Default Reasoning From Statistics

Fahiem Bacchus*
Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
N2L–3G1
fbacchus@logos.waterloo.ca

Abstract

There are two common but quite distinct interpretations of probabilities: they can be interpreted as a measure of the extent to which an agent believes an assertion, i.e., as an agent's degree of belief, or they can be interpreted as an assertion of relative frequency, i.e., as a statistical measure. Used as statistical measures probabilities can represent various assertions about the objective statistical state of the world, while used as degrees of belief they can represent various assertions about the subjective state of an agent's beliefs. In this paper we examine how an agent who knows certain statistical facts about the world might infer probabilistic degrees of beliefs in other assertions from these statistics. For example, an agent who knows that most birds fly (a statistical fact) may generate a degree of belief greater than 0.5 in the assertion that `Tweety` flies given that `Tweety` is a bird. This inference of degrees of belief from statistical facts is known as direct inference. We develop a formal logical mechanism for performing direct inference. Some of the inferences possible via direct inference are closely related to default inferences. We examine some features of this relationship.

Direct Inference

Probabilities can be used as a measure of the extent to which an agent believes an assertion, i.e., as a degree of belief, and they can also be used to represent statistical assertions of relative frequency. The key difference between these two types of probabilities is that statements of statistical probability are assertions about the objective statistical state of the world, while statements of degree of belief probability are assertions about the subjective state of an agent's beliefs.

For example, the statement "*More than 75% of all birds fly*" is a statistical assertion about the proportion of fliers among the set of birds; its truth is determined by the objective state of the world. On the other hand, the statement "*The probability that Tweety flies is greater than 0.75*" is a assertion about an agent's degree of belief; its truth is determined by the state of the agent who made the statement. The agent is using degrees of belief to quantify his uncertainty about the state of the world.

An important question is: what is the relationship between the agent's subjective degrees of belief and his knowledge about objective statistical features of the world? The answer that we will advance is that the agent's subjective degrees of belief should be *determined* by his knowledge of objective statistical facts though a mechanism that has been called *direct inference.* Direct inference is a non-deductive form of inference which has a long history in philosophy, e.g., [1, 2, 3, 4], but has received scant attention in AI.[1]

Simply stated, direct inference is the claim that the degree of belief an agent should assign to the proposition that a particular individual c possesses a particular property P, should be equal to the proportion of individuals that have P from among some class of individuals that c is known to belong to. For example, if the agent knows that `Tweety` belongs to the class of birds and that more than 75% of birds fly, then direct inference would sanction the inference that the agent's degree of belief in the proposition `Tweety flies` should be greater than 0.75. It is important to note that this is not a deductive inference: there is nothing excluding the possibility that `Tweety` is a member of the subclass of non-flying birds.

The motivation for interest in direct inference comes from its common use in applications of probability. For example, much of actuarial science is based on the principles of direct inference. When an insurance company quotes life insurance rates to particular individuals they compute those rates by equating their degree of belief (willingness to bet) in a *particular* individual's death with the *proportion* of deaths among some set of similar individuals (e.g., similar in terms of sex, age, job hazard, etc.).

*This work was supported by NSERC grant #OGP00-41848. Thanks to the referees for some useful comments.

[1] Loui [5] has used Kyburg's system of direct inference [2] as an underlying foundation in some of his systems of defeasible reasoning.

In this paper we will present a formal mechanism for performing direct inference. This mechanism is logically based, and uses a logic that can represent and reason with statistical and degree of belief probabilities. By reasoning inside of this logic we can reason about the consequences of direct inference. Although there are many other applications of direct inference, we will concentrate here on what it tells us about default reasoning by examining the relationship between some of the inferences that can be generated via direct inference and typical default inferences.

The Probability Logic

Prior to the construction of a mechanism for direct inference we need a formalism capable of representing and reasoning with both types of probabilities. In [6] a logic for statistical probabilities was developed by the author. Subsequently, Halpern used some of these ideas to develop a logic for degree of belief probabilities, and also demonstrated how the logic for statistical probabilities and the logic for degrees of belief could be combined to yield a combined probability logic capable of dealing with both types of probabilities [7].

