R-128
March 1992

Philosophy and AI
Essays at the Interface

edited by Robert Cummins and John Pollock

Judea Pearl

Probabilistic Semantics for Nonmonotonic Reasoning

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

8 Probabilistic Semantics for Nonmonotonic Reasoning

Judea Pearl

8.1 Why Probabilistic Semantics? Or, Conventions versus Norms

In nonmonotonic logics, defeasible sentences are usually interpreted as conversational conventions, as opposed to descriptions of empirical reality (McCarthy 1986, Reiter 1987). For example, the sentence "Birds fly" is taken to express a communication agreement such as: "You and I agree that whenever I want you to conclude that some bird does not fly, I will say so explicitly; otherwise you can presume it does fly." Here the purpose of the agreement is not to convey information about the world but merely to guarantee that in subsequent conversations, all conclusions drawn by the informed match those intended by the informer. Once the agreement is accepted by an agent, the meaning of the sentence acquires a dispositional character: "If x is a bird and I have no reasons to presume the contrary, then I am disposed to believe that x flies." Neither of these interpretations invokes any statistical information about the percentage of birds that fly nor any probabilistic information about how strongly the agent believes that a randomly chosen bird actually flies.

However, the probabilistic statement $P[(Fly(x)|Bird(x)] = High$ (to read: "If x is a bird, then x probably flies") offers such a clear interpretation of "Birds fly", that it is hard to refrain from viewing defeasible sentences as fragments of probabilistic information, albeit subjective in nature. With such declarative statements it is easier to define how the fragments of knowledge should be put together coherently, to characterize the set of conclusions that one wishes a body of knowledge to entail, and to identify the assumptions that give rise to undesirable conclusions, if any.

The reasons are several. First, semantics has traditionally been defined as a relation between the speaking agent and entities external to the agent. Probabilistic information is, by its very nature, a declarative summarization of constraints in a world external to the speaker. As such, it is empirically testable (at least in principle), it is often shared by many agents, and conclusions are less subject to dispute. Second, in many cases, it is the transference of probabilistic knowledge that is the ultimate aim of common conversations, not the speaker's pattern of dispositions (which are often arbitrary). In such cases, the empirical facts that caused the agent to commit to a given pattern of dispositions are more important than the dispositions themselves, because it is those empirical facts that the listening agent is about to confront in the future. Finally, being a centuries-old science, the study of probabilistic inference has accumulated a wealth of theoretical results that provide shortcuts between the semantics and the intended conclusions. This facilitates quick generation of meaningful examples and counterexamples, quick proofs of necessity or impossibility, and thus, effective communication among researchers.

But even taking the extreme position that the only purpose of default statements is to establish conversational conventions, probabilists nevertheless believe that, while we are in the process of uncovering and formulating those conventions, we cannot totally ignore their empirical origin. Doing so would resemble the hopeless task of formulating qualitative physics in total ignorance of the quantitative laws of physics, or, to use a different metaphor, designing speech-recognition systems oblivious to the laws of phonetics.

The quest for probabilistic semantics is motivated by the assumption that the conventions of discourse are not totally arbitrary, but rather, respect certain universal norms of coherence, norms that reflect the empirical origin of these conventions. Probabilistic semantics, by summarizing the reality that compelled the choice of certain conventions over others, should be capable of revealing these norms. Such norms should tell us, for example, when one convention is incompatible with another, or when one convention should be a natural consequence of another, examples of both will be illustrated in section 8.4.

The benefits of adopting probabilistic norms apply not only to syntactical approaches to nonmonotonic reasoning, but also to semantical approaches, such as those based on model preference (McCarthy 1980). Inferences based on model preference are much less disciplined than those based on probability, because the preferences induced by the various sentences in the knowledge base are not constrained a priori, and can, in general, be totally whimsical. Indeed, such a wide range of approaches to nonmonotonic reasoning can be formulated as variants of preference-based semantics (Shoham 1987), that highly sophisticated restrictions must be devised to bring the ensuing inferences in line with basic standards of rationality (Lehmann and Magidor 1988) (see section 8.6.2).

8.2 Nonmonotonic Reasoning Viewed as Qualitative Probabilistic Reasoning

To those trained in traditional logics, symbolic reasoning is the standard, and nonmonotonicity a novelty. To students of probability, on the other hand, it is symbolic reasoning that is novel, not nonmonotonicity. Dealing with new facts that cause probabilities to change abruptly from very high values to very low values is a commonplace phenomenon in almost every probability exercise and. naturally, has never attracted special attention among probabilists. The new challenge for probabilists is to find ways of abstracting out the numerical character of high and low probabilities, and cast them in linguistic terms that reflect the natural process of accepting and retracting beliefs. Thus. while nonmonotonic reasoning is commonly viewed as an extension to standard logic, it can also be viewed as an exercise in qualitative probability, much like physicists view current AI research in naive physics.

In research on qualitative physics, it is customary to discretize and abstract real quantities around a few "landmark" values (Kuipers 1986). For example, the value 0 defines the abstraction: positive, negative and zero. In probability, the obvious landmarks are $\{0, \frac{1}{2}, 1\}$, where 0 and 1 represent $FALSE$ and $TRUE$, respectively, and $\frac{1}{2}$ represents the neutral state of total ignorance. However, direct qualitative reasoning about $\{0, 1\}$ reduces to propositional logic, while reasoning with the intervals $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$ is extremely difficult—to process pieces of evidence properly and determine if a given probability should fall above $\frac{1}{2}$ requires almost the full power of numerical probability calculus (Bacchus 1988).

Following the tradition of qualitative reasoning in physics and mathematics, two avenues are still available for qualitative analysis:

1. "Perturbation" analysis, to determine the direction of CHANGE induced in the probability of one proposition as a result of learning the truth of another, and

2. An "order-of-magnitude" analysis of proximities to the landmark values.

The first approach has been pursued by Wellman (1987) and Neufeld and Poole (1988), and the second by Adams (1975), Spohn (1988), Pearl (1988), and Geffner (1988, 1989).

8.2.1 Perturbation Analyses

Both Wellman (1987) and Neufeld and Poole (1988) investigated the logic behind the qualitative relation of *influence* or *support*, namely, the condition under which the truth of one proposition would yield an increase in the probability of another. Wellman's analysis focuses on variables with ordered domains (e.g., "An increase in quantity a is likely to cause an increase in quantity b") as a means of providing qualitative aids to decisions, planning, and diagnosis. Neufeld and Poole, focused on the relation of *confirmation* between propositions (e.g., *Quaker(Nixon)* adds confirmation to *Pacifist(Nixon)*), and viewed this relation as an important component of nonmonotonic reasoning.

Both approaches make heavy use of conditional independence and its graphical representation in the form of Bayesian networks (Pearl 1988). The reason is that, if we define the relation "A supports B" (denoted $S(A, B)$) as

$$S(A, B) \quad \text{iff} \quad \mathrm{P(B \mid A) \geq P(B)}, \tag{1}$$

then this definition in itself is too weak to yield interesting inferences. For example, whereas we can easily show symmetry $S(A, B) \Leftrightarrow S(B, A)$ and contraposition $S(A, B) \Leftrightarrow S(\neg B, \neg A)$, we cannot conclude cumulativity (i.e. that $S(A \wedge B, C)$ follows from $S(A, B)$ and $S(A, C)$), nor transitivity (i.e., that $S(A, C)$ follows from $S(A, B)$ and $S(B, C)$). For the latter to hold, we must assume that C is conditionally independent of A, given B,

$$P(C \mid A, B) = P(C \mid B),$$

namely, that knowledge of A has no influence on the probability of C, once we know B.

Conditional independence is a 3-place nonmonotonic relationship that forms a *semi-graphoid* (Pearl and Verma 1987, Pearl 1988). Semi-graphoids are structures that share some properties of graphs (hence the name) but, in general, are difficult to encode completely, in a compact way. The assumption normally made in probabilistic reasoning (as well as in most nonmonotonic logics, though not explicitly) is that if we represent dependence relationships in the form of a directed (acyclic) graph, then any link missing from the graph indicates the absence of direct dependency between the corresponding variables. For example, if we are given two defeasible rules, $a \rightarrow b$ and $b \rightarrow c$, we presume that a does not have any direct bearing on c, but rather, that c is independent of a, given the value of b. An important result from the theory of graphoids states that there is indeed a sound and complete procedure (called d-separation) of inferring conditional independencies from such a graph. However, this requires that the graph be constructed in a disciplined, stratified way: Every variable x should draw arrows from all those perceived to have *direct influence* on x, that is, those that must be known to render x independent of all its predecessors in some total order (e.g., temporal). A graph (directed and acyclic) constructed in this fashion is called a *Bayesian network* (Pearl 1988). In practice, this presumes that the knowledge provider has taken pains to identify all direct influences of each variable in the system.

Neufeld and Poole have assumed that if we take isolated default statements and assemble them to form a directed graph, the resulting graph would display all the dependencies that a Bayesian network would. Unfortunately, this is not always the case, and may lead to unsound conclusions. For example, from the defaults $A \rightarrow B$, $C \rightarrow \neg B$, we will conclude (using the d-separation criterion) that A is independent of C (since there is no active path between A and C). Often, however, two classes A and C whose members differ substantially in one typical property (B vs $\neg B$) will be found dependent on one another. The language of graphs may also be insufficient for expressing some independencies that are found useful in natural discourse. For example, having been told that A supports C and B supports C, we tend to presume that A and B also supports C. This presumption. however, is not directly expressible in the language of directed graphs, nor can it be derived from the semantics of "support" given in equation (1).

