Logic, probability theory, and artificial intelligence — Part I: the probabilistic foundations of logic

CHARLES G. MORGAN

Department of Philosophy, University of Victoria, Victoria, B.C., Canada V8W 3P4

Received September 14, 1988

Revision accepted March 5, 1991

Many AI researchers have come to be dissatisfied with approaches to their discipline based on formal logic. Various alternatives are often suggested, including probability theory. This paper investigates the intimate connection between probability theory and various logics. We show that probability theory, broadly conceived, may be used as a formal semantics for virtually any monotonic logic. Thus, rather than being seen as competing, it is more appropriate to view formal logics as very special cases of probability theory, usually special cases that are computationally more tractable than the more general theory. Thus, probability theory and logic should be seen as complementary. Viewing probability theory in this abstract way may help to shed light on various recalcitrant problems in AI.

Key words: probability, semantics, logic, artificial intelligence.

De nombreux chercheurs dans le domaine de l'intelligence artificielle manifestent une certaine insatisfaction vis-à-vis certaines approches basées sur la logique formelle. Diverses solutions sont souvent proposées, y compris la théorie des probabilités. Cet article analyse la relation intime entre la théorie des probabilités et diverses logiques. Il est démontré que la théorie des probabilités, conçue de manière générale, peut être utilisée comme une sémantique formelle pour presque toute logique monotonique. Au lieu de percevoir les logiques formelles comme étant en opposition, il est plus approprié de les considérer comme des cas très spéciaux de la théorie des probabilités, habituellement plus traitables au niveau calcul que la théorie plus générale. Par conséquent, la théorie des probabilités et la logique doivent être perçues comme des éléments complémentaires. Le fait de considérer la théorie des probabilités d'une manière abstraite peut contribuer à la compréhension de divers problèmes ardus dans le domaine de l'intelligence artificielle.

Mots clés : probabilité, sémantique, logique, intelligence artificielle.

[Traduit par la rédaction]

Comput. Intell. 7, 94–109 (1991)

1. Introduction

There is a quite general perception in many quarters that research in AI has reached a plateau. Remarkable results are becoming more difficult to achieve, and many long-standing problems remain unsolved. In many cases, what appeared to be promising avenues have petered out into rabbit trails through extremely dense bush. This plateau phenomenon occurs during the early stages of major shifts in any discipline and is not unique to AI. Whenever a plateau is reached, it is common practice to search for new tools with which to attack recalcitrant problems. For at least the past decade, one of the primary tools of AI research has been some form of classical first-order predicate calculus. However, severe difficulties have been encountered by researchers in AI in trying to use the tools of classical logic to model ordinary knowledge and commonsense inferences.

Many philosophers would be surprised that AI researchers are surprised by the problems encountered in this line of endeavor. Many of the problems which have recently come to the attention of the AI community have been under examination by philosophers for many years (tens, hundreds, and, for some problems, thousands of years). The philosophical community has known for many years about the inadequacy of first-order predicate calculus for the modeling of many epistemological phenomena. The computer simply provides a convenient concrete base for testing various theories concerning knowledge acquisition and commonsense inference; in itself, computer technology does not provide any great theoretical insights. And new theoretical insights are what is required to solve the recalcitrant problems.

McDermott (1987) has reviewed some of the difficulties in what he calls the "logicist" approach. He ends on a rather pessimistic note stating that "we must resign ourselves to writing programs, and viewing knowledge representations as entities to be manipulated by the programs." I would reiterate that before successful programs can be written, theoretical insight must be achieved. We will not solve serious problems in AI by becoming better programmers, but rather by becoming better theoreticians. I do not mean to suggest that McDermott himself does not recognize the importance of theoretical insights; no doubt he simply meant to suggest that the approach of formal logic may not be the best way to implement such insights.

In a *theoretical* sense, we know that the McDermott response cannot be correct. We know that any program on any computer of current technological design can be modelled by a Turing machine, and the behaviour of any Turing machine can be described by a suitably complicated first-order theory. So, if it is possible to write a successful program to perform a given task, then there is a first-order theory for the task. Now, just as we could patch together any finite set of programs with a simple selection menu, so we can obtain a single first-order theory corresponding to any finite number of programs. So, if we followed McDermott and developed a number of problem-specific programs, there would always be a first-order theory for the collection of programs. These comments should *not* be taken to suggest that classical logic is the only appropriate theoretical environment for the solution to problems in AI. In fact, in the material below, I will suggest quite the contrary.

