PROBABILISTIC SEMANTICS FOR INHERITANCE HIERARCHIES WITH EXCEPTIONS

Judea Pearl

September 1987
CSD-870052

TECHNICAL REPORT

CSD-8700XX

R-93-I

July 1987

PROBABILISTIC SEMANTICS FOR INHERITANCE HIERARCHIES WITH EXCEPTIONS *

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.

PROBABILISTIC SEMANTICS FOR INHERITANCE HIERARCHIES WITH EXCEPTIONS

Judea Pearl

Introduction

Let Γ be a collection of default statements of the form $I(p, q)$ where $I(p, q)$ means "p is typically a q" and $I(p, \neg q)$ reads "p is typically not a q".

Our task is to draw plausible conclusions from Γ. This requires that we establish a clear semantics for the meaning of each individual statement in Γ as well as for the absence of some statements NOT contained in Γ. For example, $\Gamma = \{I(a, bird), I(bird, fly)\}$ does not contain explicitly the statement $I(a, flies)$ neither $I(a, \neg flies)$ and, since every I allows exceptions, either one of the last two statements would be logically consistent with Γ. Yet, most people would regard the absence of $I(a, \neg flies)$ as a clue for the plausibility of $I(a, flies)$ but not vice versa; $I(a, \neg flies)$ might be accepted as a surprising fact but not as a conclusion.

The purpose of this report is to propose a probabilistic formulation that faithfully accounts for people's distinction between the plausible, the possible and the surprising. The formulation is offered as yet another standard for gauging the validity of proposed non-monotonic logics.

ε-Semantics

We regard Γ as a set of elastic restrictions imposed on possible worlds. A world is a complete assignment of property values to individuals. For example, the world w_0 could be describe by $(bird(a), \neg fly(a), \neg penguin(a))$ while another world, w_1, may have the description $(\neg bird(a), fly(a), penguin(a))$, possibly referring to some penguin-shaped kite.

Since some worlds are obviously more typical than others, it is natural to regard the sentences in Γ as a reflection of what we typically find in our experience. Our task, then, amounts to inferring new typical patterns of experience from the partial list of such patterns encoded in Γ. The inference would only be possible if Γ somehow restricts the sets of worlds that one regards as typical. The most elementary restriction is between a single sentence in Γ, say $I(p, q)$, and the sets of worlds describable by the primitive predicates p and q. For instance, $I(p, q)$ renders $W_0 = \{p(a), q(a)\}$ more typical than $W_1 = \{p(a), \neg q(a)\}$. Note that W_0 and W_1 are sets of worlds rather than singleton worlds because Γ may contain other predicates beside p and q. For example, $I(bird, fly)$ renders the set of worlds $W_0 = \{bird(a), fly(a)\}$ more typical than $W_1 = \{bird(a), \neg fly(a)\}$, where

$$W_0 = \{(penguin(a), bird(a), fly(a), (\neg penguin(a), bird(a), fly(a))\}$$

$$W_1 = \{(penguin(a), bird(a), \neg fly(a)), (\neg penguin(a), bird(a), \neg fly(a)\}$$

To guarantee that the restrictions Γ imposes on sets of worlds reflect coherent patterns of experience, we resort to the calculus of probability or, more specifically, to a subset of the calculus that deals with extreme probabilities, infinitesimally removed from either 0 or 1. Thus, the

sentence $I(a, bird)$ is interpreted to state that individual a is almost surely a bird, $P(bird(a)) = 1 - \varepsilon$, and $I(bird, fly)$ stands for $P(fly(x) \mid bird(x)) = 1 - \varepsilon$, namely, given that individual x is a bird, x is very likely to have flying abilities. ε is understood to stand for an infinitesimal quantity that can be made arbitrarily small without violating the plausibility of the inferences drawn. Categorical statements can, of course, be assigned apriori $\varepsilon = 0$. However, we shall see that such distinction does not lead to new insights, neither do we gain by assigning each statement a different ε with its own rate of vanishing.