Wellman has circumvented these difficulties by starting from a well structured Bayesian network, and by defining "support" in a more restrictive way. Instead of equation (1), Wellman's definition reads:

$$S^+(a,b,G) \quad \text{iff} \quad P(B \mid A, x) \geq P(B \mid x)$$

where $S^+(a,b,G)$ stands for "a positively influences b, in the context of graph G", and the inequality should hold for every valuation x of the direct predecessors of b (in G). This stronger definition of support defines, in fact, the conditions under which inferences based on graphically derived dependencies are probabilistically sound. Compared with the system of Neufeld and Poole, soundness is acquired at the price of a more elaborate form of knowledge specification, namely, the structure of a Bayesian network

8.2.2 Infinitesimal Analysis

Spohn (1988) has introduced a system of belief revision (called OCF for Ordinal Conditional Functions) which requires only integer-value addition, and yet retains the notion of conditionalization, a facility that makes probability theory context dependent, hence nonmonotonic. Although Spohn has proclaimed OCF to be "nonprobabilistic," the easiest way to understand its power and limitations is to interpret OCF as an infinitesimal (i.e., nonstandard) analysis of conditional probabilities.

Imagine an ordinary probability function P defined over a set W of possible worlds (or states of the world), and let the probability $P(w)$ assigned to each world w be a polynomial function of some small positive parameter ε, for example, α, $\beta\varepsilon$, $\gamma\varepsilon^2$, and so on. Accordingly, the probabilities assigned to any subset A of W, as well as all conditional probabilities $P(A|B)$, will be rational functions of ε. Now define the OCF function $\kappa(A|B)$ as

$$\kappa(A|B) = \text{lowest } n \text{ such that } \lim_{\varepsilon\to 0} P(A|B)/\varepsilon^n \text{ is non-zero.} \tag{2}$$

In other words, $\kappa(A|B) = n$ iff $P(A|B)$ is of the same order as ε^n, or equivalently, $\kappa(A|B)$ is of the same order of magnitude as $[P(A|B)]^{-1}$.

If we think of n for which $P(w) = \varepsilon^n$ as measuring the degree to which the world w is disbelieved (or the degree of surprise were we to observe w), then $\kappa(A|B)$ can be thought of as the degree of disbelief (or surprise) in A, given that B is true. It is easy to verify that κ satisfies the following properties:

1. $\kappa(A) = \min\{\kappa(w) \mid w \in A\}$
2. $\kappa(A) = 0$ or $\kappa(\neg A) = 0$, or both
3. $\kappa(A \cup B) = \min\{\kappa(A), \kappa(B)\}$
4. $\kappa(A \cap B) = \kappa(A|B) + \kappa(B)$ (3)

These reflect the usual properties of probabilistic combinations (on a logarithmic scale) with *min* replacing addition, and addition replacing multiplication. The result is a probabilistically sound calculus, employing integer addition, for manipulating order of magnitudes of disbeliefs. For example. if we make the following correspondence between linguistic quantifiers and ε^n:

$P(A) = \varepsilon^0$ A is believable $\kappa(A) = 0$

$P(A) = \varepsilon^1$	A is unlikely	$\kappa(A) = 1$
$P(A) = \varepsilon^2$	A is very unlikely	$\kappa(A) = 2$
$P(A) = \varepsilon^3$	A is extremely unlikely	$\kappa(A) = 3$

then Spohn's system can be regarded as a nonmonotonic logic to reason about likelihood (contrast with the modal logic of Halpern and Rabin 1987). It takes sentences in the form of quantified conditional sentences, for example, "Birds are likely to fly" (written $\kappa(\neg f|b) = 1$), "Penguins are most likely birds" (written $\kappa(\neg b \mid p) = 2$), "Penguins are extremely unlikely to fly," (written $\kappa(f|p) = 3$), and returns quantified conclusions in the form of "If x is a penguin-bird then x is extremely unlikely to fly" (written $\kappa(f|p \wedge b) = 3$).

The weakness of Spohn's system, shared by numerical probability, is that it requires the complete specification of a distribution function before reasoning can commence. In other words, we must specify the κ associated with every world w. In practice, of course, such specification need not be enumerative, but can use the decomposition facilities provided by Bayesian networks. However, this too might require knowledge that is not readily available in common discourse. For example, we might be given the information that birds fly (written $\kappa(\neg f|b) = 1$) and no information at all about properties of non-birds, thus leaving $\kappa(f \wedge \neg b)$ unspecified. Hence, inferencial machinery is required for drawing conclusions from partially specified models, like those associating a κ with isolated default statements. Such machinery is provided by the conditional logic of Adams (1975), to be discussed next.

Adams's logic can be regarded a bi-valued infinitesimal analysis, with input sentences specifying κ values of only 0 and 1, corresponding to "likely" and "unlikely" rankings. However, instead of insisting on a complete specification of $\kappa(w)$, the logic admits fragmentary sets of conditional sentences, treats them as constraints over the distribution of κ, and infers only such statements that are compelled to acquire high likelihood in every distribution $\kappa(w)$ satisfying these constraints.

Because of its importance as a bridge between probabilistic and logical approaches, we will provide a more complete introduction to Adams's logic, using excerpts from chapter 10 of Pearl 1988. We will see that the semantics of infinitesimal probabilities (called ε-semantics in Pearl 1988) leads to a two-level architecture for nonmonotonic reasoning:

1. A conservative, consistency-preserving *core*, embodied in a semimonotonic logic, which derives only conclusions that are safe relative to the addition of new domain knowledge.

2. An adventurous *shell*, sanctioning a larger body of less grounded inferences. These inferences reflect probabilistic independencies that are not explicit in the input, yet, based on familiar patterns of discourse, are presumed to hold in the absence of explicit dependencies.

8.3 The Conservative Core

8.3.1 ε-Semantics

We consider a default theory $T = \langle F, \Delta \rangle$ in the form of a database containing two types of sentences: factual sentences (F) and default statements (Δ). The factual sentences describe findings or observations specific to a given object or a situations; for example, $p(a)$ asserts that individual a has the property p. The default statements are of the type "p's are typically q's", written $p(x) \rightarrow q(x)$ or simply $p \rightarrow q$, which is short for saying "any individual x having property p typically has property q". The properties $p, q, r \ldots$ can be compound boolean formulas of some atomic predicates $p_1, p_2, \ldots p_n$, with x as their only free variable. However, no ground defaults (e.g., $p(a) \rightarrow q(a)$) are allowed in F and no compound defaults (e.g., $p \rightarrow (q \rightarrow r)$) are allowed in Δ. The default statement S': $p \rightarrow \neg q$ will be called the *denial* of S: $p \rightarrow q$.

Nondefeasible generic statements such as "all birds are animals" can be written $Birds(x) \wedge \neg Animal(x) \rightarrow FALSE$. This facilitates the desirable distinction between a generic law-like rule "all p's are q's" (to be encoded in Δ as $p \wedge \neg q \rightarrow FALSE$) and a factual observation $p(a) \supset q(a)$, which must enter F as $\neg p \vee q$. Indeed, the theory $\langle \{p(a)\}, \{p \wedge \neg q \rightarrow FALSE\} \rangle$, will give rise to totally different conclusions (about a) than $\langle \{p(a), \neg p(a) \vee q(a)\}, \{\ \} \rangle$, in conformity with common use of conditionals. A more natural treatment of nondefeasible conditionals is given in (Goldszmidt and Pearl 1989) where a new connective $\Rightarrow$ is used to retain the law-like character of these statements. However, to simplify our discussion we will henceforth assume that all statements are defeasible.

Let L be the language of propositional formulas, and let a *truth-valuation* for L be a function t that maps the sentences in L to the set $\{1, 0\}$, (1 for $TRUE$ and 0 for $FALSE$) such that t respects the usual Boolean connectives. To define a probability assignment over the sentences in L, we regard each truth valuation t as a world w and define $P(w)$ such that $\sum_w P(w) = 1$. This assigns a probability measure to each sentence s of L via $P(s) = \sum_w P(w)\, w(s)$.

We now interpret Δ as a set of restrictions on P, in the form of *extreme* conditional probabilities, infinitesimally removed from either 0 or 1. For example, the sentence $Bird(x) \rightarrow Fly(x)$ is interpreted as $P(Fly(x) | Bird(x)) \geq 1 - \varepsilon$, where ε is understood to stand for an infinitesimal quantity that can be made arbitrarily small, short of actually being zero.

The conclusions we wish to draw from a theory $T = \langle F, \Delta \rangle$ are, likewise, formulas in L that, given the input facts F and the restrictions Δ, are forced to acquire extreme high probabilities. In particular, a propositional formula r would qualify as a *plausible conclusion* of T, written $F \mid\!\sim_\Delta r$, whenever the restrictions of Δ force P to satisfy $\lim_{\varepsilon \to 0} P(r | F) = 1$.

It is convenient to characterize the set of conclusions sanctioned by this semantics in terms of the set of facts-conclusion pairs that are entailed by a given Δ. We call this relation *ε-entailment*[1] formally defined as follows:

Definition: Let $\mathscr{P}_{\Delta,\varepsilon}$ stand for the set of distributions licensed by Δ for any given ε, i.e.,

$$P_{\Delta,\varepsilon} = \{P : P(v | u) \geq 1 - \varepsilon \text{ and } P(u) > 0 \text{ whenever } u \rightarrow v \in \Delta\} \tag{4}$$

A conditional statement $S: p \rightarrow q$ is said to be *ε-entailed* by Δ, if every distribution $P \in \mathscr{P}_{\Delta,\varepsilon}$ satisfies $P(q|p) = 1 - O(\varepsilon)$, (i.e., for every $\delta > 0$ there exists a $\varepsilon > 0$ such that every $P \in \mathscr{P}_{\Delta,\varepsilon}$ would satisfy $P(q|p) \geq 1 - \delta$).