Printed in Canada / Imprimé au Canada

In spite of these theoretical assurances, we should be prepared to be foiled by the demon of computational complexity. It may well be that the simplest theory for accomplishing a desired task is beyond the time and space bounds of our mental machinery. The problem with complexity considerations is that it is difficult to determine lower complexity bounds for interesting problems, and upper bounds can generally be established only after the development of a successful algorithm. So, before we know that we should give up on a certain problem because of complexity considerations, we need to have a theoretical solution for that problem! Thus, unless a problem can be shown to be undecidable, we may unknowingly continue to beat our heads against a wall of complexity. Certainly we can agree that the approach of attempting to obtain a first-order theory from the behaviour of a Turing machine in the manner theoretically suggested above is doomed to failure. Indeed, the mathematical existence of a first-order theory corresponding to the behaviour of an arbitrarily complex Turing machine gives very cold comfort at best. In any reasonable case, such a theory would be much too complex for humans to discover or use. Assuming a standard sheet of paper using standard print technology contains 10K or even 100K bits, it is easy to see that such a theory may well require a mass of paper greater than that produced so far in the history of our planet. So, the practical import of McDermott's comments are not to be lightly dismissed.

The responses by Woods (1987) and by Cheeseman (1988) to McDermott's pessimism are refreshing because of the relatively new approaches that they advocate. My personal sympathies are with a probabilistic approach. I very strongly feel that in the future, AI will more and more come to be dominated by what may be termed probabilistic considerations. Thus, in broad outline, I am in sympathy with Cheeseman's approach. However, in many respects I believe his characterization is misleading and much too narrowly conceived, as was indicated in Morgan (1988).

This paper is part one of a larger two-part work. I have several goals in this part. I would like to demonstrate how a formal theory of probability may be used as a metalanguage for virtually any (monotonic) logic. Thus probability theory may correctly be regarded as a very strong generalization of standard logics. Natural languages have evolved to be powerful enough to serve as their own metalanguages. However, we know that formal languages with such power lead to formal inconsistencies. We still do not have a good theory of how users of natural languages manage to cope with such potential inconsistency. In any case, since probability may be used as a metalanguage for any logic, it has much greater potential than standard logics for being able to serve as the basis for a formal analysis of many natural language locutions and inferences, including many that appear to be metalinguistic in nature. Classical first-order logic turns out to be just a very special case of probability theory, namely the case in which the values are restricted to 0 and 1. The special case is computationally more tractable than the more general theory, but only because it forces us to view the world in a very narrow way. In light of these facts, it is a serious mistake to regard logic and probability as being in any way antagonistic. I will indicate why the standard numerically based theories of probability may be inappropriate for many contexts, and I will also indicate alternatives which may be just as useful for many applications in AI as the classical theory.

In the next paper, I will use the general probabilistic framework to indicate why I believe the current interest in so-called nonmonotonic logic is at best misdirected. I will show how problems that prompted the nonmonotonic work may be handled by standard monotonic logics. The approch I advocate illuminates the source of many of the problems faced by devotees of "nonmonotonic logic."

2. Conditionalizing classical probability theory

The most important notion in standard logics is the concept of a proof or derivation from a given set of assumptions; we call this notion "syntactic entailment." Syntactic entailment is usually symbolized as $\Gamma \vdash E$, where Γ is a set of expressions from the formal language and E is a single expression; of course, Γ is the set of assumptions and E is the conclusion of the derivation. A formal semantic theory for a given logic must allow definition of a corresponding notion of semantic entailment, which is generally symbolized by $\Gamma \models E$. It is important to recognize that both syntactic and semantic entailments are conditionals from the metalanguage and not a part of the object language. Since the conditional of conditional probability theory is also part of the metalanguage rather than part of the object language to which the probability measure is applied, it does not require great insight to suspect that our probabilistic account of semantic entailment will crucially involve conditional probability theory. Thus our initial concern is to develop a coherent account of conditional probability.

Komolgoroff (1950) formulated the elementary theory of classical probability in the following way:

KP.1. P is defined on a σ-field of sets.
KP.2. $0 \leq P(\alpha)$.
KP.3. $P(U) = 1$.
KP.4. If $\alpha \cap \beta = \emptyset$ then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$.

