g. "Tweety flies").[5]

Still, the rules above are not sufficient for maintaining derived conclusions in the presence of additional, but irrelevant information. For instance, while rules 1–5 authorize to conclude that Tweety flies, given that it is a bird and that birds typically fly, they fail to preserve such conclusion upon learning, say, Tweety's color. This issue is addressed by an additional rule which appeals to the notion of irrelevance introduced above. The idea essentially is that a default $\forall\mathbf{x}.A(\mathbf{x}) \wedge \neg\mathbf{ab_i(x)} \Rightarrow B(\mathbf{x})$ permits 'jumping', say, from $A(\mathbf{a})$ to $B(\mathbf{a})$, whenever we know the relevant assumption $\neg\mathbf{ab_i(a)}$ to hold, and the new evidence does not provide an argument supporting its negation. More precisely:

Rule 6 (Irrelevance)
If $\Gamma, A(\mathbf{a}) \vdash \neg\mathbf{ab_i(a)}$ and $I(\mathbf{ab_i(a)}; E' | \Gamma, A(\mathbf{a}))$, then $\Gamma, E', A(\mathbf{a}) \vdash B(\mathbf{a})$

We argue below that what matters when testing the legitimacy of inferring $B(\mathbf{a})$ from $A(\mathbf{a})$ in context Γ when coming to know E', is not the existence of arguments in support of $\mathbf{ab_i(a)}$ but, more accurately, the existence of arguments for $\mathbf{ab_i(a)}$ in which the new information E' plays a role. These are precisely the arguments which sanction the relevance of E' to $\mathbf{ab_i(a)}$.

In order to illustrate this last point, consider for instance a candidate assumption $\neg\mathbf{ab_i(a)}$ believed in context Γ. Usually, in such context there would be different sets of assumptions AS_i logically inconsistent with $\neg\mathbf{ab_i(a)}$. For instance, in a context including information about the flying abilities of penguins and birds, the assumption that corresponds to the default instance "if Tweety is a penguin then it does not fly," will be logically inconsistent with the assumption that corresponds to the default instance "if Tweety is a bird then it flies," whenever Tweety is known to be a penguin. In such cases, *independently* of the new information E', any argument whose support includes any of the sets of assumptions AS_i inconsistent with $\neg\mathbf{ab_i(a)}$ in Γ, will automatically constitute an argument for $\mathbf{ab_i(a)}$ in the context $\{\Gamma, E'\}$. What the definition of irrelevance above simply does, is not to take those arguments into account: for E' to be relevant to $\mathbf{ab_i(a)}$ in Γ, there has to be an argument for $\mathbf{ab_i(a)}$ with a support AS', logically consistent with $\neg\mathbf{ab_i(a)}$ in Γ.

Note that in particular, if $\neg\mathbf{ab_i(a)}$ represents an assumption believed in Γ and $\neg\mathbf{ab_k(b)}$ represents an assumption logically inconsistent with $\neg\mathbf{ab_i(a)}$ in Γ, i.e. $\Gamma, \neg\mathbf{ab_i(a)} \vdash \mathbf{ab_k(b)}$, not only does L authorizes to 'ignore' the default instance $\Delta_\mathbf{k}(\mathbf{b})$ corresponding to $\neg\mathbf{ab_k(b)}$ as long as $\neg\mathbf{ab_i(a)}$ is believed[6], but to ignore such default instance even in order to evaluate the relevance of new information to $\mathbf{ab_i(a)}$. We say in thoses cases that $\neg\mathbf{ab_i(a)}$ *dominates* the assumption $\neg\mathbf{ab_k(b)}$ in Γ, and thus, the default instance $\Delta_\mathbf{k}(\mathbf{b})$.

Finally, we summarize a couple of meta-theorems that follow from the rules above, we shall later appeal to:[7]

Theorem 1 (Logical Closure)
If $E \mathrel{|\!\sim_K} H$, $E \mathrel{|\!\sim_K} H'$, and $H, H' \vdash H''$, then $E \mathrel{|\!\sim_K} H''$.