The conclusions we wish to draw from Γ are those extreme probability statements that logically follow from Γ via the axioms of probability theory. In the absence of any constraining sentences in Γ one is licensed to assume any arbitrary probability distribution P over the sets of worlds and, consequently, no statement about the world would be preferred to its negation. Once we admit statements of the form $I(p, q)$ in Γ, these force the global distribution P to exhibit extreme pairwise conditional probabilities $P(q(x) \mid p(x)) = 1 - \varepsilon$ and these, in turn, might force other conditional probabilities to become extreme, thus qualifying new sentences as "Plausible Conclusions". For example, accepting $I(bird, fly)$ and $I(a, bird)$ into Γ would render the statement $fly(a)$ plausible and $\neg fly(a)$ implausible, as shown in the following derivation:

$$P(fly(a)) = P(fly(a) \mid bird(a))\, P(bird(a)) + P(fly(a) \mid \neg bird(a))\, P(\neg bird(a))$$

$$= P(fly(x) \mid bird(x), x = a)\,(1 - \varepsilon) + P(fly(a) \mid \neg bird(a))\,\varepsilon \quad (1)$$

$$= 1 - 2\varepsilon + \varepsilon^2 + O(\varepsilon)$$

(1) The transition from $P(fly(x) \mid bird(x), x = a)$ to $P(fly(x) \mid bird(x) = 1 - \varepsilon$ will be justified in the next subsection.

$$= 1 - O(\varepsilon)$$

Thus, we see that the conclusion *fly*(a) is compelled by virtue of the fact that Γ forces every probability distributions P to yield $P(\textit{fly}(a)) = 1 - O(\varepsilon)$.

We can formalize this construction by defining the set of distributions $\mathcal{P}_{\Gamma,\varepsilon}$ licensed by Γ for any given ε:

$$\mathcal{P}_{\Gamma,\varepsilon} = \left\{ P : P(v \mid u) = \left\{ \begin{array}{ll} 1-\varepsilon & \text{if } I(u,v) \in \Gamma \\ \varepsilon & \text{if } I(u,-v) \in \Gamma \end{array} \right\} \right. \tag{1}$$

Simultaneously, we restrict the set of conclusions that logically follow from Γ to only those that hold for every P in $\mathcal{P}_{\Gamma,\varepsilon}$.

Definition: A statement $S = I(p, q)$ is said to be a *plausible conclusion* of Γ, written $\Gamma \models_{\varepsilon} S$, if $P(S) = 1 - O(\varepsilon)$ for every $P \in \mathcal{P}_{\Gamma,\varepsilon}$[1].

Having defined the validity of sentences in terms of a constrained set of probability distributions does not mean, of course, that in practice one would have to manipulate numerical probability distributions in order to issue sound conclusions. This definition can be faithfully replaced by logical inference rules (Geffner, 1987), thus facilitating the derivation of new sound sentences by direct symbolic manipulations on Γ.

[1]. That $\Gamma \models_{\varepsilon} S$ is solely a function of Γ, independent on ε, is clear from the definition of $O(\varepsilon)$; a function $f(x_1, x_2, \cdots)$ is said to be $O(\varepsilon)$ if for every arbitrarily small quantity $\delta > 0$, one can find another small quantity $\varepsilon(\delta) > 0$ such that confining each of the arguments $x_1, x_2 \cdots$ to be smaller than ε, forces $f(x_1, x_2, \cdots)$ to be smaller than δ.

If a statement S acquires $1-\varepsilon$ certainty in some $P \in \mathcal{P}_{\Gamma,\varepsilon}$ but not all, it may be advisable not to rule it out altogether. Rather, we may wish to indicate its possibly being true by saying that S "is permitted by Γ", written $\Gamma \sim_\varepsilon S$. Moreover, if a reasoning system indicates the possibility of a set of statements $\{S_\alpha\}$ simultaneously, it is important to make sure they are all supported by the same P in $\mathcal{P}_{\Gamma,\varepsilon}$.

Definition: A set of statement $\{S_\alpha\}$ is said to be permitted by Γ, $\Gamma \sim_\varepsilon \{S_\alpha\}$, iff there exists $P \in \mathcal{P}_{\Gamma,\varepsilon}$ such that $P(S_\alpha) = 1 - O(\varepsilon)$ for all α. $\Gamma \sim_\varepsilon \{S_\alpha\}$ parallels the default logic notion of statements *belonging to the same extension.*

Definition: A statement S is said to be *ambiguous,* given Γ, if both S and its negation are permitted by Γ.