In the combined probability logic, however, there is no intrinsic relationship between the two types of probabilities: they simply coexist without significant interaction (see [8]). However, the logic does provide a suitable formal framework for specifying a relationship. The main contribution of this research has been the development of such a specification: a mechanism for direct inference. This particular paper, however, will concentrate on the relationship between direct inference and default reasoning. To better understand the mechanism we will present a brief outline of the combined probability logic. For more details see [8] or [7].

The combined logic is a two sorted, first-order, modal logic. There is a sort of *objects* and a sort of *numbers*. The sort of objects consists of function and predicate symbols suitable for describing some domain of interest, while the numeric sort is used to make numeric assertions, especially assertions about probabilities. In particular, the numeric sort includes the constants 1, -1, and 0; the functions $+$ and $\times$; and the predicates $=$ and $\leq$. Additional inequality predicates and numeric constants can easily be added by definition, and we will use them freely. The formulas of the language are generated by applying standard first-order formula formation rules with the addition of two rules that generate new *numeric* terms (specifically probability terms) from existent formulas, i.e., these terms denote numbers (that correspond semantically to the values of the probability measure).

1. If α is a formula and $\vec{x}$ is a vector of n distinct object variables, then $[\alpha]_{\vec{x}}$ is a *statistical probability term.*
2. If α is a formula, then $\mathsf{prob}(\alpha)$ is a *degree of belief probability term.*

Since these constructs are terms, they can in turn be used in the generation of additional new formulas.

We can extend the language by definition to include conditional probability terms (of both types): $[\alpha|\beta]_{\vec{x}} =_{df} [\alpha \wedge \beta]_{\vec{x}}/[\beta]_{\vec{x}}$, and $\mathsf{prob}(\alpha|\beta) = \mathsf{prob}(\alpha \wedge \beta)/\mathsf{prob}(\beta)$.[2] Another useful abbreviation is: $\mathsf{cert}(\alpha) =_{df} \mathsf{prob}(\alpha) = 1$.

Semantically the language is interpreted using structures of the form

$$M = \langle \mathcal{O}, S, \vartheta, \mu_{\mathcal{O}}, \mu_S \rangle^3$$

Where:

1. $\mathcal{O}$ is a domain of objects (i.e., the domain of discourse). S is a set of states or possible worlds. ϑ is a world dependent interpretation of the symbols. The numeric symbols are interpreted as relations and function over the reals $\mathbb{R}$, with the numeric functions and predicates, $+$, $\times$, 1, -1, 0, $<$ and $=$, given their normal interpretation in every state.
2. $\mu_{\mathcal{O}}$ is a discrete probability measure over $\mathcal{O}$. That is, for every $A \subseteq \mathcal{O}$, $\mu_{\mathcal{O}}(A) = \sum_{o\in A} \mu_{\mathcal{O}}(o)$ and $\sum_{o\in\mathcal{O}} \mu_{\mathcal{O}}(o) = 1$.
3. μ_S is a discrete probability measure over S. That is, for every $S' \subseteq S$, $\mu_S(S') = \sum_{s\in S'} \mu_S(s)$ and $\sum_{s\in S} \mu_S(s) = 1$.

The formulas of the language are interpreted with respect to this semantic structure in a manner standard for modal languages. In particular, the interpretation of a formula will depend on a structure M, a current world $s \in S$, and a variable assignment function v. The probability terms are given the following interpretation:

1. $([\alpha]_{\vec{x}})^{(M,s,v)} = \mu_{\mathcal{O}}^n\{\vec{a} : (M, s, v[\vec{x}/\vec{a}]) \models \alpha\}$, where $v[\vec{x}/\vec{a}]$ is the variable assignment function identical to v expect that $v(x_i) = a_i$, and $\mu_{\mathcal{O}}^n$ is the n-fold product measure formed from $\mu_{\mathcal{O}}$.
2. $(\mathsf{prob}(\alpha))^{(M,s,v)} = \mu_S\{s' : (M, s', v) \models \alpha\}$.

So we see that $[\alpha]_{\vec{x}}$ denotes the measure of the set of satisfying instantiations of $\vec{x}$ in α and $\mathsf{prob}(\alpha)$ denotes the measure of the set of worlds that satisfy α. Another way of interpreting the statistical probability terms is to consider them as representing the probability that a *randomly selected*[4] vector of individuals $\vec{x}$ will satisfy α. We will sometimes call the variables, $\vec{x}$, bound by the statistical terms *random designators.*

The use of a product measure to assign probabilities to vectors means that we are treating the individuals in the vector as if they were selected independently of each other. Hence *marginal* probabilities

[2] For ease of exposition we will ignore the technicalities of dealing with division by zero. See [8] for details.