Theorem 2 (Exceptions)
If $E \mathrel{|\!\sim_K} H$ and $E, H' \mathrel{|\!\sim_K} \neg H$ then $E \mathrel{|\!\sim_K} \neg H'$.

2.3 Examples

Example 1. Let us first consider a background context K in which it is known that both penguins (P) and circus-birds (CB) are birds (B), and that most birds fly (F), though most penguins do not (Fig. 1):

$$\forall\mathbf{x}.\mathtt{B}(\mathbf{x}) \wedge \neg\mathbf{ab_1(x)} \Rightarrow \mathtt{F}(\mathbf{x})$$
$$\forall\mathbf{x}.\mathtt{P}(\mathbf{x}) \wedge \neg\mathbf{ab_2(x)} \Rightarrow \neg\mathtt{F}(\mathbf{x})$$
$$\forall\mathbf{x}.\mathtt{P}(\mathbf{x}) \Rightarrow \mathtt{B}(\mathbf{x})$$
$$\forall\mathbf{x}.\mathtt{CB}(\mathbf{x}) \Rightarrow \mathtt{B}(\mathbf{x})$$

Let us now say we learn about a penguin called Tim. We can then conclude by means of rule 5 that $\neg\mathbf{ab_2(Tim)}$ holds in context $\{\mathtt{P(Tim)}\}_K$, i.e. $\mathtt{P(Tim)} \mathrel{|\!\sim_K} \neg\mathbf{ab_2(Tim)}$. Likewise, being L closed under logical implication (Theorem 1), we can further conclude $\mathtt{P(Tim)} \mathrel{|\!\sim_K} \neg\mathtt{F(Tim)}$.

Note that extending the context $\{\mathtt{P(Tim)}\}_K$ to include $\mathtt{B(Tim)}$, does not affect either conclusion since, by means of rule 2 and the fact that $\mathtt{P(Tim)} \mathrel{|\!\sim_K} \mathtt{B(Tim)}$ follows (rule 1), formulas that hold in context $\{\mathtt{P(Tim)}\}_K$, can also be shown to hold in the enhanced context $\{\mathtt{P(Tim)}, \mathtt{B(Tim)}\}_K$. In particular thus, we obtain

$$\mathtt{P(Tim)}, \mathtt{B(Tim)} \mathrel{|\!\sim_K} \neg\mathtt{F(Tim)}.$$

L does not authorize reasoning in the opposite direction though. While $\mathtt{B(Tim)} \mathrel{|\!\sim_K} \neg\mathbf{ab_1(Tim)}$ and, as a consequence, $\mathtt{B(Tim)} \mathrel{|\!\sim_K} \mathtt{F(Tim)}$ can be derived, the conclusion $\mathtt{B(Tim)}, \mathtt{P(Tim)} \mathrel{|\!\sim_K} \mathtt{F(Tim)}$ cannot. Nor is $\mathtt{P(Tim)}$ irrelevant to $\mathbf{ab_1(Tim)}$ in context $\{\mathtt{B(Tim)}\}_K$, as the presence of an argument for $\mathbf{ab_1(Tim)}$ in $\{\mathtt{B(Tim)}, \mathtt{P(Tim)}\}_K$ with support $\{\neg\mathbf{ab_2(Tim)}\}$ suggests, nor is $\mathtt{P(Tim)}$ a consequence of $\mathtt{B(Tim)}$.

Interestingly, we also have that, in context $\{\mathtt{P(Tim)}\}_K$, the assumption $\neg\mathbf{ab_1(Tim)}$ is *dominated* by the assumption $\neg\mathbf{ab_2(Tim)}$. That is, we have both

$$\mathtt{P(Tim)} \mathrel{|\!\sim_K} \neg\mathbf{ab_2(Tim)}$$

and

$$\mathtt{P(Tim)}, \neg\mathbf{ab_2(Tim)}, K \vdash \mathbf{ab_1(Tim)}.$$

[3] Note that rule 5 permits deriving $B(\mathbf{a})$ from $A(\mathbf{a})$, but not $\neg A(\mathbf{a})$ from $\neg B(\mathbf{a})$. What amounts to say that the two logically equivalent sentences $\forall\mathbf{x}.A(\mathbf{x}) \wedge \neg\mathbf{ab_i(x)} \Rightarrow B(\mathbf{x})$ and $\forall\mathbf{x}.\neg B(\mathbf{x}) \wedge \neg\mathbf{ab_i(x)} \Rightarrow \neg A(\mathbf{x})$ are interpreted by L as encoding two different defaults. More about default contraposition in sections 3 and 4.