For our purposes, a σ-field may be defined to be a set of sets; the σ-field must have as elements both the empty set $\emptyset$ and the universal set U, and it must be closed under (finite) unions, intersections, and complements. A more parsimonious definition is possible but of no interest in the present discussion. Komolgoroff and a host of others speak of probability as being defined over a σ-field of sets of *events*. Intuitive examples of events are draws of cards, rolls of dice, and deaths of male Caucasians over the age of 40. We may think of information about the world as constraining the set of events. Thus zero information corresponds to the universal set of events, while contradictory information corresponds to the empty set of events.

Of extreme interest for us will be Komolgoroff's definition of conditional probability.

Definition DKP.1
If $P(\beta) \neq 0$ then $P(\alpha, \beta) = P(\alpha \cap \beta)/P(\beta)$.

The symbolism $P(\alpha, \beta)$ is read "the probability of α on the condition β," or, more simply, "the probability of α *given* β." We will sometimes refer to the position occupied by β as the "assumption" position and sometimes we will refer to it as the "evidence" position. We will refer to the position occupied by α as the conclusion position. We will use the phrase "Komolgoroff classical probability theory" to refer to KP.1–KP.4 plus the definition of conditional probability. It is well known that from these simple principles, a very rich account of probability may be derived.

Of particular interest to us will be the following standard theorems:

Theorem TKP.1
$P(\alpha) \leq 1$.

Theorem TKP.2
$P(\alpha \cup \beta) = P(\alpha) + P(\beta) - P(\alpha \cap \beta)$.

Theorem TKP.3
$P(\alpha^c) = 1 - P(\alpha)$.

Note that we are using α^c to designate the set-theoretic complement of α.

There is a slight problem with conditional probability theory (as given by definition DKP.1) for the case when the probability of the conditioning set is 0. What should we take $P(\alpha, \beta)$ to be when $P(\beta) = 0$? One standard (but naive) answer is just to define the conditional probability to be 1 in such cases. However, if we adopt this course, then the conditional version of KP.4 would fail. This result is unfortunate, since we want our conditional probability measure to be a probability measure. Another standard suggestion is to say that the conditional probability is simply undefined when the conditioning set has measure 0. This suggestion is also somewhat unsatisfactory, since the conditional probability functions would not be defined over the entire σ-field. At best, this result would require severe hedging of many important theorems.

There is another way out of the problem of conditioning on sets of measure 0; the technique is essentially due to Popper (1965). Although he did not formulate his conditional probability theory in the way we will here, we will nonetheless call the following constraints Popper classical conditional probability theory:

PP.1. P is defined on ordered pairs from a σ-field of sets.
PP.2. $0 \leq P(\alpha, \beta) \leq 1$.
PP.3. $P(U, \alpha) = 1$.
PP.4. If $\alpha \cap \beta = \emptyset$ then $P(\alpha \cup \beta, \gamma) = P(\alpha, \gamma) + P(\beta, \gamma)$ unless for all δ, $P(\delta, \gamma) = 1$.
PP.5. $P(\alpha \cap \beta, \gamma) = P(\alpha, \gamma) \times P(\beta, \alpha \cap \gamma)$.

Principles KP.1 and PP.1 directly correspond. For simplicity, we have incorporated the conditional version of theorem TKP.1 into PP.2. Principle PP.3 is just the conditional version of KP.3. Principle PP.5 is just the conditional version of the usual definition of conditional probability stated by Komolgoroff. (Of course, sets of constraints other than PP.1–PP.5 could be used to pick out the same functions. Our set was chosen because these constraints parallel nicely the Komolgoroff constraints KP.1–KP.4 and the definition of conditional probability DKP.1.)

The Popperian innovation is included in the "unless" clause of PP.4 and deserves some comment. The intuitive idea is that there are some "events" that are so bizarre that if asked to assume that they have occurred, I would be unable to reject anything. Standard examples of such events generally include (but may not be limited to) contradictory events like finding a cubical sphere or observing that it is both raining and not raining in a given spot at a given time. Sets containing such events are said to be "abnormal." We formally adopt the following definition:

Definition DPP.1
A set γ is said to be *P-abnormal* on Popper classical conditional probability distribution P if and only if for all δ, $P(\delta, \gamma) = 1$. A set γ is *P-normal* if and only if it is not P-abnormal.