[3] Halpern has called such structures type III probability structures.

[4] That is, the probability of selecting any particular vector of individuals $\vec{o} \in \mathcal{O}^n$ is $\mu_{\mathcal{O}}^n\{\vec{o}\}$.

will display independence. For example, we have that $[P(x) \wedge R(y)]_{\langle x,y\rangle} = [P(x)]_x \times [R(y)]_y$. This says that if we select at random two individuals, the probability that the first has property P and the second has property R will be the product of the two probabilities: the two individuals are selected independently of each other. However, this not mean that the *conditional* probabilities will be independent. In particular, if we have knowledge that x and y are related, say through relation Q, we will no longer have independence, e.g., in general, $[P(x) \wedge R(y)|Q(x,y)]_{\langle x,y\rangle}$ is not decomposable into the product of simpler probability terms. For example, if we know that x and y are two randomly selected birds that we know are brothers, the probability that x is a flier will *not* be independent of the probability of y being a flier. However, if we have *no* knowledge of any relationship between x and y then the probabilities of each being a flier will be independent. For more on this point see [8].

Here are some examples to help the reader understand the language and its semantics.

1. $[\texttt{fly(x)}|\texttt{bird(x)}]_{\texttt{x}} > 0.75$. A particular triple (M, s, v) satisfies this formula if the probability that a randomly selected bird flies is greater than 0.75. That is, if more that 75% of the birds in the world 's' fly.
2. $\mathsf{prob}(\texttt{fly(Tweety)}) > \mathsf{prob}(\texttt{swim(Tweety)})$. This formula is satisfied if the measure of the set of worlds in which `Tweety` flies is greater than the measure of the set of worlds in which he swims. That is, the agent believes `Tweety` more likely to be a flier than a swimmer.
3. $\mathsf{prob}([\texttt{grey(x)}|\texttt{elephant(x)}]_{\texttt{x}} > 0.75) > 0.9$. This formula is satisfied if the measure of the set of worlds in which a randomly selected elephant has a greater that 75% probability of being grey, is greater than 0.9. That is, the agent has a high degree of belief that more that 75% of elephants are grey. Such knowledge might come from traditional statistical inference over sample information.

A powerful proof theory for this language can be constructed which is complete for various special cases [7, 8]. We simply note that all of the examples of reasoning we give here can be performed within this proof theory.

We need to extend the language beyond that developed by Halpern to include an expectation operator. Specifically, if t is a *statistical probability term*, then $\mathsf{E}(t)$ is a new numeric term whose denotation is defined as follows:

$$(\mathsf{E}(t))^{(M,s,v)} = \sum_{s' \in S} \mu_S(s') \times t^{(M,s',v)}.$$

That is, the expected value of a term is the weighted (by μ_S) average of its denotation over the set of possible worlds. The expected value as the advantage that it has the same denotation across all possible worlds (it is *rigid*), unlike the raw statistical term.

The Direct Inference Formalism

With the combined probability logic in hand we can describe the formalism of direct inference as a theory, i.e., as a collection of sentences, of that logic. Direct inference works by using an agent's base of accepted knowledge to assign degrees of belief in assertions that are *not* deductive consequences of the knowledge base.

We will call a formula of our logic *objective* if it does not contain any 'prob' or 'E' operators. Such formulas do not depend on any world except the current one for their interpretation. We suppose that the agent's knowledge base is represented by a finite collection of objective formulas, and we let `KB` denote the formula generated by conjoining all of the formulas in the knowledge base. `KB` will usually include information about particular individuals, e.g., `bird(Tweety)`, general logical relationships between properties, e.g., $\forall\texttt{x.bird(x)} \rightarrow \texttt{animal(x)}$, and *statistical information*, e.g., $[\texttt{fly(x)}|\texttt{bird(x)}]_{\texttt{x}} > .75$.