In essence, this definition guarantees that an ε-entailed statement S is rendered highly probable whenever all the defaults in Δ are highly probable. The connection between ε-entailment and plausible conclusions, is simply:

$$F \mid\!\sim_\Delta r \text{ iff } (F \rightarrow r) \text{ is } \varepsilon\text{-entailed by } \Delta \tag{5}$$

8.3.2 Axiomatic Characterization

The conditional logic developed by Adams (1975) faithfully represents this semantics by qualitative inference rules, thus facilitating the derivation of new sound sentences by direct symbolic manipulations on Δ. The essence of Adams's logic is summarized in the following theorem, restated for default theories in (Geffner and Pearl 1988).

[1] Adams (1975) named this p-entailment. However, ε-entailment better serves to distinguish this from weaker forms of probabilistic entailment, section 4.

THEOREM 1: Let $T = \langle F, \Delta \rangle$ be a default theory where F is a set of ground proposition formulas and Δ is a set of default rules. r is a plausible conclusion of F in the context of Δ, written $F \mathrel{|\!\!\sim}_\Delta r$, iff r is derivable from F using the following rules of inference:

Rule 1 (Defaults) $(p \rightarrow q) \in \Delta \Rightarrow p \mathrel{|\!\!\sim}_\Delta q$

Rule 2 (Logic Theorems) $p \mathrel{|\!\!\sim} q \Rightarrow p \mathrel{|\!\!\sim}_\Delta q$

Rule 3 (Cumulativity) $p \mathrel{|\!\!\sim}_\Delta q, p \mathrel{|\!\!\sim}_\Delta r \Rightarrow (p \wedge q) \mathrel{|\!\!\sim}_\Delta r$

Rule 4 (Contraction) $p \mathrel{|\!\!\sim}_\Delta q, (p \wedge q) \mathrel{|\!\!\sim}_\Delta r \Rightarrow p \mathrel{|\!\!\sim}_\Delta r$

Rule 5 (Disjunction) $p \mathrel{|\!\!\sim}_\Delta r, q \mathrel{|\!\!\sim}_\Delta r \Rightarrow (p \vee q) \mathrel{|\!\!\sim}_\Delta r$

Rule 1 permits us to conclude the consequent of a default when its antecedent is all that has been learned and this permission is granted regardless of the content of Δ. Rule 2 states that theorems that logically follow from a set of formulas can be concluded in any theory containing those formulas. Rule 3 (called *triangularity* in Pearl 1988 and *cautious monotony* in Lehmann and Magidor 1988) permits the attachment of any established conclusion (q) to the current set of facts (p), without affecting the status of any other derived conclusion (r). Rule 4 says that any conclusion (r) that follows from a fact set (p) augmented by a derived conclusion (q) also follows from the original fact set alone. Finally, rule 5 says that a conclusion that follows from two facts also follows from their disjunction.

Some Meta-Theorems:

T-1 (Logical Closure) $p \mathrel{|\!\!\sim}_\Delta q, p \wedge q \mathrel{|\!\!\sim} r \Rightarrow p \mathrel{|\!\!\sim}_\Delta r$

T-2 (Equivalent Contexts) $p \equiv q, p \mathrel{|\!\!\sim}_\Delta r \Rightarrow q \mathrel{|\!\!\sim}_\Delta r$

T-3 (Exceptions) $p \wedge q \mathrel{|\!\!\sim}_\Delta r, p \mathrel{|\!\!\sim}_\Delta \neg r \Rightarrow p \mathrel{|\!\!\sim}_\Delta \neg q$

T-4 (Right Conjunction) $p \mathrel{|\!\!\sim}_\Delta r, p \mathrel{|\!\!\sim}_\Delta q \Rightarrow p \mathrel{|\!\!\sim}_\Delta q \wedge r$

Some Non-Theorems:

(Transitivity) $p \supset q, q \mathrel{|\!\!\sim}_\Delta r \Rightarrow p \mathrel{|\!\!\sim}_\Delta r$

(Left Conjunction) $p \mathrel{|\!\!\sim}_\Delta r, q \mathrel{|\!\!\sim}_\Delta r \Rightarrow p \wedge q \mathrel{|\!\!\sim}_\Delta r$

(Contraposition) $p \mathrel{|\!\!\sim}_\Delta r \Rightarrow \neg r \mathrel{|\!\!\sim}_\Delta \neg p$

(Rational Monotony)

$$p \mathrel{|\!\!\sim}_\Delta r, \text{NOT}(p \mathrel{|\!\!\sim}_\Delta \neg q) \Rightarrow p \wedge q \mathrel{|\!\!\sim}_\Delta r \qquad (6)$$

This last property (similar to CV of conditional logic) has one of its antecedents negated, hence, its consequences cannot be derived from Δ using the five rules above. It is, nevertheless, a desirable feature of a consequence relation, and was proposed by Makinson as a standard for nonmonotonic logics (Lehmann and Magidor 1988). Rational monotony can be restored within ε-semantics if we limit out attention to families of distributions P_ε that are parameterized by ε and are analytic in ε (Goldszmidt et al 1990). Alternatively, rational monotony obtains if we interpret ε as a nonstandard infinitesimal (in a non-standard analysis), which also amounts to interpreting $p \rightarrow q$ as an OCF constraint $\kappa(q|p) < \kappa(\neg q|p)$ (see section 8.4.1 and Pearl 1990).

The reason transitivity, positive conjunction, and contraposition are not sanctioned by the ε-semantics is clear: There are contexts in which they fail. For instance, transitivity fails in the penguin example—all penguins are birds, birds typically fly, yet penguins do not. Left conjunction fails when p and q create a new condition unshared by either p or q. For example, if you marry Ann (p) you will be happy (r), if you marry Nancy (q) you will be happy as well (r), but if you marry both ($p \wedge q$), you will be miserable ($\neg r$). Contraposition fails in situations where $\neg p$ is incompatible with $\neg r$. For example, let $p \rightarrow r$ stand for *Birds* $\rightarrow$ *Fly*. Now imagine a world in which the only nonflying objects are a few sick birds. Clearly, *Bird* $\rightarrow$ *Fly* holds, yet if we observe a nonflying object we can safely conclude that it is a bird, hence $\neg r \rightarrow p$, defying contraposition.

THEOREM 2 (Δ-monotonicity): The inference system defined in theorem 1 is monotonic relative to the addition of default rules, that is,

$$\text{if } p \mathrel{|\!\sim}_{\Delta} r \text{ and } \Delta \subseteq \Delta', \text{ then } p \mathrel{|\!\sim}_{\Delta'} r \tag{7}$$

The proof follows directly from the fact that $\mathscr{P}_{\Delta',\varepsilon} \subseteq \mathscr{P}_{\Delta,\varepsilon}$ because each default statement imposes a new constraint on $\mathscr{P}_{\Delta,\varepsilon}$. Thus, the logic is nonmonotonic relative to the addition of new facts (in F) and monotonic relative to the addition of new defaults (in Δ). Full nonmonotonicity will be exhibited in section 8.4, where we consider stronger forms of entailment.

8.3.3 Consistency and Ambiguity

An important feature of the system defined by rules 1–5 is its ability to distinguish theories portraying inconsistencies (e.g., $\langle p \rightarrow q, p \rightarrow \neg q \rangle$), from those conveying ambiguity (e.g., $\langle p \wedge q, p \rightarrow r, q \rightarrow \neg r \rangle$), and those conveying exceptions (e.g., $\langle p \rightarrow q, p \wedge r \rightarrow \neg q \rangle$).

Definition: Δ is said to be *ε-consistent* if $\mathscr{P}_{\Delta,\varepsilon}$ is non-empty for every $\varepsilon > 0$, else, Δ *is ε-inconsistent*. Similarly, a set of default statements $\{S_\alpha\}$ is said to be *ε-consistent with* Δ if $\Delta \cup \{S_\alpha\}$ is ε-consistent.

Definition: A default statement S is said to be *ambiguous*, given Δ, if both S and its denial are consistent with Δ.

THEOREM 3 (Adams 1975): If Δ is ε-consistent, then a statement $S: p \to q$ is ε-entailed by Δ iff its denial $S': p \to \neg q$ is ε-inconsistent with Δ.

In addition to rules 1–5 of theorem 1, the logic also possesses a systematic procedure for testing ε-consistency (hence, ε-entailment), involving a moderate number of propositional satisfiability tests.

Definition: Given a truth-valuation t, a default statement $p \to q$ is said to be *verified* under t if t assigns the value 1 to both p and q. $p \to q$ is said to be *falsified* under t if p is assigned a 1 and q is assigned a 0. A default statement $S: p \to q$ is said to be *tolerated* by a set Δ' of such statements if there is a t that verifies S and does not falsify any statement in Δ'.

THEOREM 4 (Adams 1975): Let Δ be a finite set of default statements. Δ is ε-consistent iff in every non-empty subset Δ' of Δ there exists at least one statement that is tolerated by Δ'.

COROLLARY 1 (Goldszmidt and Pearl 1989): Consistency (hence entailment) can be tested in $|\Delta|^2/2$ propositional satisfiability tests.