A classical example of a statement left ambiguous by Γ is the "Nixon diamond":

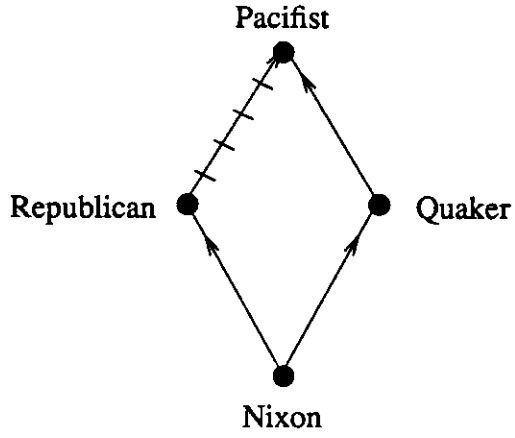

Figure 1

Here

$\Gamma = \{Quaker\ (Nixon), Republican\ (Nixon), I(Quaker, Pacifist), I(Republican, \neg Pacifist)\}$

Since $P(Pacifist \mid Republican, Quaker)$ is not constrained by either $P(Pacifist \mid Republican)$ or $P(Pacifist \mid Quaker)$, there exist a P in $\mathcal{P}_{\Gamma,\varepsilon}$ that yields $P(Pacifist(Nixon)) = 1 - \varepsilon$ and another $P' \in \mathcal{P}_{\Gamma,\varepsilon}$ yielding $P'(Pacifist(Nixon)) = \varepsilon$, thus rendering the statement $Pacifist(Nixon)$ ambiguous.

This example also demonstrates the importance of representing exceptions by keeping ε small but positive. Were we to treat Γ as a set of categorical statements, a contradiction would have resulted, from which any conclusion whatsoever could be derived. The ε-semantic, on the other hand, treats the conflict between Republicanism and Quakerism as a local ambiguity rather than a contradiction; humbly indicating the need for additional information regarding properties of Republican-Quakers, but making no claims regarding conclusions which do not critically depend on this specific information.

However, the real power of ε-semantics lies in cases where the machinery of probability calculus can be harnessed to *resolve* conflicts of property inheritance. A classical example of such cases is represented by the "Penguin triangle" of Figure 2. Here Γ comprises the sentences:

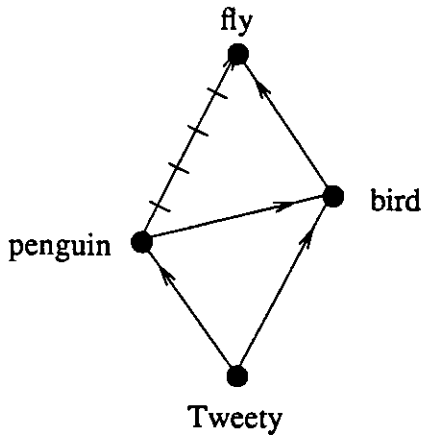

Figure 2

$$\Gamma = \{penguin(Tweety), bird(Tweety), I(penguin, \neg fly), I(bird, fly), I(penguin, bird)\}$$

It is similar in structure to the Nixon diamond except for the extra link between "penguin" and "bird," indicating that penguins are a subclass of birds.

Early inheritance systems (e.g., FRL [Roberts and Goldstein, 1978] and NETL [Fahlman, 1979]) have resolved the conflict between $I(penguin, \neg fly)$ and $\{I(penguin, bird), I(bird, fly)\}$ by appealing to the "shortest path" criterion, which correctly prefers the direct conclusion $I(penguin, \neg fly)$ over the inferred sentence $I(penguin, fly)$. However, as observed by Touretzky [Touretzky, 1984], the "shortest path" criterion does not always provide the desired preference of more specific defaults over less specific defaults. For example, Tweety inherits both the default "*fly*" from "*bird*" and "$\neg$*fly*" from "*penguin*" along paths of equal length. We shall now demonstrate how the desired conclusions follow directly from the ε-semantics attributed to Γ, making no reference to topological considerations neither to intuition about subclass specificity.