[4] The consistency test is for discarding from K some of the default instances otherwise implicit in the default 'schemas' in K, and its role should not be confused with the role consistency plays in other formalisms (e.g. [Reiter, 80; McDermott and Doyle, 80]). That convention allows us to write a 'unique-name hypothesis' , for instance, as: $\forall\mathbf{x}.\forall\mathbf{y}.\neg\mathbf{ab_1(x,y)} \Rightarrow \mathbf{x} \neq \mathbf{y}$, without implying those default instances in which $\mathbf{x} = \mathbf{y}$.

[5] More about this distinction in section 3.

[6] Since, in such case, we can show $\Gamma \vdash \mathbf{ab_k(b)}$ by means of rules 1 and 3.

[7] See [Geffner and Pearl, 87] for proofs.

Rule 6, as we discussed above, can then be understood as asserting that the default instance $\Delta_1(\mathtt{Tim})$ can be ignored in order to evaluate whether it is legitimate to 'jump' from $\mathtt{P(Tim)}$ to $\neg\mathtt{F(Tim)}$ in the presence of new facts. Or, more graphically, that the link connecting B to F, in what Tim is concerned, can be ignored as long as $\neg\mathbf{ab}_2(\mathtt{Tim})$ is believed. In particular then, we have that $\mathtt{CB(Tim)}$ turns out to be irrelevant to $\mathbf{ab}_2(\mathtt{Tim})$ in context $\{\mathtt{P(Tim)}\}_K$ and, thus, we obtain

$$\mathtt{P(Tim)}, \mathtt{CB(Tim)} \mid\!\sim_K \neg\mathtt{F(Tim)}.$$

L might be also regarded as legitimizing a weak form of contraposition. For instance, by virtue of Theorem 3 and the fact that we can derive both

$$\mathtt{B(Tim)} \mid\!\sim_K \mathtt{F(Tim)}$$

and

$$\mathtt{B(Tim)}, \mathtt{P(Tim)} \mid\!\sim_K \neg\mathtt{F(Tim)},$$

we have that

$$\mathtt{B(Tim)} \mid\!\sim_K \neg\mathtt{P(Tim)},$$

also follows. That is, if we assume a bird to fly, though we know that penguin-birds do not fly, we are implicitly assuming that the bird is not a penguin. Stronger forms of contraposition, as deriving $\neg\mathtt{B(Tim)}$ from $\neg\mathtt{F(Tim)}$ however, are not sanctioned by L.

Example 2. Consider the background context K given by the defaults:

$$\forall\mathbf{x}.\mathtt{P}(\mathbf{x}) \wedge \neg\mathbf{ab}_1(\mathbf{x}) \Rightarrow \mathtt{Q}(\mathbf{x})$$
$$\forall\mathbf{x}.\mathtt{Q}(\mathbf{x}) \wedge \neg\mathbf{ab}_2(\mathbf{x}) \Rightarrow \mathtt{R}(\mathbf{x})$$
$$\forall\mathbf{x}.\mathtt{S}(\mathbf{x}) \wedge \neg\mathbf{ab}_3(\mathbf{x}) \Rightarrow \neg\mathtt{R}(\mathbf{x})$$

Clearly, for an individual a, we can derive $\mathtt{P}(\mathbf{a}) \mid\!\sim_K \neg\mathbf{ab}_1(\mathbf{a})$ and, thus, $\mathtt{P}(\mathbf{a}) \mid\!\sim_K \mathtt{Q}(\mathbf{a})$. It turns out however, that the conclusion $\mathtt{Q}(\mathbf{a})$ results defeated if $\neg\mathtt{R}(\mathbf{a})$ is learned in such context. This is due to the fact that $\neg\mathtt{R}(\mathbf{a})$ does provide an argument for $\mathbf{ab}_1(\mathbf{a})$ supported by the assumption $\neg\mathbf{ab}_2(\mathbf{a})$, and thus, $I_K(\mathbf{ab}_1(\mathbf{a}); \neg\mathtt{R}(\mathbf{a}) | \mathtt{P}(\mathbf{a}))$ does not hold.