The procedure is simply to find every default statement that is tolerated by Δ, remove those from Δ, and repeat the process on the remaining set of statements, until there are no more default statements left. If this process leads to an empty set then Δ is ε-consistent, else it is inconsistent.

If the material counterpart of $p \supset q$ of each statement $p \to q$ in Δ is a Horn expression, then consistency can be tested in time quadratic with the number of literals in Δ.

When Δ can be represented as a network of default rules, the criterion of theorem 4 translates into a simple graphical test for consistency, generalizing that of Touretzky (1986):

COROLLARY 2 (Pearl 1987a): Let Δ be a *default network*, that is, a set of default statements $p \to q$ where both p and q are atomic propositions (or negation thereof). Δ is consistent iff for every pair of conflicting arcs $p_1 \to q$ and $p_2 \to \neg q$

1. p_1 and p_2 are distinct, and
2. There is no cycle of positive arcs that embraces both p_1 and p_2.

Theorems 3, 4, and their corollaries are valid only when Δ consists of purely defeasible conditionals. For mixtures of defeasible and nondefeasible statements, consistency and entailment require a slightly modified procedure (Goldszmidt and Pearl 1989). This procedure attributes a special meaning to a strict conditional $a \Rightarrow b$, different than the material implication $a \supset b$. For example, conforming to common usage of conditionals, it proclaims $\{a \Rightarrow b, a \Rightarrow \neg b\}$ as inconsistent and will entail $a \Rightarrow b$ from $\neg b \Rightarrow \neg a$ but not from $\neg a$. Another extension of ε-semantics, permitting defaults to be given different strength, is treated in (Goldszmidt and Pearl 1991).

8.3.4 Illustrations

To illustrate the power of ε-semantics and, in particular, the syntactical and graphical derivations sanctioned by theorems 1, 3, and 4, consider the celebrated "Penguin triangle" of figure 8.1. T comprises the sentences:

$$F = \{Penguin\ (Tweety), Bird(Tweety)\}, \tag{8}$$

$$\Delta = \{Penguin \rightarrow \neg fly, Bird \rightarrow Fly, Penguin \rightarrow Bird\}; \tag{9}$$

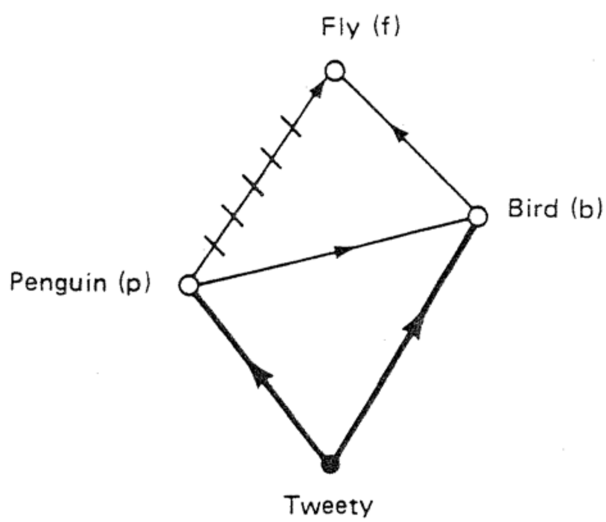

Figure 8.1
A network representing the knowledge base of equations (8) and (9). Heavy arcs represent evidence about individuals, thin arcs represent default statements, slashed arcs represent default denials. The arc between Penguin and Bird imposes specificity preference, yielding the conclusion "Tweety does not fly."

Although Δ does not specify explicitly whether penguin-birds fly, the desired conclusion is derived in three steps, using rules 1 and 3 of theorem 1:

1. *Penguin* (*Tweety*) $\mid\!\sim_\Delta$ $\neg$*Fly* (*Tweety*) (from Rule 1)
2. *Penguin* (*Tweety*) $\mid\!\sim_\Delta$ *Bird* (*Tweety*) (from Rule 1)
3. *Penguin* (*Tweety*), *Bird* (*Tweety*) $\mid\!\sim_\Delta$ $\neg$*Fly* (*Tweety*)
   (Applying rule 3 to lines 1, 2)

Note that preference toward subclass specificity is maintained despite the defeasible nature of the rule *Penguin* $\rightarrow$ *Bird*, which admits exceptional penguins in the form of non-birds.

We can also derive this result using theorems 3 and 4 by showing that the denial of the conclusion $p \wedge b \rightarrow \neg f$ is ε-inconsistent with

$$\Delta = \{p \rightarrow \neg f, b \rightarrow f, p \rightarrow b\}. \tag{10}$$

Indeed, no truth-valuation of $\{p, b, f\}$ can verify any sentence in

$$\Delta' = \{p \rightarrow \neg f, p \rightarrow b, p \wedge b \rightarrow f\} \tag{11}$$

without falsifying at least one other sentence.

Applying theorem T-3 to the network of figure 8.1 yields another plausible conclusion. *Bird* $\rightarrow$ $\neg$*Penguin*, stating that when one talks about birds one does not have penguins in mind, that is, penguins are exceptional kind of birds. It is a valid conclusion of Δ because every P in $\mathscr{P}_{\Delta,\varepsilon}$ must yield $P(p \mid b) = O(\varepsilon)$. Of course, if the statement *Bird* $\rightarrow$ *Penguin* is artificially added to Δ, inconsistency results; as ε diminishes below a certain level (1/3 in our case), $\mathscr{P}_{\Delta,\varepsilon}$ becomes empty. This can be predicted from purely topological considerations (corollary 2), since adding the arc *Bird* $\rightarrow$ *Penguin* would create a cycle of positive arcs embracing "bird" and "penguin," and these sprout two conflicting arcs toward "fly." Moreover, theorem 3 implies that if a network becomes inconsistent by the addition of S, then that network ε-entails its denial, S'. Hence, the network of Figure 8.1 ε-entails *Bird* $\rightarrow$ $\neg$*Penguin*. By the same graphical method one can easily show that the network also ε-entails the natural conclusion, *Fly* $\rightarrow$ $\neg$*Penguin*. This contraposition of *Penguin* $\rightarrow$ $\neg$*Fly* is sanctioned only because the existence of flying objects that are not penguins (i.e., normal birds) is guaranteed by the other rules in Δ.

8.4 The Adventurous Shell

The preceding adaptation of Adams's logic of conditionals yields a system of defeasible inference with rather unique features:

1. The system provides a formal distinction between exceptions, ambiguities, and inconsistencies and offers systematic methods of testing and maintaining consistency.
2. Multiple extensions do not arise, and preferences among arguments (e.g., toward higher specificity) are respected by natural deduction.
3. There is no need to specify abnormality relations in advance (as in circumscription); such relations (e.g., that penguins are abnormal birds) are automatically inferred from the knowledge base.

However, default reasoning requires two facilities: one which forces conclusions to be retractable in the light of new refuting evidence; the second which protects conclusions from retraction in the light of new but irrelevant evidence. Rules 1–5 excel on the first requirement but fail on the second. For instance, in the example of figure 8.1, if we are told that Tweety is also a blue penguin, the system would retract all previous conclusions (as ambiguous), even though there is no rule which in any way connects color to flying. (The opposite is true in default logics—they excel on the second requirement but do not retract conclusions refuted by more specific information, unless all exceptions are enumerated in advance [Reiter 1987].)

The reason for this conservative behavior lies in our insistence that any issued conclusion attains high probability in *all* probability models licensed by Δ and one such model reflects a world in which blue penguins do fly. It is clear that if we want the system to respect the communication convention that, unless stated explicitly, properties are presumed to be *irrelevant* to each other, we need to restrict the family of probability models relative to which a given conclusion must be checked for soundness. In other words, we should consider only distributions which minimize dependencies relative to Δ, that is, they embody dependencies which are absolutely implied by Δ, and no others.

8.4.1 System Z

One way of suppressing irrelevant properties is to restrict our attention to the "most normal" or "least surprising" probability models that comply with the constraints in Δ. This can be most conveniently done within the nonstandard analysis of Spohn (see section 8.2.2), where $\kappa(w)$ represents the degree of surprise associated with world w. To ratify a sentence $p \to q$ within this paradigm, we must first find the minimal κ distribution permitted by the constraints in Δ and, then, test whether $\kappa(q|p) < \kappa(\neg q|p)$ holds in this distribution.

Translating the constraints of equation (4) to the language of nonstandard analysis, yields

$$\kappa(v \wedge u) < \kappa(\neg v \wedge u) \quad \text{if } u \to u \in \Delta \tag{12.a}$$

where κ of a formula f is given by

$$\kappa(f) = \min_w \{\kappa(w) : w \models f\} \tag{12.b}$$

Remarkably, these constraints admit a unique minimal κ distribution whenever Δ is ε-consistent. Moreover, finding this minimal distribution, which was named **Z**-ranking in Pearl 1990, requires no more computation than testing for ε-consistency according to corollary 1. We first identify all default statements in Δ that are tolerated by Δ, assign to them a **Z**-rank of 0, and remove them from Δ. Next we assign a **Z**-rank of 1 to every default statement that is tolerated by the remaining set, and so on. Continuing in this way, we form an ordered partition of $\Delta = (\Delta_0, \Delta_1, \Delta_2, \ldots, \Delta_K)$, where Δ_i consists of all statements tolerated by $\Delta - \Delta_0 - \Delta_1 - \ldots \Delta_{i-1}$. This partition uncovers a natural priority among the default rules in Δ, and represents the relative "cost" associated with violating any of these defaults, with preference given to the more specific classes.