Our problem is to determine the range of $P[fly(Tweety) \mid bird(Tweety), penguin(Tweety)]$ permitted by the sentences in Γ or, more abstractly, we wish to examine the degree to which the probability $P(f \mid b, p)$ is constrained by the inputs:

$$P(f \mid b) = 1 - \varepsilon, \quad P(f \mid p) = \varepsilon, \quad P(b \mid p) = 1 - \varepsilon.$$

Conditioning $P(f \mid p)$ on both b and $\neg b$, one obtains

$$P(f \mid p) = P(f \mid p, b)\, P(b \mid p) + P(f \mid p, \neg b)\, [1 - P(b \mid p)]$$

$$\geq P(f \mid p, b)\, P(b \mid p)$$

Thus,

$$P(f \mid p, b) \leq \frac{P(f \mid p)}{P(b \mid p)} = \frac{\varepsilon}{1 - \varepsilon} = O(\varepsilon)$$

and

$$P(\neg f \mid p, b) = 1 - O(\varepsilon).$$

We see that the conclusion $\neg fly(Tweety)$ can be issued with almost certainty, i.e., $\Gamma \models_{\varepsilon} I(Tweety, \neg fly)$, even if penguins are not a strict subclass of birds. All that is required is the probabilistic condition $P(b \mid p) = 1 - \varepsilon$ which is secured by the sentence $I(penguin, bird)$, meaning that exceptions in the form of non-bird penguins are rather rare.

This is a slight generalization of a well known result in probability theory stating that, while $P(x \mid y, z)$ is, in general, unconstrained by $P(x \mid y)$ and $P(x \mid z)$, the one exception is when one of the conditioning arguments subsumes the other, say $y \rightarrow z$, in which case

$P(x \mid y, z) = P(x \mid y)$. Translated to the graphical descriptions of Figures 1 and 2, this result states that ambiguities among conflicting defaults can be resolved if a direct arrow exists between the tails of the corresponding conflicting arrows. Whenever such an arrow exists (e.g., *penguin* $\rightarrow$ *bird* in Figure 2), ambiguities are resolved in favor of the property labeling the tail of the arrow (e.g., "*penguin*" in Figure 2). Whenever the two tails, b and p, are not connected directly, the ambiguity can be resolved if the required condition $\Gamma \models_\varepsilon I(p, b)$, (or, alternatively, $\Gamma \models_\varepsilon I(b, p)$), can be inferred from indirect paths between b and p, applying the criterion recursively. Otherwise, if neither $\Gamma \models_\varepsilon I(p, b)$ nor $\Gamma \models_\varepsilon I(b, p)$ can be derived, ambiguity remains, while if both prevail, Γ is inconsistent.

This criterion constitutes the probabilistic basis of Touretzky's "inheritance distance" [Touretzky, 1984] and the recent "skeptical" algorithm for inheritance reasoning by Horty, Thomasson and Touretzky [1987] which rectifies the deficiencies of the "shortest path" heuristic. The probabilistic justification of this criterion renders a refined version of the skeptical algorithm *sound* relative to the ε-semantic introduced. A detailed description and a soundness proof of this algorithm will be given elsewhere [Pearl, in preparation].

It is not hard to show that the network of Figure 2 yields another plausible conclusion, $I(bird, \neg penguin)$, stating that when one talks about birds one does not have penguins in mind, i.e., penguins are exceptional kind of birds. It is a valid conclusion of Γ because every P in $\mathcal{P}_{\Gamma, \varepsilon}$ must yield $P(p \mid b) = O(\varepsilon)$. Of course, if the statement $I(bird, penguin)$ is artificially added to Γ, inconsistency results; as ε diminishes below a certain level (1/3 in our case), $\mathcal{P}_{\Gamma, \varepsilon}$ becomes empty. It can be shown that if Γ is acyclic and all arrows emanate from positive properties (i.e., precluding $I(-p, q)$), then Γ is consistent iff it does not contain coflicting pairs

$\{I(p, q)\ \&\ I(p, -q)\}$. Algorithms for testing consistency in general inheritance networks will be discussed elsewhere.