Definition 1 [Randomization] Let α be an objective formula. If $\langle c_1, \ldots, c_n\rangle$ are *all* the n distinct object constants that appear in $\alpha \wedge \texttt{KB}$ and $\langle v_1, \ldots, v_n\rangle$ are n distinct object variables that do not occur in $\alpha \wedge \texttt{KB}$, then let $\texttt{KB}^{\mathbf{v}}$ $(\alpha^{\mathbf{v}})$ denote the new formula which results from textually substituting c_i by v_i in `KB` (α), for all i.

Definition 2 [Direct Inference Principle] For any objective formula α the agent's degree of belief in α should be determined by the equality

$$\mathsf{prob}(\alpha) = \mathsf{E}\Big([\alpha^{\mathbf{v}}|\texttt{KB}^{\mathbf{v}}]_{\vec{v}}\Big).^5$$

In addition, the agent must fully believe that $[\texttt{KB}^{\mathbf{v}}]_{\vec{v}} > 0$, i.e., $\mathsf{cert}([\texttt{KB}^{\mathbf{v}}]_{\vec{v}} > 0)$.[6]

The collection of all instances of the direct inference principle specifies a theory which characterizes the agent's reasoning with direct inference.

Definition 3 [The Base Theory] Let D_0 be the set of formulas consisting of $\mathsf{cert}([\texttt{KB}^{\mathbf{v}}]_{\vec{v}} > 0)$ along with all instances of the direct inference principle. Let T_0 denote the closure of D_0 under logical consequence. T_0 is the agent's base theory given her initial knowledge base `KB`.

Before going further we note some important theorems which demonstrate that our mechanism is coherent.

[5] The justification for our use of the expected value of the statistical term is given in [8]. A pragmatic reason is that the expectations terms are rigid, as noted above.

[6] There is a very reasonable justification for this assumption [8], but here it is sufficient to note that it allows us to avoid division by zero.

Theorem 1

1. The degrees of belief given by the direct inference principle are in fact probabilities.
2. If $\mathtt{KB} \wedge [\mathtt{KB}^{\mathbf{V}}]_{\vec{v}} > 0$ is satisfiable, then so is T_0 (i.e., T_0 is consistent).
3. $\mathsf{prob}(\alpha) = 1$ is a logical consequence of the direct inference principle if and only if $\models \mathtt{KB} \rightarrow \alpha$.
4. The agent's full beliefs, i.e., the formulas which are assigned degree of belief 1 in T_0, have the same logical behavior as a collection of *KD45* beliefs with cert acting as a belief modality.

These theorems give us the following picture of our mechanism. First, the agent's degrees of belief are probabilistic. Second, the mechanism is consistent. And third, the mechanism generalizes the most common non-probabilistic model of an agent's beliefs, the modal logic *KD45* [9, 10]. In particular, the agent fully believes all logical consequences of his initial accepted collection of beliefs (which includes all of KB), and his full beliefs are closed under positive and negative introspection. In accord with the *KD45* model, his beliefs are not, however, necessarily accurate (although the agent thinks that the probability of his full beliefs being false is zero).

Examples

With these theoretical guarantees we can proceed to examine the nature of the inferences generated by direct inference through some examples. As will be become clear in this section, direct inference sanctions conclusions that are very similar to those conclusions sought after in default reasoning. We only have space for a couple examples, but see [8] for the detailed working out of many other examples.

The direct inference principle requires us to condition on the entire knowledge base. In actual applications this store of knowledge can be very large. The following equality is useful for reducing the conditioning formula.

* $[\alpha|\beta \wedge \lambda]_{\vec{x}} = [\alpha|\beta]_{\vec{x}}$, if no $x_i \in \vec{x}$ which appears in $\alpha \wedge \beta$ is free in λ.[7]

This equality is valid, i.e., true in every possible world in every model. Hence, it holds even when inside of a cert context. The equality follows from our use of a product measure. It says that the chance of a vector of individuals satisfying α will be influenced only by their known properties and by properties of other individuals that are known to be related them. For example, the fact that Fido is a dog will have no influence on the chance of Tweety being a flier.

[7] Freedom and bondage is extended in our language to include the clause that the random designators are bound. Also it can be proved that we can rename the random designators without changing the values of the statistical terms [11].

As our first example, consider Tweety the bird and Opus the penguin. We will use c as a numeric constant with any denotation greater than 0.5.