What this indicates is that while L does not consider default contrapositives to be strong enough as to authorize deriving the negation of the antecedent from the negation of the consequent, L does consider default contrapositives to be strong enough to make the latter relevant to the former, and thus precluding certain inferences to take place. In terms of Nute [86], contrapositives are treated in L only as *defeaters*.

Indeed, not only does L preclude deriving $\mathtt{Q}(\mathbf{a})$ from $\mathtt{P}(\mathbf{a})$ when $\neg\mathtt{R}(\mathbf{a})$ is learned, but even when $\mathtt{S}(\mathbf{a})$ is. We find this latter type of behavior counterintuitive though.[8] In the next subsection we shall propose a refinement of the definition of the irrelevance predicate $I(\cdot)$ given above which distinguishes between the two situations.

2.4 Contrapositives

The way L handles contraposition of defaults departs from other frameworks known to the author. Except for a weak form of contraposition, L does not permit to infer the negation of a default antecedent from the negation of its consequent, though it makes the latter relevant to the former, thus precluding certain dubious derivations to take place.

Still, as we discussed above, contrapositives appear sometimes to interfere with derivations that appear to be intuitively valid. These situations usually arise from the conflict of two 'expectation-evoking' defaults with incompatible consequents. Here we propose a simple refinement of the definition of irrelevance given above, which draws on the ideas of [Pearl, 88a], and which leaves those derivations undisturbed.

Pearl essentially argues that causality should play a distinctive role in default reasoning, and that, in particular, reasoning chains involving 'expectation-evoking' defaults (e.g. "if it rained, the grass is wet") followed by 'explanation-evoking' defaults (e.g. "if the grass is wet, the sprinkler was on") should not be authorized.

In our case, due to the fact that we assume defaults to be 'expectation-evoking,' and their contrapositives to be 'explanation-evoking,'[9] all we need to do, in order to enforce Pearl's maxim, is to prevent such chains of reasoning when computing the irrelevance predicate $I(\cdot)$.

The definition of $I(\cdot)$ above, amounts to sanction a set of sentences E' to be relevant to a sentence H in context E_K, whenever there is an argument for H in context $\{E \cup E'\}_K$ with support AS, consistent with E_K. The extra requirement we add is simply that, whenever $\Delta_{\mathbf{i}}(\mathbf{a})$ and $\Delta_{\mathbf{k}}(\mathbf{b})$ represent two 'expectation-evoking' default instances with consequents inconsistent in K, then AS does not simultaneously include the assumptions $\neg\mathbf{ab}_{\mathbf{i}}(\mathbf{a})$ and $\neg\mathbf{ab}_{\mathbf{k}}(\mathbf{b})$.

This simple proviso significantly improves the original account of irrelevance given above, and, in particular, correctly accounts for the type of counterintuitive behavior mentioned above.

From now on we will use $I(\cdot)$ to stand for this improved definition and will refer to the pair of conflicting 'expectation-evoking' defaults as forming a *causal fork*. The new definition can then be understood simply as preventing relevance to 'flow' through causal forks. We will also refer to the assumptions that correspond to defaults forming a causal fork, as *conflicting assumptions*.

We illustrate next how such refinement endows L with the ability to properly handle the "Yale Shooting Problem."[10]

Example 3. We consider next a version of the now famous "Yale Shooting Problem," presented in [Hanks and McDermott, 86] as an example in which both Reiter's logic and circumscription yield weaker conclusions than expected.