Except when stating theorems, we will simply use "normal" and "abnormal" instead of "P-normal" and "P-abnormal." It is useful to point out that the constant function assigning all pairs the value 1 satisfies constraints PP.1–PP.5. This function corresponds to the limiting case in which all sets are abnormal. The universal set occupies a rather special role with respect to abnormality, as stated in the following theorem:

Theorem TPP.1
If the universal set is P-abnormal, then every set is P-abnormal.

Except for the extreme situation in which all sets are abnormal, it is not difficult to show that the abnormal sets all take probability 0 given the universal set, i.e., they take probability 0 on the basis of zero information about the world.

Theorem TPP.2
For any set α, if α is P-abnormal then $P(\alpha, U) = 0$, provided that there is at least one P-normal set.

The conditional in theorem TPP.2 cannot be replaced by a biconditional. That is, there may be events that are initially assigned a probability of 0, even though they are not abnormal. We will look at a simple example below in Sect. 6. Very briefly, the example concerns a simple coin toss experiment. It is not unreasonable to assign a probability of 0 to the statement that the tossed coin came to rest on edge. However, given the statement that the tossed coin has come to rest on edge, I would assign probability of 0 to the claim that on the same toss the coin had come to rest heads up; hence the statement that the tossed coin came to rest on edge is not abnormal.

For our purposes, there are a number of additional theorems which will be quite important. We will simply list them and then say a brief word about each.

Theorem TPP.3
$P(\alpha \cup \beta, \gamma) = P(\alpha, \gamma) + P(\beta, \gamma) - P(\alpha \cap \beta, \gamma)$.

Theorem TPP.4
$P(\alpha^c, \beta) = 1 - P(\alpha, \beta)$ provided β is P-normal.

Theorem TPP.5
If $\beta \subseteq \alpha$ then $P(\alpha, \beta) = 1$.

Theorem TPP.3 is just the conditionalized version of theorem TKP.2 and states the general "sum" rule. Note that we cannot simply replace restriction PP.4 by theorem TPP.3, as the resulting theory would be substantially weaker. A simple counter-example will suffice. Consider the σ-field of just U and $\emptyset$. Consider the function P^* which assigns the following values:

$P^*(\emptyset, U) = r, \quad 0 < r < 1$
$P^*(\emptyset, \emptyset) = 1$
$P^*(U, U) = 1$
$P^*(U, \emptyset) = 1$

This function satisfies PP.1, PP.2, PP.3, PP.5, and theorem TPP.3. However, it does not satisfy condition PP.4, since $P(U \cup \emptyset, U) \neq P(U, U) + P(\emptyset, U)$ although $P(\emptyset, U) \neq 1$.

Theorem TPP.4 corresponds to theorem TKP.3; note the inclusion of the requirement that the assumption set be normal, which results from the change from KP.4 to PP.4. Theorem TPP.5 is actually equivalent to principle PP.3, in the context of the other principles. Recall that the larger the set of events, the fewer the constraints placed on the universe; hence if $\beta \subseteq \alpha$ then the α-constraints are less restrictive than the β-constraints and in fact must be included in the β-constraints.

We will now outline a very general technique for the construction of functions that satisfy PP.1–PP.5. This technique will prove to be quite useful in our deliberations. In order to state the technique, we must first define a key concept:

Definition DPP.2

An *elementary weighting function* w on a σ-field of sets is any function mapping the field into the real numbers which satisfies the following:
(a) $w(\emptyset) = 0$,
(b) $w(\alpha \cup \beta) = w(\alpha) + w(\beta) - w(\alpha \cap \beta)$, and
(c) if $\alpha \subseteq \beta$ then $w(\alpha) \leq w(\beta)$.

At this point, it will be good for the reader to try to understand the intuitions behind the concept of elementary weighting function. In the first instance it is best to think of the σ-field as containing sets of universe designs; a universe design is just a total specification of one way the universe might be. Any statement may be compatible with many universe designs or perhaps with only a few. A logically absurd proposition would be compatible with no universe designs. Initially we may think of the weighting function as representing simple cardinality (normal counting). The number of universe designs in the empty set is of course 0, and hence constraint (a). For constraint (b), note that the number of universe designs in the union of two sets will be the sum of the numbers in the two sets separately, minus the number in the overlap; we must subtract the number in the overlap because they will have been counted twice, once when we count the number in the first set and once when we count the number in the second set. For constraint (c), note that the number of elements in a subset is never greater than the number in the superset. So the three constraints obviously hold if we are talking about counting elements. To generalize a bit, we could think of each element as itself being a representative of some other class. An element may represent only one item, while another may represent 552 items, and so on. So, we could think of the weighting function as counting the items represented. Because we wish to be able to handle infinite sets, we do not wish to restrict ourselves to simple counting, so we generalize to the notion of elementary weighting functions.