Example 1 [Tweety, but not Opus, flies] Let KB =

$$\mathtt{bird(Tweety)} \wedge \mathtt{bird(Opus)} \wedge \mathtt{penguin(Opus)}$$
$$\wedge[\mathtt{fly(x)}|\mathtt{bird(x)}]_{\mathtt{x}} > c \wedge \forall \mathtt{x}.\mathtt{penguin(x)} \rightarrow \mathtt{bird(x)}$$
$$\wedge[\mathtt{fly(x)}|\mathtt{penguin(x)}]_{\mathtt{x}} < 1 - c.$$

$\mathtt{KB}^{\mathbf{V}}$ differs from KB only with respect to the first three conjuncts (none of the others have constant symbols). We let $\mathtt{v}_1$ replace Tweety and $\mathtt{v}_2$ replace Opus.

First we can examine prob(Tweety):

$$\mathsf{prob}(\mathtt{fly(Tweety)}) = \mathsf{E}([\mathtt{fly}(\mathtt{v}_1)|\mathtt{KB}^{\mathbf{V}}]_{\langle \mathtt{v}_1,\mathtt{v}_2\rangle}) \quad \text{(a)}$$
$$\mathsf{cert}([\mathtt{fly}(\mathtt{v}_1)|\mathtt{KB}^{\mathbf{V}}]_{\langle \mathtt{v}_1,\mathtt{v}_2\rangle} = [\mathtt{fly}(\mathtt{v}_1)|\mathtt{bird}(\mathtt{v}_1)]_{\mathtt{v}_1}) \quad \text{(b)}^{8}$$
$$\mathsf{E}([\mathtt{fly}(\mathtt{v}_1)|\mathtt{KB}^{\mathbf{V}}]_{\langle \mathtt{v}_1,\mathtt{v}_2\rangle} = [\mathtt{fly}(\mathtt{v}_1)|\mathtt{bird}(\mathtt{v}_1)]_{\mathtt{v}_1}) \quad \text{(c)}$$
$$\mathsf{cert}([\mathtt{fly}(\mathtt{v}_1)|\mathtt{bird}(\mathtt{v}_1)]_{\mathtt{v}_1} > c) \quad \text{(d)}$$
$$\mathsf{E}([\mathtt{fly}(\mathtt{v}_1)|\mathtt{bird}(\mathtt{v}_1)]_{\mathtt{v}_1}) > c \quad \text{(e)}$$
$$\mathsf{prob}(\mathtt{fly(Tweety)}) > c$$

Equation (a) is the relevant instance of the direct inference principle; (b) is the result of applying equation *; (c) and (e) follow from the relationship between certainty and expectation;[9] and (d) is a consequence of Theorem 1.3.

The general form of the derivation is applicable to most examples. First, we apply the particular instance of the direct inference principle we are interested in, using equation * to remove irrelevant parts of KB. These parts include formulas with no constants and formulas with constants that stand in no relationship with the constants in the formula of interest (in this first example, the constant Opus). We can then proceed to reason about the value of the resulting expectation term in the agent's base theory by examining the statistical knowledge in KB. All of this knowledge (and every other deductive consequence of KB) will be certain in the agent's base theory.

The case for Opus also yields the expected result that $\mathsf{prob}(\mathtt{fly(Opus)}) < 1-c$. This follows from the fact that penguins are known to be a subset of birds. Therefore, the chance that a randomly selected penguin-bird is a flier will be equal to the chance that a randomly selected penguin is a flier: these two sets are identical. This example shows that there is a natural subset preference between defaults encoded as statistical assertions.

[8] This inference also depends on the fact that $[\alpha]_{\langle x,y\rangle} = [\alpha]_x$ if y does not appear free in α. This fact is an easy consequence of our semantics: if y is not free its instantiation will not matter.

[9] Clearly if two terms have equal value in every world of non-zero probability (i.e., are certainly equal), then their expected values over the worlds must also be equal. Similar comments hold for the cases of their values being certainly related by $>$, $<$, $\geq$, or $\leq$.

This preference is a direct consequence of the semantics of such an encoding. That is, there is no need for a meta-logical preference criterion under the statistical interpretation.