The Principle of Mediated Inheritance

The ε-semantics defined above is sufficient to guarantee that any issued conclusion is truly dictated by Γ. However, it does not capture *all* the assumptions people make in normal discourse. It turns out that probability theory permits such a rich set of distributions in each $\mathcal{P}_{\Gamma, \varepsilon}$, that many expected conclusions would cautiously be proclaimed ''ambiguous'' by the system describe thus far. For example, consider the statements ''birds are winged-animals'' and ''winged-animals fly'' encoded as

$$\Gamma = \{I(bird, WA), I(WA, fly)\}. \tag{2}$$

In ordinary discourse we would expect to draw the plausible conclusion $S = I(bird, fly)$, yet, the ε-semantic defined in (1) would not sanction S as a legitimate conclusion of Γ. The reason is that $\mathcal{P}_{\Gamma, \varepsilon}$ as defined in (1) also contains a distribution yielding $P(fly \mid bird) = \varepsilon$, just in case birdness constitutes an impediment to flying. That Γ *should* contain such a distribution is clearly seen by replacing ''bird'' with ''penguin''; a world in which penguin constitutes an exception to flying and in which

$$\Gamma' = \{I(penguin, WA)\, I(WA, fly)\} \tag{3}$$

holds, certainly exists. Indeed, we cannot expect a reasoning system to issue the conclusions $\Gamma \models_{\varepsilon} I(bird, fly)$ and $\Gamma' \models_{\varepsilon} I(penguin, \neg fly)$ unless additional information is supplied regarding birds, penguins and their flying abilities. In the case of penguins, we expect the knowledge base

to contain an explicit statement making flying penguins an exception, i.e., $I(penguin, \neg fly)$ as in Figure 2, while in the case of birds we expect to use a default assumption that, unless stated otherwise, birds should inherit the property "fly" via the intermediate predicate "*WA*". This assumption, which might be termed "mediated inheritance," corresponds to the celebrated probabilistic assumption of conditional independence and is best described using graphical terminology. (Pearl, 1986). If we map the properties and statements in Γ, respectively, to the vertices and arcs of some network, then the mediated inheritance assumption can be formulated by requiring that, in addition to satisfying (1), every $P \in \mathcal{P}_{\Gamma,\varepsilon}$ should also be a *Markov field* relative to Γ.

Definition: P is said to be a *Markov field* relative to Γ iff whenever Z is a set of vertices (predicates) separating p from q in Γ then

$$P(q \mid p, Z) = P(q \mid Z) \tag{4}$$

For example, the network corresponding to Γ in (2) is shown in Figure 3

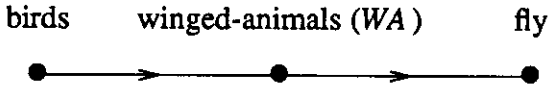

Figure 3

and, since $Z = \{WA\}$ separates "fly" from "birds," (4) translates to

$$P(fly(x) \mid WA(x), bird(x)) = P(fly(x) \mid WA(x)).$$

The meaning of such assumption is fairly clear; the flying properties of an individual x, known to be both a bird and a winged animal are solely determined by the flying property of the mediating class "winged animals". The same argument should apply to Γ' in (3) except that in this

case we expect the exceptional feature of penguins to be captured by an explicit statement $I(penguin, \neg fly)$, yielding the network Γ'' of Figure 4,

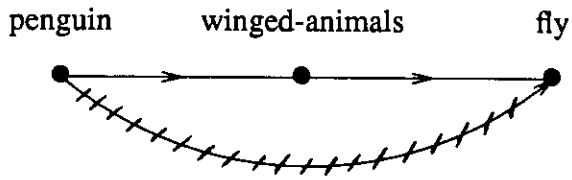

Figure 4

where "penguin" is no longer separated from "fly".

It is easy to show that if $\mathcal{P}_{\Gamma,\varepsilon}$ is further restricted by the assumption of mediated inheritance then the networks corresponding to Γ and Γ'' yield the expected conclusions, i.e., $\Gamma \models_{\varepsilon} I(bird, fly)$ and $\Gamma'' \models_{\varepsilon} I(penguin, \neg fly)$.