The puzzle says that people alive ($\mathtt{A}(t)$) typically remain alive ($\mathtt{A}(t+1)$) unless shot ($\mathtt{S}(t)$) with a loaded gun ($\mathtt{L}(t)$). Likewise, loaded guns ($\mathtt{L}(t)$) typically remain loaded ($\mathtt{L}(t+1)$):[11]

$$\forall\mathtt{t}.\mathtt{L}(\mathtt{t}) \wedge \neg\mathbf{ab}_1(\mathtt{t}) \Rightarrow \mathtt{L}(\mathtt{t}+1)$$
$$\forall\mathtt{t}.\mathtt{A}(\mathtt{t}) \wedge \neg\mathbf{ab}_2(\mathtt{t}) \Rightarrow \mathtt{A}(\mathtt{t}+1)$$
$$\forall\mathtt{t}.\mathtt{S}(\mathtt{t}) \wedge \mathtt{L}(\mathtt{t}) \wedge \neg\mathbf{ab}_3(\mathtt{t}) \Rightarrow \neg\mathtt{A}(\mathtt{t}+1)$$

[8] This type of behavior is also exhibited by circumscription and by Reiter's default logic, when defaults are encoded as to allow contraposition (see [Morris, 87]).

[9] Poole [87] makes a similar assumption.

[10] See also [Pearl, 88a].

[11] For clarity, we do not follow Hanks' and McDermott's use of a reified situation calculus. The formulation we use appears more comprehensible to us, while still serves to illustrate the difficulty detected by Hanks and McDermott in both circumscription and Reiter's default logic.

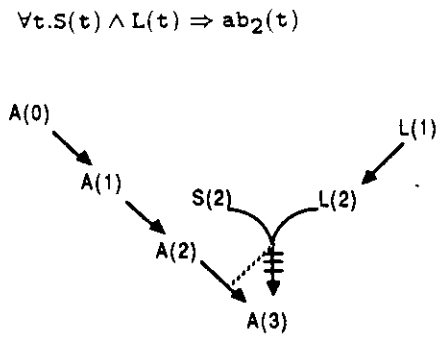

Figure 2: A version of the "Yale Shooting Problem"

We want to show that the person in question, called Fred, will most likely stop being alive if he is shot at time $t = 2$, with a gun loaded at $t = 1$, even if he was alive at time $t = 0$, i.e. we want to prove $A(0), L(1), S(2) \mid\!\sim_K \neg A(3)$. First notice that by virtue of rule 5, we have

$$S(2), L(2) \mid\!\sim_K \neg ab_3(2),$$

from which we can further infer, by means of rule 6

$$A(0), L(1), S(2), L(2) \mid\!\sim_K \neg A(3).$$

This in turn follows from the fact that

$$I_K(ab_3(2); A(0), L(1) | S(2), L(2))$$

holds, as a result of the assumption $\neg ab_2(2)$ being dominated by the assumption $\neg ab_3(2)$ in context $\{S(2), L(2)\}_K$.

Similar results would be obtained by circumscription and default logic. The behavior of **L** departs from these formalisms however, in which **L** is capable of further establishing[12] $L(2)$ from $L(1)$, $S(2)$ and $A(0)$, and, thus, by rule 3, the expected conclusion

$$A(0), L(1), S(2) \mid\!\sim_K \neg A(3).$$

Notice first that, by means of rule 5, we have that

$$L(1) \mid\!\sim_K \neg ab_1(1).$$

In order to evaluate whether $L(2)$ can be concluded upon learning both $S(2)$ and $A(0)$, we need to test whether $I_K(ab_1(1); A(0), S(2) | L(1))$ holds. In particular, we need to verify whether there is an argument for $ab_1(1)$ in the resulting context whose support does not include conflicting assumptions.

It turns out that the only argument for $ab_1(1)$ in context $\{L(1), A(0), S(2)\}_K$, does appeal to the conflicting assumptions $\neg ab_2(2)$ and $\neg ab_3(2)$ in its support, and, therefore, does not render $E' = \{A(0), S(2)\}$ relevant to $ab_1(1)$ in $\{L(1)\}_K$. It follows then by rule 6 that

$$A(0), L(1), S(2) \mid\!\sim_K L(2)$$

which, together with the previous result, leads by means of rule 3, to the expected conclusion

$$A(0), L(1), S(2) \mid\!\sim_K \neg A(3).$$

Let us finally remark that the derivation presented does not rest on a preference for 'reasoning forward' in time as opposed to 'reasoning backwards'. In particular, had we learned in addition that Fred survived the shot ($A(3)$), **L** would correctly have failed to authorize the conclusion that the gun was loaded at the time of the shooting ($L(2)$).