Example 2 [Clyde likes Tony but not Fred.] Let KB =

$$\begin{aligned}&\texttt{elephant(Clyde)} \;\wedge\; \texttt{zookeeper(Tony)}\\ \wedge\;& \texttt{zookeeper(Fred)}\\ \wedge\;& [\texttt{likes(x,y)}|\texttt{elephant(x)} \wedge \texttt{zookeeper(y)}]_{\langle \texttt{x,y}\rangle} > c\\ \wedge\;& [\texttt{likes(x,Fred)}|\texttt{elephant(x)}]_{\texttt{x}} < 1-c\end{aligned}$$

Although there is no space for the details, this example also yields the expected results: it is likely that Clyde likes Tony, but unlikely that he likes Fred. What happens here is that since the defaults are not separated from the rest of the knowledge, i.e., they are represented as formulas of KB, we can condition on specialized defaults, like the one particular to Fred. This has the desired effect on the resulting degree of belief. Hence, T_0 has a natural "specificity" preference when dealing with defaults specific to a particular individual.

Statistically Encoded Defaults

By encoding defaults as qualitative statistical assertions we have demonstrated the similarity between the style of inference generated by our system of direct inference and that sought after by default inference. Others have argued in favor of a statistical interpretation for defaults [12, 5] and have discussed the role of probabilities in nonmonotonic reasoning [13, 14], but such an encoding of defaults remains controversial. It does, however, match our basic intuitions that defaults are rules of thumb about what is usually, but not always, the case: this is precisely the semantics of a statistical assertion of majority.

It is often claimed that there are many other, non-probabilistic, notions involved in the meaning of defaults. Unfortunately, this position is difficult to argue for or against, as very little work has been done on specifying exactly what defaults are. Our current understanding of defaults remains mostly a collection of intuitions as to how they should behave, and typically systems of default reasoning are evaluated by their behavior rather than by any argument that they are founded on a reasonable semantics for defaults.

I will not claim that a statistical interpretation provides a complete picture of the their semantics, but I would argue that such an interpretation provides an *essential* component of their meaning (in my opinion, the major component). Thus, to specify this component precisely, as is the goal of this work, is an essential step in producing a complete specification. I will give two arguments to support my position. It should be noted, however, that I view default reasoning as reasoning that is used to draw plausible or useful conclusions that are not sound in the strict logical sense. There are many other sources of nonmonotonicity in an agent's reasoning, e.g., changing the theory in which reasoning is being performed. My claims about the statistical interpretation only pertain to default reasoning, not to all forms of nonmonotonic reasoning (although my intuition is that statistics play an important role in these other forms as well).

The strongest argument in favor of the statistical interpretation comes from its flexibility and how well it obeys the afore mentioned intuitions about how defaults should behave. See [8] for precise details as to how the behaviors describe below follow from the statistical semantics. Besides imparting the natural subset and specificity preferences the statistical interpretation and our system of direct inference includes the following features:

1. Through its assignment of degrees of belief, the system formally distinguishes default and deductive inferences (Theorem 1.3). As a consequence it avoids the paradoxes of unlimited scope in nonmonotonic reasoning [15] which include the lottery paradox [16].
2. The statistical semantics allows us to reason naturally with the defaults. For example, we can conclude that certain defaults are contradictory, e.g., "birds typically fly" contradicts "birds typically don't fly." In fact, all of the reasoning with defaults (and more) that make conditional logic approaches attractive [17] can be duplicated within the statistical semantics.
3. Contraposition is handled correctly: we are not forced to contrapose defaults. With some defaults like "birds typically fly" the contraposition "non-fliers are typically not birds" is a reasonable inference, but with others it is clearly counter-intuitive, e.g., from "living things are typically not fliers" to "fliers are typically not living." Using the statistical semantics one can demonstrate that in the case of birds, contraposition is sanctioned by the fact that birds are more likely to be fliers than non-birds. Living things, on the other hand, are not more likely to be non-fliers than non-living things; hence contraposition is blocked in this case.
4. The system is capable of reasoning with many different forms of statistical information. For example, instead of encoding defaults as statements of statistical majority, as in the examples above, one could encode them as qualitative influences which increase probability, [18, 19]. Each form sanctions a different collection of inferences. Furthermore, by including in the specification statements of conditional independence (easily expressed in the logic) additional inferences can be captured. For example, statements of statistical majority cannot be chained [20], but with conditional independences weak chaining can occur where the strength of the conclusion diminishes as more defaults are used. Intuitively this makes, to me, more sense than systems that do not distinguish between inferences that rely on a very large number

of defaults and those that only require one or two defaults, as each use of a default represents the introduction of a possible error.