The derivation of $\Gamma \models_{\varepsilon} I(bird, fly)$ is as follow:

$$P(f \mid b) = P(f \mid b, WA)\,P(WA \mid b) + P(f \mid b, \neg WA)\,P(\neg WA \mid b)$$

$$= P(f \mid WA)\,P(WA \mid b) + P(f \mid b, \neg WA)\,P(\neg WA \mid b)$$

$$= (1-\varepsilon)(1-\varepsilon) + P(f \mid b, \neg WA)\,\varepsilon$$

$$= 1 - O(\varepsilon)$$

The penguin triangle still yields the same derivation of $P(fly \mid penguin) = \varepsilon$ as before because no subset of vertices (e.g., WA) separates "fly" from "penguin". The Nixon diamond, on the other hand, will remain ambiguous because, although the set $Z = \{Quaker, Republican\}$ separates "*Nixon*" from "*Pacifist*," $P\{Pacifist \mid Quaker, Republican\}$ remains unconstrained

by $P(Pacifist \mid Quaker) = 1 - \varepsilon$ and $P(Pacifist \mid Republican) = \varepsilon$.

Formulating the assumption of mediated dependency in probabilistic terms endows the topology of inheritance networks with meaningful semantics, open to public discussion and scrutiny. In particular, it clearly highlights the significance of links *missing* from these networks and it explains why, contrary to logical deduction, induced links ought to be treated differently than those found originally in Γ [Sandewall, 1986]. The next subsection unleashes the detective powers of probability theory to uncover another tacit assumption underlying inheritance reasoning.

The Principle of Positive Conjunction

The two assumptions introduced so far, ε-semantics and mediated inheritance, still lack one ingredient necessary for producing all the plausible conclusions we desire. Assume we have two positive paths leading from p to q via the intermediate nodes r and s, as in Figure 5.

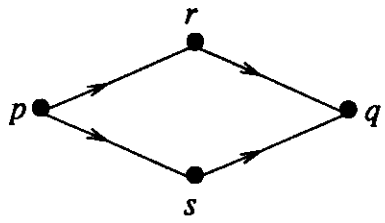

Figure 5

The assumption of mediated inheritance dictates

$$P(q \mid p) = P(q \mid r, s, q)\, P(r, s \mid q) + O(\varepsilon)$$

$$= P(q \mid r, s)(1 - \varepsilon) + O(\varepsilon)$$

Yet, since bare probability theory imposes no restriction on $P(q \mid r, s)$ given $P(q \mid r) = 1 - \varepsilon$ and $P(q \mid s) = 1 - \varepsilon$, $\mathcal{P}_{\Gamma, \varepsilon}$ may contain a P with arbitrary small $P(q \mid r, s)$. Thus, we face a paradoxical situation where the presence of multiple inheritance paths between p and q, instead of reinforcing the natural conclusion $I(p, q)$, actually causes us to reserve judgement.

This skeptical behavior of probability theory is not totally without reason. While it appears paradoxical in the interpretation $p = birds, r = have\text{-}wings, s = have\text{-}feathers, q = fly$, it is certainly justified in the interpretation: $p = any\ man, r = who\ marries\ Ann$, $s = who\ marries\ Sue, q = will\ be\ happy$ since bigamy is occasionally regarded as an impediment to happiness. Nevertheless, since such cancellation effects are relatively rare, it makes sense to institute the following default principle: In the absence of information to the contrary, assume

$$I(r, q) \in \Gamma \ \&\ I(s, q) \in \Gamma \Longrightarrow P(q \mid r, s) = 1 - \varepsilon \tag{5}$$

This principle is tacitly assumed by all systems of multiple inheritance [e.g., Touretzky (1986)] by permitting two non-preempted paths from p to q to sanction the conclusion $I(p, q)$. It reflects the attitude that cancellation is a rare occurrence. Moreover, the diagrammatic language of inheritance networks does not permit one to express the existence of cancellation (e.g., the bigamy example); hypernetwork are needed for that purpose, corresponding to semi-normal default rules in default logic [Etherington and Reiter, 1983].

We now combine the 3-principles above by stating a weaker condition for a statement to qualify as a conclusion of Γ.