Once we establish the **Z**-ranking on defaults, the minimal ranking on worlds is given as follows:

THEOREM 5 (Pearl 1990): Out of all ranking functions $\kappa(w)$ satisfying the constraints in equation (12) the one that achieves the lowest κ for each world w is unique and is given by

$$\mathbf{Z}(w) = \min\{n : w \models (v \supset u), \mathbf{Z}(v \to u) \geq n\} \tag{13}$$

In other words, $\mathbf{Z}(w)$ is equal to 1 plus the rank of the highest-ranked default statement falsified in w.

Given $\mathbf{Z}(w)$, we can now define a useful extension of ε-entailment, which was called *1-entailment* in Pearl 1990.

Definition (1-entailment): A formula g is said to be *1-entailed* by f, in the context Δ, (written $f \mid\!\sim_1 g$), if g holds in all minimal-**Z** worlds satisfying f. In other words,

$$f \mid\!\sim_1 g \quad \text{iff} \quad \mathbf{Z}(f \wedge g) < \mathbf{Z}(f \wedge \neg g) \tag{14}$$

Note that ε-entailment is clearly a subset of 1-entailment.

Lehmann (1989) has extended ε-entailment in a different way, syntactically closing it under the *rational monotony* rule of equation 6, thus obtaining a new consequence relation which he called *rational closure*. Goldszmidt and Pearl (1990) have shown that 1-entailment and rational closure are identical whenever Δ is ε-consistent. Thus, the procedure of testing ε-consistency also provides a $|\Delta^2|/2$-time procedure for testing entailment in rational closure.

Figure 8.2 represents a knowledge base formed by adding three rules to that of figure 8.1:

1. "Penguins live in the antarctic" $p \to a$
2. "Birds have wings" $b \to w$
3. "Animals that fly are mobile" $f \to m$

The numerical labels on the arcs stand for the **Z**-ranking of the corresponding rules. The following are examples of plausible consequences that can be drawn from Δ by the various systems discussed in this section (ME will be discussed in section 8.4.2):

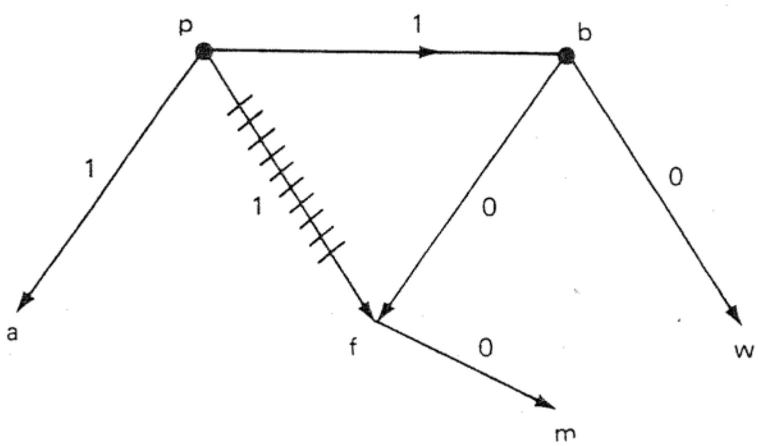

Figure 8.2
A knowledge base containing six defaults together with their **Z**-labels.

ε-entailed	1-entailed	ME-entailed
$b \wedge p \mid\!\sim_{\Delta} \neg f$	$\neg b \mid\!\sim_1 \neg p$	$p \mid\!\sim_{\mathrm{ME}} w$
$f \mid\!\sim_{\Delta} \neg p$	$\neg f \mid\!\sim_1 \neg b$	$p \wedge \neg a \mid\!\sim_{\mathrm{ME}} \neg f$
$b \mid\!\sim_{\Delta} \neg p$	$\neg f \mid\!\sim_1 m$	$p \wedge \neg a \mid\!\sim_{\mathrm{ME}} w$
$p \wedge a \mid\!\sim_{\Delta} b$	$\neg m \mid\!\sim_1 \neg b$	
	$p \wedge \neg w \mid\!\sim_1 b$	

1-entailment sanctions many plausible inference patterns that are not ε-entailed, among them chaining, contraposition and discounting irrelevant features. For example, from the knowledge base of figure 8.2 we can now conclude that birds are mobile, $b \mid\!\sim_1 m$, and that immobile objects are non-birds, $\neg m \mid\!\sim_1 \neg b$, and that green birds still fly. On the other hand, 1-entailment does not permit us to conclude that penguins who do not live in the antarctic still do not fly, $p \wedge \neg a \rightarrow \neg f$.

The main weakness of 1-entailment is its inability to sanction property inheritance from classes to exceptional sub-classes. For example, from $\Delta = \{a \rightarrow b, c \rightarrow d\}$ we cannot conclude $a \wedge \neg b \wedge c \rightarrow d$. Likewise, given the knowledge base of figure 8.2, 1-entailment will not sanction the conclusion that penguins have wings ($p \rightarrow w$) by virtue of being birds (albeit exceptional birds). The reason is that according to the **Z**-ranking procedure all statements conditioned on p should obtain a rank of 1, and this amounts to proclaiming penguins an exceptional type of birds in *all* respects, barred from inheriting *any* bird-like properties (e.g., laying eggs, having beaks). To sanction property inheritance across exceptional classes, a more refined ordering is required which also takes into account the *number* of defaults falsified in a given world, not merely their rank orders. One such refinement is provided by the maximum-entropy approach (Goldszmidt, Morris, and Pearl 1990) where each world is ranked by the sum of weights on the defaults falsified by that world. Another refinement is provided by Geffner's conditional entailment (Geffner 1989), where the priority of defaults induces a *partial* order on worlds. These two refinements will be summarized next.

8.4.2 The Maximum Entropy Approach

The maximum-entropy (ME) approach (Pearl 1988) is motivated by the convention that, unless mentioned explicitly, properties are presumed to be independent of one another, such presumptions are normally embedded in probability distributions that attain the maximum entropy subject to a set of constraints. Given a set Δ of default rules and a family of probability distributions that are admissible relative the constraints conveyed by Δ (i.e., $P(\beta_r \rightarrow \alpha_r) \geq 1 - \varepsilon \, \forall r \in \Delta$), we single out a distinguished distribution $P^*_{\varepsilon,\Delta}$ having the greatest entropy $-\sum_w P(w)\log(w)$, and define entailment relative to this distribution by

$$f \mathrel{|\!\!\sim}_{\mathrm{ME}} g \quad \text{iff} \quad P^*_{\varepsilon,\Delta}(g \mid f) \underset{\varepsilon \to 0}{\rightarrow} 1. \tag{15}$$

An infinitesimal analysis of the ME approach also yields a ranking function κ on worlds, where $\kappa(w)$ corresponds to the lowest exponent of ε in the expansion of $P^*_{\varepsilon,\Delta}(w)$ into a power series in ε. It can be shown that this ranking function can be encoded parsimoniously by assigning an integer weight κ_r to each default rule $r \in \Delta$ and letting $\kappa(w)$ be the sum of the weights associated with the rules falsified by w. The weight κ_r, in turn, reflects the "cost" we must add to each world w that falsifies rule r, so that the resulting ranking function would satisfy the constraints conveyed by Δ, namely,

$$\min\{\kappa(w)\colon w \models \alpha_r \wedge \beta_r\} < \min\{\kappa(w)\colon w \models \alpha_r \wedge \neg\beta_r\}, \quad r : \alpha_r \rightarrow \beta_r \in \Delta. \tag{16}$$

These considerations lead to a set of $|\Delta|$ nonlinear equations for the weights κ_r which, under certain conditions, can be solved by iterative methods. Once the rule weights are established, ME-entailment is determined by the criterion of equation (15), translated to

$$f \mathrel{|\!\!\sim}_{\mathrm{ME}} g \quad \text{iff} \quad \min\{\kappa(w)\colon w \models f \wedge g\} < \min\{\kappa(w)\colon w \models f \wedge \neg g\} \tag{17}$$

where

$$\kappa(w) = \sum_{r\colon w \models \alpha_r \wedge \neg\beta_r} \kappa_r$$

We see that ME-entailment requires minimization over worlds, a task that is NP-hard even for Horn expressions (Ben-Eliyahu 1990). In practice, however, this minimization is accomplished quite effectively in network type databases, yielding a reasonable set of inference patterns. For example, in the database of figure 8.2, ME-entailment will sanction the desired consequences $p \mathrel{|\!\!\sim} w$, $p \wedge \neg a \mathrel{|\!\!\sim} \neg f$ and $p \wedge \neg a \mathrel{|\!\!\sim} w$ and, moreover, unlike 1-entailment it will conclude $c \wedge p \mathrel{|\!\!\sim}_1 \neg f$ from $\Delta \cup \{c \rightarrow f\}$, where c is an irrelevant property.

An interesting feature of the ME approach is its sensitivity to the format in which the rules are expressed. This is illustrated in the following example. From $\Delta = \{$Swedes are blond, Swedes are well mannered$\}$, ME will conclude that dark-haired Swedes are still well mannered, while no such conclusion will be drawn from $\Delta = \{$Swedes are blond and well mannered$\}$. This sensitivity might sometimes be useful for distinguishing fine nuances in natural discourse, indicating, for example, that behavior and hair color are two independent qualities (as opposed, say, to hair color and eye color). However, it stands at variance with most approaches to default reasoning, where $a \rightarrow b \wedge c$ is treated as a shorthand notation of $a \rightarrow b$ and $a \rightarrow c$.