Carefully note that standard Komolgoroff probability functions are elementary weighting functions; however, not all elementary weighting functions will be Komolgoroff probability functions, since elementary weighting functions need not be bounded above by 1. We can use any well-ordered sequence of elementary weighting functions to define a conditional probability function by the technique described in the following theorem. We could give a slightly more general theorem concerning other sorts of ordered sequences of elementary weighting functions, but there is no need to do so in the present context. —

Theorem TPP.6

Let $\{w_1, w_2, \ldots\}$ be a well-ordered sequence of elementary weighting functions, all defined on the same σ-field, and let the function P be defined as follows:

$$P(\alpha, \beta) = 1, \quad \text{if } w_i(\beta) = 0 \text{ for all } w_i$$
$$= w_i(\alpha \cap \beta)/w_i(\beta), \quad \text{for the first } w_i \text{ such that } w_i(\beta) \neq 0$$

Then P so defined satisfies PP.1–PP.5.

The reader will appreciate part of the importance of theorem TPP.6 by bearing in mind the intuitions concerning the weighting functions discussed above and considering a "sequence" containing only one weighting function. Now, consider a simple situation covered by relative frequency considerations, such as rolling a standard die. In the usual treatments, we consider only six distinct possibilities, namely one for each of the faces of the die uppermost. We may think of these possibilities as classes of universe designs. For the purposes of our problem, we ignore all other aspects of the universe designs except which face of the die lands uppermost; that is, we act as though there really are only six distinct universe designs. If we believe the die is fair, we assign equal weight to all six possibilities; but if the die is weighted or shaved, we will assign weights to the six possibilities to reflect the biases in the die. We assess probability (e.g., the probability that the die is showing a 6, given that it is showing an even number) along relative frequency lines exactly in accord with theorem TPP.6. In fact, all simple relative frequency cases may be analyzed in this way. But agreement with simple relative frequency is not the only important point. As we will show by example a bit later, theorem TPP.6 allows us to "paste together" a complex conditional probability function from any sequence of elementary Komolgoroff probability functions to arrive at a more rational function than would be obtained by simply assigning 1 whenever the assumption set has initial probability 0.

Historically, probability theory has been regarded as at least one part of inductive logic. Further, inductive logic has been regarded as the poor second cousin of deductive logic; and there have been attempts to use classical deductive logic as a foundation for probability theory, the most notable example being Carnap (1950). However, more recent developments have shown that this traditional view may be turned on its head, so to speak, and that probability theory may be used as a foundation for deductive logic. We will now show how our conditional probability theory may be used as a formal semantics for almost any deductive logic.

3. Constraints on the logics

In our discussion from now on we will assume to be given a language $\mathcal{L}$ consisting of some countable set of expressions. We will use capital letters A, B, and so on to designate expressions in our language. We will use capital Greek letters Γ, Δ, and so on to designate sets of expressions. Unless specified explicitly, we will not assume that the language contains any particular connectives, quantifiers, or other syntactic structure, since we want our discussion to be as general as possible.

In order to be extremely general, we will assume that our logics are formulated in "sequent" style. A logic L for our language consists of a set of axioms and a set of inference

rules, which together will be used to define the notion of syntactic entailment; we use the standard notation $\Gamma \vdash_L A$ to mean that there is a derivation of expression A from the set of assumptions Γ using the axioms and inference rules of logic L. We will sometimes just write $\Gamma \vdash A$ when the logic in question is understood. We assume that our inference rules can all be stated in the following general way:

IR.i. If $\Gamma_1 \vdash A_1$ and ... and $\Gamma_j \vdash A_j$ then $\Gamma \vdash A$, where Γ, A, the Γ_i, and the A_i satisfy conditions COND.

We use COND to stand for any English sentence specifying special conditions which must be satisfied. For example, one standard version of the rule of universal quantifier introduction could be stated as follows:

IR.EX1. If $\Gamma \vdash A$, then $\Gamma \vdash (\forall x)A$, provided there are no free occurrences of the variable x in any member of Γ.