Definition: A statement $S = I(p, q)$ is said to be a *plausible-conclusion* of Γ, written $\Gamma \models_{\varepsilon} S$, if $P(S) = 1 - O(\varepsilon)$ for every $P \in \mathcal{P}_{\Gamma, \varepsilon}$, where:

$$\mathcal{P}_{\Gamma, \varepsilon} = \left\{ P : P(v \mid u) = \begin{cases} 1-\varepsilon & \text{if } I(u, v) \in \Gamma \\ \varepsilon & \text{if } I(u, \neg v) \in \Gamma \end{cases} , \right.$$

P is Markov relative to Γ, and

$$\left. I(u, w) \in \Gamma \ \& \ I(v, w) \in \Gamma \Longrightarrow P(w \mid u, v) = 1 - \varepsilon \right\} \tag{6}$$

Identical conditions for $\mathcal{P}_{\Gamma, \varepsilon}$ should be used in the definitions of permitted and ambiguous conclusions.

Discussion

Inheritance hierarchies represent one of the simplest form of nonmonotonic reasoning and yet, as Sandewall [1986] has observed, "the combined structure, multiple inheritance with exceptions, offers a number of unpleasant and challenging surprises". Research in the past ten years has been guided by a collection of clever, intuition-loaded examples and has led to the development of algorithms that cover, more or less, the examples accumulated. In the absence of a more global guiding principle, Sandewall goes as far as proposing "that we consider such

collections of structure types *as* the definition of the semantics, for the time being."

One result of lacking a more principled semantics is that it took over five years [Touretzky, 1984] to discover that the "shortest-distance" heuristic used by earlier systems [Roberts et al. (1977), Fahlman (1979)] occasionally produce implausible conclusions. Currently available remedies are still incapable of distinguishing some ambiguous conclusions from plausible ones. For example, the "skeptical" algorithm of Horty et al. [1987], would issue statements predicated on Nixon's not being a pacifist (Figure 1) with the same conviction as those predicated on penguins not flying (Figure 2).

Can the more powerful nonmonotonic logics, such as circumscription and default theories, be of assistance to inheritance hierarchies? This prospect seems to be hindered by two hurdles. First, available nonmonotonic logics have semantic problems of their own. They capture a person's intuition about how he/she is disposed to react to any local chunk of information but do not guarantee that the sum total of these dispositions would lead to desirable, formally specified net results. Second, they do not incorporate the implicit assumptions underlying the unique structure of inheritance systems as built-in features of the logics. For example, formulating inheritance systems in semi-normal default rules [Etherington and Reiter, 1983] requires labeling each default rule with the names of all its exceptions. The natural rule $I(bird, fly)$ of Figure 2 should be written as

$$\frac{bird(x):\ fly(x)\ \&\ \neg penguin(x)}{fly(x)}$$

meaning, if x is a bird then, unless it leads to a contradiction to assume that x flies and that x is not a penguin, asserts that x flies. As Touretzky [1984] and Sandewall [1986] pointed out, the

need to write exceptional cases explicitly into the inference rules that may be affected by them, is very impractical; the very point with non-monotonic reasoning and exception links is that we should not have to perform that chore.

Formulating inheritance hierarchies in normal default theory [Etherington, 1987] properly handles implicit exceptions but, since the theory generates multiple extensions, one must appeal to Touretzky's "inferential distance" in order to sort plausible conclusions [e.g., $I(penguin, \neg fly)$, Figure 2] from implausible ones [e.g., $I(penguin, fly)$, Figure 2]. This still leaves plausible and some ambiguous conclusions indistinguishable, unless one is willing to enlist and intersect all "credulous" extensions.

The probabilistic semantics offered in this report promises to overcome some of the difficulties mentioned. First, it bases its decisions on denotational rather than operational semantics. Second, plausibility criteria are formally defined and are not subject to subjective disputation. Third, it explicates the assumptions underlying inheritance reasoning and renders them empirically testable. Fourth, it leads to derivational algorithms whose correctness can be verified formally.

Some readers may object to the very idea of basing common-sense reasoning on probability calculus. The usual argument is that "typicality" has nothing to do with frequency of occurrence, it is more of a mental disposition along the line of "In the absence of any information to the contrary, assume the form ... " [Brachman, 1985]. Thus, why should dispositions be combined like frequencies?