The failure to respond to causal information (see Pearl 1988: 463, 519 and Hunter 1989) prevents the ME approach from properly handling tasks such as the Yale shooting problem (Hanks and McDermott 1986), where rules of causal character should be given priority over other rules. This weakness may perhaps be overcome by introducing causal operators into the ME formulation, similar to the way causal operators are incorporated within other formalisms of nonmonotonic reasoning (e.g., Shoham 1986, Geffner 1989).

8.4.3 Conditional Entailment

Geffner (1989) has overcome the weaknesses of 1-entailment by introducing two new refinements. First, rather than letting rule priorities dictate a ranking function on worlds, a partial order on worlds is induced instead. To determine the preference between two worlds, w and w', we examine the highest priority default rules that distinguish between the two, that is, that are falsified by one and not by the other. If all such rules remain unfalsified in one of the two worlds, then this world is the preferred one. Formally, if $\Delta[w]$ and $\Delta[w']$ stand for the set of rules falsified by w and w', respectively, then w is preferred to w' (written $w < w'$) iff $\Delta[w'] \neq \Delta[w']$ and for every rule r in $\Delta[w] - \Delta[w']$ there exists a rule r' in $\Delta[w'] - \Delta[w]$ such that r' has a higher priority than r (written $r \prec r'$). Using this criterion, a world w will always be preferred to w' if it falsifies a proper subset of the rules falsified by w'. Lacking this feature in the Z-ordering has prevented 1-entailment from concluding $p \mid\!\sim w$ in the example of figure 8.2.

The second refinement introduced by Geffner is allowing the rule-priority relation, $\prec$, to become a partial order as well. This partial order is determined by the following interpretation of the rule $\alpha \rightarrow \beta$; if α is all that we know, then, regardless of other rules that Δ may contain, we are authorized to assert β. This means that $r\colon \alpha \to \beta$ should get a higher priority than any argument (a chain of rules) leading from α to $\neg\beta$ and, more generally, if a set of rules $\Delta' \subset \Delta$ does not tolerate r, then at least one rule in Δ' ought to have a lower priority than r. In figure 8.2, for example, the rule $r_3\colon p \wedge \neg f$ is not tolerated by the set $\{r_1\colon p \to b, r_2\colon b \to f\}$, hence, we must have $r_1 \prec r_3$ or $r_2 \prec r_3$. Similarly, the rule $r_1\colon p \to b$ is not tolerated by $\{r_2, r_3\}$, hence, we also have $r_2 \prec r_1$ or $r_3 \prec r_1$. From the asymmetry and transitivity of $\prec$, these two conditions yield $r_2 \prec r_3$ and $r_2 \prec r_1$. It is clear, then, that this priority on rules will induce the preference $w < w'$ whenever w validates $p \wedge b \wedge \neg f$ and w' validates $p \wedge b \wedge f$; the former falsifies r_2, while the latter falsifies the higher priority rule r_3. In general, we say that a proposition g is *conditionally entailed* by f (in the context of Δ) if g holds in all the preferred worlds of f induced by every priority ordering admissible with Δ.

Conditional entailment bridges the conditional and dispositional approaches to default reasoning. It rectifies the shortcomings of 1-entailment and ME-entailment. However, having been based on model minimization as well as on enumeration of subsets of rules, its computational complexity might be overbearing. A proof theory for conditional entailment and its unification with causal theories can be found in Geffner 1989.

8.4.4 Dialectic Approaches

Dialectic approaches attempt to supplement the probabilistic interpretation of defaults with a set of assumptions about conditional independence drawn on the basis of the syntactic structure of Δ. For a default $p \to q$, these approaches assume the probability of q to be high not only when p is all that is known, but also in the presence of an additional body of evidence which does not provide an *argument* against q (Loui 1987, Pollock 1987). This interpretation is closer in spirit to the syntactic approaches to nonmonotonic reasoning proposed by Reiter (1980) and McDermott and Doyle (1980), which allow us to infer q from p in the absence of "proofs" for $\neg q$.

In the systems reported in Geffner and Pearl 1988 and Geffner 1988, these ideas take the form of an additional inference rule, similar to:

RULE 6: Irrelevance

If $p \to r \in \Delta$ and $I_\Delta(q, \neg r \mid p)$, then $p \wedge q \mathrel{|\!\sim}_\Delta r$,

where the predicate $I_\Delta(q, \neg r \mid p)$, which reads: "$q$ is irrelevant to $\neg r$ given p," expresses the conditional independence $P(r \mid p) \approx P(r \mid p, q)$.

The mechanism for evaluating the irrelevance predicate $I_\Delta(q, \neg r|p)$ appeals to the set ψ' of wffs formed by converting each default $p \rightarrow q$ in Δ into a corresponding material implication $p \supset q$. In essence, q is then said to be *relevant to* $\neg r$ given p, if there is a a set ψ' of implications in ψ which permit an *argument* for $\neg r$ to be constructed, i.e if $\psi', p, q \mid\!\sim \neg r$, with the set of wffs $\psi' \cup p, q$ being logically consistent. The set ψ' is called the *support* of the argument for $\neg r$. If q is not relevant to $\neg r$ given p, then q is assumed to be *irrelevant to* $\neg r$ in that context, and $I_\Delta(q, \neg r|p)$ thus holds.

This simple extension permits us to infer, for instance, that red birds are likely to fly given a default stating that birds fly, as "redness" does not induce any argument in support of not flying. Further refinements are installed to insure that arguments for $\neg r$ that are blocked by p (or its consequences) do not bear on the predicate I_Δ. With this refinement, most examples analyzed in the literature yield the expected results.

Dialectic approaches constitute an alternative way of extending the inferential power of the core set of probabilistic rules. An advantage of these approaches over those based on maximum entropy is intelligibility: derivations under this approach can usually be justified in a more natural fashion. On the other hand, these approaches lack the foundational basis of a principle like maximum entropy, making it difficult to justify and make precise the form these extensions should take.

8.5 Do People Reason with Extreme Probabilities (or Lotteries and other Paradoxes of Abstraction)

Neufeld and Poole (1988) have raised the following objection (so-called Dingo Paradox) in connection with the theorem of exceptions (T-3). We saw that the penguin triangle (fig. 8.1) sanctions the conclusion $Bird \rightarrow \neg Penguin$ by virtue of the fact that penguins are an exceptional class of birds (relative to flying). Similarly, if "sandpipers" are birds that build nests in sand, we would conclude $Bird \rightarrow \neg Sandpiper$. Continuing in this manner through all types of birds, and assuming that every subclass of birds has a unique, distinguishing trait, we soon end up with the conclusion that birds do not exist—birds are not penguins, not sandpipers, not canaries, and so on—thus ruling out all types of birds.

This paradox is a variant of the celebrated Lottery Paradox (Kyburg 1961): Knowing that a lottery is about to have one winner is incompatible with common beliefs that each individual ticket is, by default, a loser. Indeed, the criterion provided by ε-semantics would proclaim the overall set of such statements ε-inconsistent, since the set of conditional probabilities

$$P(Loser(i) \mid Ticket(i)) \geq 1 - \varepsilon \qquad i = 1, 2, \ldots, N$$

cannot be satisfied simultaneously for $\varepsilon < 1/N$. Perlis (1987) has further shown that every default logic is bound to suffer from some version of the lottery paradox if we insist on maintaining deductive closure among beliefs.

Are these paradoxes detrimental to ε-semantics, or to nonmonotonic logics in general? I would like to argue that they are not. On the contrary, I view these paradoxes as healthy reminders that in all forms of reasoning we are dealing with simplified abstractions of real-world knowledge, that we may occasionally step beyond the boundaries of one abstraction and that, in such a case, a more refined abstraction should be substituted.

Predicate logic and probability theory are two such abstractions, and ε-semantics offers an abstraction that is somewhere between logic and probability. It requires less input than probability theory (e.g., we need not specify numerical probabilities), but more input than logic (e.g., we need to distinguish between defaults, $a \rightarrow b$, and observations, $\neg a \vee b$). It is more conservative than logic (e.g., it does not sanction transitivity), but more adventurous than probability theory (e.g., it admits conclusions even if their probabilities approach 1 very slowly, such as $= (1 - \varepsilon)^N$.

Each abstraction constitutes an expedient simplification of reality, tailored to serve a specialized set of tasks. Each simplification is supported by a different symbol processing machinery and by a set of norms, to verify whether the simplification and its supporting machinery are still applicable. The lottery paradox represents a situation where ε-semantics no longer offers a useful abstraction of reality. Fortunately, however, the consistency norms help us identify such situations in advance, and alert us (in case our decisions depend critically on making extensive use of the disjunction axiom) that a finer abstraction should be in order (perhaps full-fledged probability theory).

Probabilities that are infinitesimally close to 0 and 1 are very rare in the real world. Most default rules used in ordinary discourse maintain a non-vanishing percentage of exceptions, simply because the number of objects in every meaningful class is finite. Thus, a natural question to ask is, why study the properties of an abstraction that applies only to extreme probabilities? Why not develop a logic that characterizes moderately high probabilities, say probabilities higher than 0.5 or 0.9—or more ambitiously, higher than α, where α is a parameter chosen to fit the domains of the predicates involved? Further, why not develop a logic that takes into account utility information, not merely probabilities, thus formalizing reasoning about actions, in addition to beliefs (Doyle 1988)?