Of course these rules are stated in the metalanguage. The Γ_i and A_i may be either metalinguistic constants or variables. Metalinguistic variables are presumed to be universally quantified (e.g., "for all Γ and all A"). Note that this inference rule form will not in general be suitable for nonmonotonic inferences. (For one example, we may wish to make the derivation of A dependent on what things *cannot* be derived from the assumption set, as well as on what things can be derived from the assumption set. We will explicitly discuss nonmonotonic inferences in the next paper.) We do not really have to consider axioms separately, as an axiom AX_i is just a special case of an inference rule with vacuous antecedent:

IR.AX_i. $\Gamma \vdash AX_i$

There may be other inference rules with vacuous antecedents but that do not correspond to axioms, as in the following example:

IR.EX2. $\Gamma \cup \{A\} \vdash A$

A "sequent" consists of a set designator followed by the symbol $\vdash$ followed by a sentence designator. A derivation in a sequent logic consists of a finite ordered list of sequents, each one of which may be justified on the basis of previous members of the list (if any) and one of the inference rules. Of course, a derivation must begin with a rule with vacuous antecedent (e.g., an axiom) because there are no previous members of the list to use in the justification.

Instead of thinking of probability theory as applying to sets of events, we may instead think of it as a very general theory of models (in the logical sense of the term "model"). If we are ultimately to use probability theory as a formal semantics, we do not want to require some intervening traditional account of models. Standard completeness proofs almost always require the consideration of a "maximal" set of expressions at some stage of the argument. Taking our cue from this technique, we will adopt the view that our probability theory is to be applied to a σ-field of sets of "maximal" sets of expressions, where "maximal" is defined as follows:

Definition DL.1

Γ is *maximal with respect to A* if and only if both of the following are satisfied:

(a) not $\Gamma \vdash A$, and
(b) if not $\Gamma \cup \{B\} \vdash A$, then $B \in \Gamma$.

Γ is *maximal* if and only if there is some expression A with respect to which Γ is maximal.

To allow the proofs of some important theorems, we must place some restrictions on the possible logics under consideration. We will state the restrictions and then discuss each in turn.

R.1. If $A \in \Gamma$ then $\Gamma \vdash A$.
R.2. If $\Gamma \vdash A$ then for some finite subset Δ of Γ, $\Delta \vdash A$.
R.3. If $\Gamma \cup \{A\} \vdash B$ and $\Gamma \vdash A$ then $\Gamma \vdash B$.
R.4. If $\Gamma \vdash A$ then $\Gamma \cup \Delta \vdash A$.

Restriction R.1 simply says that any member of our assumption set is a theorem of that assumption set. Restriction R.2 is just the requirement of proof compactness; we do not permit infinitary rules of inference. Restriction R.3 is sort of a cut rule; it allows us to think of proofs as proceeding by adding further lines to existing proofs. Restriction R.4 is the requirement of monotonicity. We should note that (almost) all standard logics satisfy R.1–R.4.

Each logic will be determined by its own formation rules, axioms, and inference rules. Of course, changes in logic will result in changes in the collection of maximal sets. We will use L to stand for an arbitrary logic satisfying R.1–R.4. Given these four restrictions and the fact that our language is denumerable, the following two important theorems about L are easily established.

Theorem TL.1

If $\Gamma \vdash A$ then for every maximal superset Δ of Γ, $A \in \Delta$.

Theorem TL.2

If not $\Gamma \vdash A$ then there is a superset Δ of Γ which is maximal with respect to A.

Theorem TL.1 follows directly from R.3 and R.4. The proof of theorem TL.2 requires R.1 and R.2 in addition, and directly parallels the standard proof of Lindenbaum's theorem using Zorn's lemma; see Robinson (1974), for example. We need not really be restricted to denumerable languages, as theorem TL.2 can be proved for any well-ordered language; see Chang and Keisler (1973).