True, if one is utterly determined to confine oneself to covert mental dispositions, oblivion to how they come about, one is welcome to treat the sentence "$P(fly \mid bird) = 1 - \varepsilon$" just as such, meaning that in the absence of information to the contrary, Tweety's birdness evokes readiness to presume it can fly. Fortunately, the sentence also offers us the option of occasionally going one step further and relating the readiness evoked to our external experience, confirming that, indeed, most birds do fly.

In other words, what probability theory offers, that alternative logics of mental dispositions have so far ignored, is to connect such dispositions to their experiential origin (e.g., frequency of events) and, more importantly, to propose a calculus of dispositions that mirrors the features of that origin. Since the latter is open, well understood, coherent and free of contradictions and/or surprises, the hope is that conclusions drawn from such calculus will follow suit.

It is true that "typical" is not the same as "usual" and that the two are different than "likely". But the differences do not mask the overriding commonality of these quantifiers. Consider, for example, the four sentences:

1. Elephants typically have trunks ($e \rightarrow t$)

2. Trunk-animals normally love honey ($t \rightarrow h$)

3. Elephants usually hate honey ($e \rightarrow \neg h$)

4. Most trunk-animals are elephants ($t \rightarrow e$)

No matter how one chooses to distinguish the "most" from the "typical" and the "usual" from the "normal," intuition dictates that the first three statements convey an exception while the

four convey a contradiction. The fact that probabilistic semantics, unlike other nonmonotonic logics, formally captures this intuitively sound distinction, demonstrates that even simplistic reliance on frequency interpretation might some time payoff. Additional payoffs lie in providing clear formal guidance as to what conclusions we wish to draw and whether the algorithms proposed deliver the results expected.

It is hoped, therefore, that the probabilistic semantics proposed in this report will play a useful role in the development of common-sense reasoning systems.

Acknowledgment

The possibility that non-monotonic reasoning can serve "as a very streamlined expression of probabilistic information when numerical probabilities, especially conditional probabilities, are unobtainable" was recognized by McCarthy [1984] who also suggested infinitesimal probabilistic interpretation to circumscription. Preliminary attempts along such lines are reported in [Rich, 1983].

I am grateful to Hector Geffner for calling my attention to the current state of inheritance systems, for many stimulating discussions and for pointing out the need of adopting positive-conjunction as a separate default principle. I also appreciate the discussions I had with Danny Bobrow, Ben Grossof, Vladimir Lifschitz and Dave Touretzky at the AAAI-87 conference, in Seattle. And of course, Jackie Trang for making all the penguins, the Nixons and the birds fly in the right direction.

References

[Brachman, 1985] Brachman, R. J., "I Lied About The Trees," or, "Defaults and Definitions in Knowledge Representation," *AI Magazine,* Vol. 6, No. 3, 1985, pp. 80-93.

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

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

[Fahlman, 1979] Fahlman, S.E. *NETL: A System for Representing and Using Real-World Knowledge.* MIT Press, Cambridge, Massachusetts, 1979.

[Geffner, 1987] Geffner, H., "Sound Defeasible Inference," *Technical Report R-94,* Cognitive Systems Laboratory, UCLA. In preparation.

[Horty *et al.,* 1987] J. Horty, R. Thomason, and D. Touretzky. "A Skeptical Theory of Inheritance in Nonmonotonic Semantic Networks." *Proceedings, AAAI-87,* Seattle, Washington 1987, pp. 358-363.

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

[Pearl, 1986] Pearl, J., "Markov and Bayes Networks: a Comparison of Two Graphical Representations of Probabilistic Knowledge." UCLA Cognitive Systems Laboratory *Technical Report (R-46),* October, 1986.

[Rich, 1983] Rich, E., "Default Reasoning as Likelihood Reasoning." *Proceedings, IJCAI-83,* 1983, pp. 348-351.

[Roberts *et al.,* 1977] Roberts, R., and Goldstein, I. *The FRL Manual.* MIT AI Memo 409, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, September, 1977.

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

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

[Touretzky, 1986] Touretzky, D.S., *The Mathematics of Inheritance Systems*. Morgan Kaufmann, Los Alton, California, 1986.