[12]Thanks to the improved definition of $I(\cdot)$. Otherwise **L** would have exhibited the same limitation.

3 Discussion

We have presented a system of defeasible inference motivated on probabilistic grounds and notions of relevance which provides an alternative basis for default reasoning. We have illustrated through examples how such an approach appears to overcome some of the weaknesses exhibited by other non-monotonic logics proposed in AI. In this section we want to propose some refinements and discuss some of the open issues.

1. Supported Propositions Circumscription and default logic appeal to either minimal models or fixed-point constructions in order to characterize the set of defeasible conclusions authorized in a given context. In particular, formulas that hold in a minimal model or extension of a given default theory, represents propositions which enjoy certain degree of support, while formulas which hold in none, stand for propositions with no support at all.

The classical example, is the "Nixon diamond:" we know quakers to be pacifists, republicans to be non-pacifists and Nixon to be both a quaker and a republican. Neither circumscription nor default logic expresses in this case any preference for believing either that Nixon is a pacifist or that he is a non-pacifist. Still, both formalisms distinguish between "Nixon is a pacifist," and, say, "Nixon is a soccer fan." The first proposition fails to be sanctioned because of conflicting evidence; the second, due to lack of support.

L does not appeal to either minimal models or fixed-point constructions and, therefore, does not account for such a distinction: neither proposition is derivable by its rules.[13] Still, a simple account for such a distinction can be constructed on top of **L**. Let us say that a proposition H is *supported* in context E_K, if there is a candidate assumption set AS not ruled out by the evidence, i.e. $E \not\mid\!\sim_K \neg AS$, such that $E, AS \mid\!\sim_K H$.

From such a definition it is possible to show that "Nixon is a pacifist" is supported, while "Nixon is a soccer fan" is not. More interestingly, it can be shown by means of Theorem 2, that if H is derivable from E_K, then no proposition inconsistent with H in such a context will be regarded as supported.

2. Background and Evidence. **L** naturally handles implicit exceptions. We have seen in the example 1 how subclasses override conflicting superclasses properties, without having explicated the corresponding 'abnormalities.' This results from the probabilistic interpretation of defaults embedded in the rules of **L**, together with the distinction between the formulas taking part of the background K from the formulas taking part of the evidential set E.

The latter distinction is specially important; "penguins," for instance, would not have overridden "birds" with respect to "flying" if we had stated the fact that "penguins are birds" in E rather than in K (see [Geffner and Pearl, 87]). As we pointed out in section 2, K should contain those sentences whose truth does not depend on context, and "penguins are birds" is one such sentence.[14]

[13]This point was raised by D. Etherington in relation to [Geffner and Pearl, 87].

[14]There are other frameworks for default reasoning that have appealed to distinctions of this sort. Two such examples are Poole's [85] scheme for comparing conflicting default theories

We might also regard K as *defining* the vocabulary which is *used* in E to characterize a particular context. As such, K encodes information about classes with no commitment at all about what their members are.

This distinction for instance, in the framework of inheritance networks, amounts to include in K the expressions that correspond to links among classes, leaving in E those which correspond to links connecting individuals to classes.

The question that remains to be addressed is whether such criterions for distinguishing K from E are sufficient for validating rule 5. While a number of examples here and in [Geffner and Pearl 87] suggest so, we expect a more general answer to evolve.

3. Soundness. Rules 1–4 represent the core of **L**. They share the inferential power of a probabilistic sound and complete system of rules developed by Adams [66] for capturing what he called the probabilistic consequences of a set of indicative conditionals.[15] The addition of rules 5 and 6 amounts to augmenting the probabilistic interpretation of defaults embedded in rules 1–4 with a set of conditional independence assumptions, drawn on the basis of the syntactic structure of the knowledge base.