The answer is that any such alternative logic would be too complicated; it would need to invoke many of the axioms of arithmetic, and would require more information than is usually available. Almost none of the patterns of reasoning found in common conversation will remain sound relative to such semantics. Take, for example, the logic of "majority," namely, interpreting the default rule $a \rightarrow b$ to mean "The majority of *a*s are *b*s," or $P(b|a) > 0.5$. Only the first two axioms of theorem 1 remain sound in this interpretation. Even the cumulativity axiom (rule 3, theorem 1), which is rarely disputed as a canon of default reasoning, is flatly violated by some proportions (e.g., $P(q|p) = 0.51$ and $P(r|p) = 0.51$ could yield $P(r|p \wedge q) = 0.02$, in case $P(\neg q \wedge \neg r) = 0$.)

How, then, do people reason qualitatively about properties and classes, proportions and preferences? It appears that, if the machinery invoked by people for such tasks stems from approximating numerical information by a set of expedient abstractions, then the semantics of extreme probabilities is one of the most popular among these abstractions. The axioms governing this semantics (i.e., rules 1–5, theorem 1) appears to have been thoroughly entrenched as inference rules in plausible reasoning. For example, from the sentences "Most students are males" and "Most students will get an A," the cumulativity axiom would infer "Most male students will get an A." This conclusion can be grossly incorrect, as shown in the paragraph above, yet it is a rather common inference made by people, given these two inputs. In conclusion, it appears that the machinery of plausible reasoning reflects a remarkable agreement with the rules of "almost all" logic.[2]

8.6 Relations to Other Nonmonotonic Systems

The logic closest in spirit to the probabilistic approaches presented in the preceding sections are those based on model preferences, where conclusions

[2] In an earlier version of this paper, as well as in Pearl 1988, I have speculated that this agreement indicates that plausible reasoning is more in line with the rules of "almost all" logic than with those of "support" or "majority" logics. I am now in the opinion that this agreement is more reflective of tacit assumptions of independence than of the type of logic chosen by people to reason about proportions. In our example, the conclusion "most students will receive an A" reflects the assumption that grades are independent of gender. The same holds for my analysis of the Simpson paradox (Pearl 1988).

are sanctioned relative to the minimal (or least abnormal) models of the theory. The reason for this closeness is that, if we regard probability weights as measures of "normality," then probability theory is essentially a theory of preferences among models. The main point of departure, however, is that in probability theory formulae are scored by the *sum of the weights* on models that validate the formulae while in model-preference logic a formula assumes the weight of its *maximal-weight* model. The practical implications of this difference is best illustrated by comparing qualitative and probabilistic approaches to abduction.

8.6.1 Probabilistic and Qualitative Abduction

In the probabilistic approach, abduction is considered the task of finding the "most probable explanation" of the evidence observed, namely, seeking an instantiation of a set of explanatory variables that attains the highest probability, conditioned on the evidence observed. Assume that we have a probability function P defined on the set W of possible worlds. One way of accounting for abductive beliefs is to posit that, at any state of knowledge, beliefs are fully committed to a world that has the highest probability. In other words, a proposition A is *believed* if A holds in some $w^* \in W^*$, where $W^* \subseteq W$ is the set of most probable worlds,

$$W^* = \{w^* \mid P(w^*) \geq P(w) \; \forall w \in W\}.$$

To maintain coherence, we also demand that any set of propositions that are simultaneously believed, must hold in the same w^*. Nonmonotonic behavior is obtained by conditionalization; given a body of evidence (facts) e, the probability function $P(w)$ shifts to $P(w|e)$ and this yields a new set of most probable worlds

$$W_e^* = \{w^* \mid P(w^* \mid e) \geq P(w \mid e) \; \forall w \in W\}.$$

which, in turn, results in a new set of beliefs.

This approach to abduction was explored in Pearl 1987b, where a world was defined as an assignment of values to a set of interdependent variables (e.g., assignment of *TRUE-FALSE* values to a set of diseases in medical diagnosis), and the worlds in W^* were called *most probable explanations* (MPE). It was shown that the task of finding a most probable explanation of a body of evidence is no more complex than that of computing the probability of a single proposition. In singly connected networks (directed trees with unrestricted orientation) the task can be accomplished in linear time, using a parallel and distributed message-passing process. In multiply connected networks, the problem is NP-hard; however, clustering, conditioning and stochastic simulation techniques can offer practical solutions in reasonable time if the network is relatively sparse. Applications to circuit and medical diagnosis are described in Geffner and Pearl 1987, Pearl 1988, and Peng and Reggia 1987.

The MPE approach provides a bridge between probabilistic reasoning and nonmonotonic logic. Like the latter, the method provides systematic rules that lead from a set of factual sentences (the evidence) to a set of conclusion sentences (the accepted beliefs) in a way that need not be truth-preserving. However, whereas the input-output sentences are categorical, the medium through which inferences are produced is numerical, and the parameters needed for complete specification of $P(w)$ may not be readily available. In modeling man-made systems such as digital circuits this problem is not too severe, because all internal relationships are provided with the system's specifications. However, in medical diagnosis, as well as in reasoning about everyday affairs, the requirement of specifying a complete probabilistic model is too cumbersome and can be justified only in cases where critical decisions are at stake.

The qualitative approaches demand fewer judgments in constructing the knowledge base, but suffer from the lack of rating among competing explanations and, closely related to it, the lack of rating among pending information sources. To overcome this deficiency, the qualitative approaches make explicit appeal to explanatory scenarios, and seek scenarios that are both coherent and parsimonious.

A major challenge facing and the qualitative approaches is to enforce an appropriate separation between the *prospective* and *retrospective* modes of reasoning so as to capture the intuition that predictions should not trigger suggestions. To use my favorite example: "Sprinkler On" predicts "Wet Grass," "Wet Grass" suggests "Rain," but "Sprinkler On" should not suggest "Rain." In the probabilistic approach such separation is enforced via the patterns of independencies that are assumed to accompany causal relationships. In the qualitative approaches the separation is accomplished in two ways. One is to label sentences as either *causally established* (i.e., explained) or *evidentially established* (i.e., conjectured) and subject each type to a different set of inference rules (Pearl 1988a; Geffner 1989). The second method is to regard abduction as a specialized metaprocess that operates on a causal theory (Poole 1987; Reiter 1987).

In qualitative theories simplicity is enforced by explicitly encoding the preference of simple theories over complex ones, where simple and complex are given syntactical definitions, for example, smallest number of (cohesive) propositions (Thagard 1989) or minimal covering (Reiter 1987; Reggia et al. 1983). These syntactic ratings do not always coincide with the notion of plausibility, for example, two common diseases are often more plausible than a single rare disease in explaining a given set of symptoms (Reggia 1989). In probabilistic theories, coherence and simplicity are managed together by one basic principle—maximum posterior probability.

8.6.2 Relation to Model Preference Semantics

The model preference approach to nonmonotic reasoning (Shoham 1987) leaves room for widely different interpretations of defaults, ranging from the adventurous to the conservative. The adventurous approach takes the statement $A \rightarrow B$ to mean: Every world where $A \wedge B$ holds has a prima facie preference over the corresponding world where $A \wedge \neg B$ holds, everything else being equal (the terms "world" and "models" are used interchangeably in the literature). Conflicts are later resolved by extra logical procedures (Selman and Kautz 1988). The conservative school (Lehmann and Magidor 1988, DelGrande 1988) takes $A \rightarrow B$ to be a faint reflection of a preexisting preference relation, saying merely: B holds in all the most preferred worlds compatible with A. Whether a collection of such faint clues is sufficient to reveal information (about the preference relation) that entails a new statement $x \rightarrow y$, depends on the type of restrictions the preference relation is presumed to satisfy.

Lehmann and Magidor (1988) have identified the class of preference relations, whose consequence relation satisfies a reasonable set of rationality requirements including, for example, cumulativity and rational monotony. In essence, the restriction is that states of worlds be *ranked* (e.g., by some numerical score r) such that a state of lower rank is preferred to a state of higher rank. Lehmann and Magidor proved that the consequence relation induced by this class of ranked worlds coincides exactly with Adams's ε-entailment relation defined in equation (4) and, of course, its properties coincide with rules 1–5 of theorem 1.

It is remarkable that two totally different interpretations of defaults yield identical sets of conclusions and identical sets of reasoning machinery. Note that, even if we equate rank with probability, the interpretation $P(B|A) > 1 - \varepsilon$ is different from the model preference interpretation, because, for any finite ε, the former permits the most probable world of A to be incompatible with B. Fortunately, the two interpretations coincide in the language of non-standard analysis (see sections 8.2.2 and 8.4.1). Based on this coincidence, it is now possible to transport shortcuts and intuitions across semantical lines. For example, theorem 3 establishes a firm connection between preferential entailment and preferential consistency. Similarly, theorems 4 and 5 determine the complexity of proving entailment in model preference semantics.

Perhaps the deepest point of tension between probability and traditional nonmonotonic logics revolves around the issue of specificity-based arguments, that is, finding ways to ensure that inferences be based on the most specific classes for which information is available (e.g., the inference that a penguin cannot fly must override the inference that a bird can fly). In the case of Reiter's default logic (Reiter 1980), this requires semi-normal rules, which explicitly enumerate exceptions (e.g. birds fly, unless they are penguins or ostriches or ...). In the case of circumscription, we must supply priorities among abnormalities (McCarthy 1986). Touretzky (1986) has argued that the enumeration of exceptions places an impractical burden on the management of inheritance networks, and he showed how attention to "inferential distance" in the network can assure priority for more specific arguments without such explicit enumeration. In section 8.3, we saw that specificity-based priorities were obtained naturally from probability theory, even if numerical probabilities are not used, provided that we interpret defaults as statements of high conditional probability, infinitesimally close to one. Identical facility is provided by the conditional logics of Lehmann and Magidor (1988) and of DelGrande (1988).