As an aside it is interesting to contrast these behaviors with that of ϵ-probabilities [21]. The approach of ϵ-probabilities does not appeal to a statistical semantics; rather, it is more of an attempt to duplicate traditional nonmonotonic approaches using probabilistic notions. It has the advantages that it retains the context sensitivity of probability thus capturing the subset preference, and it allows some reasoning with the defaults. However, by forcing all probabilities to go to the limit it gives up the ability to assign degrees of belief, thus falling prey to the lottery paradox; it forces contraposition in all cases, intuitive or not; it lacks the ability to deal with different forms of probabilistic information, e.g., while the formalism outlined above can include exact qualitative evidential reasoning ϵ-probabilities forces one to "switch logics" [14]; and finally, it gives up the relatively clear notion of statistics for a more obscure notion of probabilities infinitesimally removed from 1.

The second argument for the statistical interpretation comes from an appealing view of default reasoning as rational inference, put forward by Doyle and Wellman [22]. In this view an agent adopts a belief (draws a default conclusion) if the expected utility of holding it exceeds the expected utility of not holding it. Under such a view it is very natural that utility should be divided into separate notions of cost and likelihood, and that the notion of likelihood required would be a graded one (even if roughly graded). For example, accepting a falsehood yields a tremendous decrease in cost of reasoning—subsequent reasoning in the inconsistent knowledge base can be done by always answering yes, a constant time procedure—but it has zero utility precisely because the likelihood of correctness is zero.

An important reason for separating cost and likelihood is that these notions can be fairly independent. Computational cost will not usually be affected by the acceptance or rejection of a plausible conclusion, rather it will be more affected by the acceptance or rejection of theories. For example, for reasons of computational cost an agent might decide to accept a Horn theory description of a domain rather than a more accurate first-order description. However, within the Horn theory he may still need to infer a new Horn formula that is plausible but does not follow deductively.[10]

A statistical approach using direct inference provides these grades of likelihood in a natural way. Furthermore, these grades, degrees of belief, are probabilistic, and thus they lead naturally to well developed models of expected utility. It is quite possible that statistical information could be compiled into more traditional default rules that carry an index of caution with them. This index, derived from the statistical information, would be used to indicate under what conditions of risk the default could be used (cf. [23]). Under such a scheme most of the behavior of these rules, like subset preference, would follow directly from the underlying statistical semantics.

[10] If computational cost is solely determined by acceptance or rejection of theories, which is a source of nonmonotonicity in reasoning, then default reasoning, circumscribed to be the generation of plausible inferences within a particular theory, would be solely determined by likelihood. That is, one could argue that although computational cost is important in general nonmonotonic reasoning, default reasoning is purely based on likelihood.

Conclusions

We have pointed out that direct inference is a useful form of non-deductive inference that takes us from statistics to predictions about particular cases. And we have provided a formal mechanism for performing this kind of inference. Our system makes clear the important similarities between reasoning from statistics in this manner and default reasoning as studied in AI.

An issue we have not addressed is what has been called the conditional interpretation [24]. This is the problem that the system is unable to reason to default conclusions in the face of irrelevant information, much like other approaches that adopt a conditional interpretation (e.g., ϵ-probabilities and conditional logics).[11] For example, if we know that `Tweety` is a yellow bird, we cannot use our statistics about birds; rather, we will require statistics about yellow birds. The system cannot conclude that yellowness is irrelevant without explicit statistical information.

In [8] a system is developed that allows the agent to extend his base theory. These extensions are generated as in Reiter's default logic [25] by adding explicit assertions about statistical irrelevance. Thus it uses a tradition approach to non-monotonic reasoning in a non-traditional manner: to make default assumptions about relevance and irrelevance rather than the final default conclusions. This preserves the statistical semantics of the ordinary defaults thus retaining the reasonable constraints imposed by those semantics.

In joint work a more satisfying approach is under investigation [26]. This work approaches the problem semantically rather than syntactically, building models in which all possible statistical independencies hold. It is related to other principles of maximal independence like maximum-entropy, but it has the advantage of providing a fairly natural semantic construction which captures the reasonable assumptions of irrelevance that the agent can make.