Other syntactic and non-syntactic means of determining reasonable conditional independence assumptions must be possible. We have illustrated for instance how a refinement of the definition of $I(\cdot)$ originally provided, which takes into account the nature of the defaults involved, permitted certain reasonable inferences, otherwise precluded, to take place. Further refinements might be needed in order to capture other subtle aspects associated with causal defaults.

Another aspect that remains open, is a characterization of the provable consistent theories in the light of **L**. Though we expect such characterization to comprise most of the default 'benchmarks' reported in the literature, we are specially interested in those theories which can be mapped to graphs, and in which, reasoning, even in the presence of inconsistency, can be done 'meaningfully' and efficiently.

Acknowledgments

Many of the intuitions that led to this work originated in conversations with Judea Pearl. I also want to thank him, M. Fuenmayor and M. Goldszmidt for comments on earlier drafts of this paper.

This work was partially supported by the National Science Foundation grant IRI 86-10155.

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.

[Delgrande, 87] Delgrande J., "An Approach to Default Reasoning Based on a First-Order Conditional Logic". *Proceedings AAAI-87*, Seattle, 1987, pp 340-345.

[Geffner and Pearl, 87] Geffner H. and Pearl J., "A Framework for Reasoning with Defaults", TR-94b, October 1987, Cognitive Systems Lab., UCLA.

[Ginsberg, 87] Ginsberg M., editor. *Readings in Non-Monotonic Logics*, Morgan Kaufmann, Palo Alto, 1987.

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

[McCarthy 80] McCarthy J., "Circumscription–A Form of Non-Monotonic Reasoning", *Artificial Intelligence 13*, 1980, pp 27-39.

[McCarthy 86] McCarhty J., "Applications of Circumscription to Formalizing Commonsense Knowledge", *Artificial Intelligence 28*, 1986, pp 89-116.

[McDermott and Doyle, 80] McDermott D. and Doyle J., "Non-Monotonic Logic I", *Artificial Intelligence 13*, 1980, pp 41-72.

[Morris 87] Morris P., "Curing Anomalous Extensions", *Proceedings of the AAAI-87*, Seattle, 1987, pp 437-442.

[Nute 86] Nute D.,"LDR: a Logic for Defeasible Reasoning", ACMC Research Report 01-0013, University of Georgia, Athens, 1986.

[Pearl and Verma, 87] Pearl J. and Verma T. " The Logic of Representing Dependencies by Directed Graphs" *Proceedings AAAI-87*, Seattle, 1987, pp 374-379. Also in [Pearl, 88b].

[Pearl, 88a] Pearl J., "Embracing Causality in Default Reasoning", *Artificial Intelligence 35*, 1988, pp 259-271. Also in [Pearl, 88b].

[Pearl, 88b] Pearl J., *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, Los Altos, 1988.

[Pearl and Geffner, 88] Pearl J. and Geffner H., "Probabilistic Semantics for a Subset of Default Reasoning", TR-93-III, March 1988, Cognitive Systems Lab., UCLA. Also in [Pearl, 88b].

[Poole, 85] Poole D., "On the Comparison of Theories: Preferring the Most Specific Explanation", *Proceedings of the IJCAI-85*, Los Angeles, 1985, pp 144-147.

[Poole, 87] Poole D., "Defaults and Conjectures: Hypothetical Reasoning for Explanation and Prediction", Report CS-87-4, October 1987, University Waterloo.

[Reiter, 80] Reiter. R., "A Logic for Default Reasoning" *Artificial Intelligence 13*, 1980, pp 81-132.

[Reiter and Criscuolo, 81] Reiter R. and Criscuolo G., "On Interacting Defaults", *Proceedings IJCAI-81*, Vancouver, 1981, pp 270-276.

and Delgrande's [87] extended conditional logic, in which a distinction is made between sentences expressing necessary truths from those expressing contingent truths.

[15] Indicative conditionals of the form $a \rightarrow b$ are interpreted by Adams as asserting that the probability of b given a is infinitesimally close to one. See also [Pearl and Geffner, 88].