What sets these systems apart from circumscription and Reiter's default logics is the distinction between *knowledge* and *facts* (Δ and F) a distinction that, for some reason has not been totally accepted throughout the nonmonotonic community. Intuitively, the knowledge component specifies the tendency of things to happen, that is, relations that hold true in all worlds, while the facts or "observations" describe that which actually happened, that is, one particular world. In other words, the knowledge base Δ contains information that is equivalent to meta inference rules, telling us how to process "observations" to get conclusions about a particular situation or a particular individual.

In section 8.3 we saw, for example, that it makes a profound difference whether the sentence "all penguins are birds" is treated as a rule $p \rightarrow b$ in the knowledge base Δ, or as observational formula $\neg p \vee b$ in F. The latter would represent the English sentence "It has been observed that Tweety is either a non-penguin or a bird." The former is treated as constraint that shapes the set of admissible probability distributions (or κ rankings) while the latter serves as evidence upon which the admissible distributions are to be conditioned. The former gives the intended results, properly treating penguins as subclass of birds. The latter does not, because the observation $\neg p \vee b$ can be totally subsumed by other observations, say $p \wedge b$, thus yielding identical conclusions regardless of whether penguins are a subclass of birds or birds are a subclass of penguins.[3]

Acknowledgments

I am grateful to Ernest Adams, Hector Geffner, Moises Goldszmidt, Daniel Lehmann, David Makinson, Menachen Magidor, and Paul Morris for providing helpful comments on several topics of this paper. This work was supported in part by National Science Foundation Grant #IRI-8821444 and Naval Research Laboratory Grant #N00014-87-K-2029. An earlier version of this paper was presented at the First International Conference on Principles of Knowledge Representation and Reasoning (KR '89), Toronto, May 1989.

References

Adams, E. 1975. *The logic of conditionals.* Dordrecht, Netherlands: D. Reidel.

Bacchus, F. 1988. Representing and reasoning with probabilistic knowledge. Ph.D. dissertation, Dept. of Computer Science, University of Alberta, Edmonton.

Ben-Eliyahu, R. 1990. Minimizing Horn-Clause violations is NP-hard UCLA Cognitive Systems Laboratory, Computer Science Dept., *Technical Report* R–158.

DelGrande, J. P. 1988. An approach to default reasoning based on first-order conditional logic; revised report. *Artificial Intelligence*, 36:63–90.

Doyle, J. 1988. On universal theories of defaults. *Technical Report* CMU–CS–88–11. Carnegie Mellon University.

Geffner, H., and J. Pearl. 1988. A framework for reasoning with defaults. UCLA Cognitive Systems Laboratory, *Technical Report* 870058 (R–94), March 1988. Also in Kyburg et al. (eds.) *Knowledge Representation and Defeasible Reasoning*, Kluwer, pp. 61–87 1990.

Geffner, H. 1988. On the logic of defaults. *Proceedings of the National Conference on AI (AAAI–88)*, St. Paul, 449–454.

[3] The story of the paper (Geffner and Pearl 1988) exemplifies the traditional resistance to distinguishing between knowledge and observations. This paper was rejected by the CSCSI–88 Conference Committee, because the referee would not allow conclusions to change as sentences move from F to Δ.

Geffner, H. 1989. Default reasoning: causal and conditional theories. Ph.D. dissertation, UCLA Computer Science Department, Cognitive Systems Laboratory *Technical Report* (R-137).

Goldszmidt, M., and J. Pearl. 1989. On the consistency of defeasible databases. *Proceedings of the 5th Workshop on Uncertainty in AI*. Windsor, Canada, 134–141. Also, *Uncertainty in AI-5*. North Holland, 1990, 87–97.

Goldszmidt, M., and J. Pearl. 1990. On the relation between rational closure and System-Z. *Proceedings, Third International Workshop on Nonmonotonic Reasoning*. S. Lake Tahoe, CA, May 1990, 130–140.

Goldszmidt, M., and J. Pearl. 1991. System Z^+: A formalism for reasoning with variable-strength defaults. *Proceedings of the National Conference on AI (AAA-91)*. Anaheim, CA.

Goldszmidt, M., P. Morris, and J. Pearl. 1990. A maximum entropy approach to non-monotonic reasoning. *Proceedings of the National Conference on AI (AAAI-90)*. Boston, MA, pp. 646–652.

Halpern, J. Y., and M. Rabin. 1987. A logic to reason about likelihood. *Artificial Intelligence* 32 (3): 379–406.

Hanks, S., and D. V. McDermott. 1986. Default reasoning, nonmonotonic logics, and the frame problem. *Proceedings of the 5th National Conference on AI (AAAI-86)*. Philadelphia, 328–333.

Hunter, D. 1989. Causality and maximum entropy updating. *International Journal of Approximate Reasoning* 3(1): 87–114.

Jaynes, E. T. 1979. Where do we stand on maximum entropy? In *The maximum entropy formalism*. R. D. Levine and M. Tribus, eds. Cambridge: MIT Press.

Kuipers, B. 1986. Qualitative simulation. *Artificial Intelligence*, 29: 289–338.

Kyburg, H. E. 1961. *Probability and the logic of rational belief*. Middleton, Conn.: Wesleyan University Press.

Lehmann, D., and Magidor, M. 1988. Rational logics and their models: a study in cumulative logics. Department of Computer Science, Hebrew University, Jerusalem, Israel, *Technical Report* TR-88-16.

Lehmann, D. 1989. What does a conditional knowledge base entail? *Proceedings of the First International Conference on Principles of Knowledge, Representation, and Reasoning*, 212–222. San Mateo: Morgan Kaufmann.

Loui, R. 1987. Defeat among arguments: A system of defeasible inference. *Computational Intelligence*.

McCarthy, J. 1980. Circumscription—A form of non-monotonic reasoning. *Artificial Intelligence* 13(1), 27–70.

McCarthy, J. 1986. Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence* 28(1): 89–116.

McDermott, D. V., and J. Doyle. 1980. Non-monotonic logic 1. *Artificial Intelligence* 13(1, 2): 41–72.

Neufeld, E., and D. Poole. 1988. Probabilistic semantics and defaults. *Proceedings of the 4th AAAI Workshop on Uncertainty in AI*. Minneapolis. 275–281.

Pearl, J. 1987. Deciding consistency in Inheritance Networks. UCLA Cognitive Systems Laboratory, *Technical Report* 870053 (R-96).

Pearl, J. 1987. Distributed revision of composite beliefs. *Artificial Intelligence* 33(2): 173–215.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* San Mateo: Morgan Kaufmann.

Pearl, J., and T. S. Verma. 1987. The logic of representing dependencies by directed graphs. *Proceedings of the 6th National Conference on AI (AAAI–87)*, Seattle, 374–379.

Pearl, J. 1990. System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning. In R. Parikh (ed.), *Theoretical Aspects of Reasoning about Knowledge* (TARK–III). San Mateo: Morgan Kaufmann, pp. 121–135.

Peng, Y., and J. A. Reggia. 1987. Plausibility of diagnostic hypotheses. *Proceedings of the 5th National Conference on AI (AAAI–86)*, Philadelphia, 140–145.

Perlis, D. 1987. On the consistency of common sense reasoning. *Computational Intelligence*, 2:180–190.

Poole, D. L. 1987. Defaults and conjectures: Hypothetical reasoning for explanation and prediction. *Research Report* CS–87–54, University of Waterloo.

Pollock, J. 1987. Defeasible Reasoning, *Cognitive Science*, 11:481–518.

Reggia, J. A. 1989. Measuring the plausibility of explanatory hypotheses. *Behavioral and Brain Sciences*. 12(3):486–487.

Reggia, J. A., D. S. Nau, and Y. Wang. 1983. Diagnostic expert systems based on a set-covering model. *International Journal of Man-Machine Studies* 19:437–460.

Reiter, R. 1980. A logic for default reasoning. *Artificial Intelligence* 13:81–132.

Reiter, R. 1987. A theory of diagnosis from first principles. *Artificial Intelligence* 32(1):57–95.

Reiter, R. 1987. Nonmonotonic reasoning. In *Annual review of computer science* 2, 147–186. Palo Alto, Calif.: Annual Reviews.

Selman, B. and H. Kautz. 1988. The complexity of model-preference default theories. *Proc. CSCSI–88*, Seventh Canadian Conference on Artificial Intelligence, Edmonton, Alberta, 102–109.

Shoham, Y. 1986. Chronological ignorance: Time, nonmonotonicity, necessity, and causal theories. *Proceedings of the 5th National Conference on AI (AAAI–86)*, Philadelphia, 389–393.

Shoham, Y. 1987. Nonmonotonic logics: meaning and utility. *Proceedings of International Joint Conference on AI (IJCAI–87)*, Milan, 388–393.

Spohn, W. 1988. A general non-probabilistic theory of inductive reasoning, *Proceedings of the 4th Workshop on Uncertainty in AI*, Minneapolis, 315–322. Also, in W. L. Harper and B. Skyrms, eds. *Causation in Decision, Belief Change, and Statistics*, Vol. 2. Reidel, Dordrecht, Netherlands, 1988, 105–134.

Thagard, P. 1989. Explanatory coherence. *Behavioral and Brain Sciences* 12(3):435–468.

Touretzky, D. 1986. *The mathematics of inheritance systems.* Los Altos, Calif.: Morgan Kaufmann.

Wellman, M. 1987. Probabilistic semantics for qualitative influences. *Proceedings of the National Conference on AI (AAAI–87)*, Seattle, 660–664.