[11] Actually the issue of irrelevance appears in other approaches to default reasoning, e.g., default logic, autoepistemic logic, circumscription. These approaches include implicit, built-in, assumptions of irrelevance. These built-in assumptions, however, may not always be appropriate.

References

[1] Hans Reichenbach. *Theory of Probability*. University of California Press, Berkeley, 1949.

[2] Henry E. Kyburg, Jr. *The Logical Foundations of Statistical Inference*. D. Reidel, Dordrecht, Netherlands, 1974.

[3] John L. Pollock. *Nomic Probability and the Foundations of Induction*. Oxford University Press, Oxford, 1990.

[4] Isaac Levi. *The Enterprise of Knowledge*. MIT-Press, Cambridge, Massachusetts, Cambridge, Massachusetts, 1980.

[5] Ronald P. Loui. *Theory and Computation of Uncertain Inference and Decision*. PhD thesis, The University of Rochester, September 1987.

[6] Fahiem Bacchus. Lp, a logic for representing and reasoning with statistical knowledge. *Computational Intelligence*, 6(4):209–231, 1990.

[7] Joseph Y. Halpern. An analysis of first-order logics of probability. In *Proc. International Joint Conference on Artifical Intelligence (IJCAI)*, pages 1375–1381, 1989.

[8] Fahiem Bacchus. *Representing and Reasoning With Probabilistic Knowledge*. MIT-Press, Cambridge, Massachusetts, 1990.

[9] Ronald Fagin and Joseph Y. Halpern. Belief, awareness and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[10] Hector J. Levesque. A logic of implicit and explicit belief. In *Proc. AAAI National Conference*, pages 198–202, 1984.

[11] Fahiem Bacchus. *Representing and Reasoning With Probabilistic Knowledge*. PhD thesis, The University of Alberta, 1988.

[12] Henry E. Kyburg, Jr. Objective probabilities. In *Proc. International Joint Conference on Artifical Intelligence (IJCAI)*, pages 902–904, 1987.

[13] Benjamin N. Grosof. Non-monotonicity in probabilistic reasoning. In L. N. Kanal and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence Vol I*, pages 91–98. North-Holland, Amsterdam, 1986.

[14] Judea Pearl. Probabilistic semantics for nonmonotonic reasoning: A survey. In Brachman et al. [27], pages 505–516.

[15] David W. Etherington, Sarit Kraus, and Donald Perlis. Nonmonotonicity and the scope of reasoning. In *Proc. AAAI National Conference*, pages 600–607, 1990.

[16] David Poole. What the lottery paradox tells us about default reasoning. In Brachman et al. [27], pages 333–340.

[17] James P. Delgrande. A first-order conditional logic for prototypical properties. *Artificial Intelligence*, 33:105–130, 1987.

[18] Michael P. Wellman. *Formulation of Tradeoffs in Planning Under Uncertainty*. Research Notes in Artificial Intelligence. Morgan Kaufmann, San Mateo, California, 1990.

[19] Eric Neufeld. Defaults and probabilities; extensions and coherence. In Brachman et al. [27], pages 312–323.

[20] Fahiem Bacchus. A modest, but semantically well founded, inheritance reasoner. In *Proc. International Joint Conference on Artifical Intelligence (IJCAI)*, 1989.

[21] Hector Geffner and Judea Pearl. A framework for reasoning with defaults. In Henry E. Kyburg, Jr., Ronald Loui, and Greg Carlson, editors, *Knowledge Representation and Defeasible Inference*. Kluwer Academic Press, London, 1989.

[22] Jon Doyle and Michael P. Wellman. Impediments to universal preference-based default theories. In Brachman et al. [27], pages 94–102.

[23] Henry E. Kyburg, Jr. Full beliefs. *Theory and Decision*, 25:137–162, 1988.

[24] Hector Geffner. Conditional entailemnt: Closing the gap between defaults and conditionals. In *AAAI Workshop on Nonmonotonic Reasoning (unpublished collection of articles)*, pages 58–72, 1990.

[25] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[26] Fahiem Bacchus, Joseph Halpern, Adam Grove, and Daphne Koller. From statistics to beliefs, 1991. In preparation.

[27] Ronald J. Brachman, Hector J. Levesque, and Raymond Reiter, editors. *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*. Morgan Kaufmann, San Mateo, California, 1989.