Intuitively, a maximal set of expressions says everything it is possible to say without being inconsistent (relative to the logic L); that is, each maximal set specifies one way the universe might be. So we may think of maximal sets as alternative universe designs, where every universe design must accord with the principles of logic L. A single expression E will correspond to a set (possibly empty) of maximal sets, namely all those of which E is a member; this set of maximal sets may be regarded as the collection of all universe designs in which E would hold. Similarly, a set Γ of expressions will pick out a set (possibly empty) of maximal sets of expressions, namely the set of maximal supersets of Γ; as before, the set of maximal supersets of Γ may be thought of as the collection of designs for universes in which all members of Γ would hold. With these intuitions, it should not be surprising that we can construct conditional probability functions by assigning weights to the sets of universe designs.

As a technical aside, note that sets of "universe designs" do not exactly correspond to possible worlds structures

familiar from modal logics. We assume no relations of "nearness" of "accessibility" between maximal sets or between sets of maximal sets as is done in possible worlds semantics. Nonetheless, our probabilistic techniques will easily accommodate any of the nonclassical logics for which possible worlds semantics can be formulated. But more importantly, our probabilistic account will easily serve even for those logics for which no possible worlds semantics is possible, such as those in Fine (1974) and Thomason (1974). For extensions of classical logic, these results will be established by the end of Sect. 5. For logics weaker than classical logic, these results will be established by the end of Sect. 7.

4. Probabilistic pseudo-semantics

Let U_M be the set of all maximal sets of the logic. Then the set of all subsets of U_M, designated by $\mathcal{P}(U_M)$ is a σ-field of sets. For an arbitrary set of expressions Γ, we use the notation $M(\Gamma)$ to stand for the set of all maximal supersets of Γ. Obviously $M(\Gamma) \in \mathcal{P}(U_M)$. By theorem TPP.6, we know that any sequence of elementary weighting functions on $\mathcal{P}(U_M)$ defines a conditional probability function satisfying PP.1–PP.5. Our goal is to use such probability functions as the basis of a formal semantics for L.

For a given probability function P defined on the σ-field $\mathcal{P}(U_M)$, we may isolate a corresponding probabilistic notion of entailment P_M as follows:

$$P_M(E, \Gamma) = P(M(\{E\}), M(\Gamma))$$

The function P_M gives the probability of expression E, given the assumption set Γ. It is tempting to think that a value of 1 for P_M corresponds to logical entailment, but that would not be correct for monotonic logics. As we will show in an example below (Sect. 6), it is possible to have $P_M(E, \Gamma) = 1$ but, for some Δ, to have $P_M(E, \Gamma \cup \Delta) = 0$, which would not correspond to proof theory. For E to be logically certain, given Γ, it is not enough that Γ just happens to make us certain of E; rather, it must be the case that given Γ, no additional evidence could ever cause us to doubt E. Thus we define our probabilistic notion of semantic entailment as follows:

Definition DL.2

The set of expressions Γ *semantically implies* expression E with respect to the maximal sets of logic L, symbolically $\Gamma \Vdash_{M(L)} E$, if and only if for every probability function P constructed on the basis of a sequence of elementary weighting functions on $\mathcal{P}(U_M)$, we have $P(M(\{E\}), M(\Gamma \cup \Delta)) = 1$ for every set of expressions Δ.

We earlier made the claim that conditional probability theory could be viewed as a general theory of models. Our claim is firmly established by the following two theorems:

Theorem TLP.1 (pseudo-soundness)
If $\Gamma \vdash_L E$ then $\Gamma \Vdash_{M(L)} E$.

Theorem TLP.2 (pseudo-completeness)
If $\Gamma \Vdash_{M(L)} E$ then $\Gamma \vdash_L E$.

The proof of theorem TLP.1 depends on theorem TL.1 in the obvious way. The proof of theorem TLP.2 follows the standard Henkin argument and depends on theorem TL.2. If we suppose that not $\Gamma \vdash_L E$, then we must construct a probability function so that for some Δ, $P(M(\{E\}), M(\Gamma \cup \Delta)) \neq 1$. We may take Δ to be empty. The required probability function may then be constructed from a single weighting function. We pick some superset of Γ which is maximal with respect to E, as guaranteed by theorem TL.2; the required weighting function assigns a weight of 1 (or any nonzero value) to all sets containing the chosen maximal superset and 0 to all other sets.

It is important to note that theorems TLP.1 and TLP.2 hold for *any* logic that satisfies restrictions R.1–R.4. The differences from one logic to another are buried in the notion of "maximal set." Different logics may have different syntactic components and hence different expressions; but even when two logics have the same expressions, two distinct logics will always define different collections of maximal sets. And if two logics determine different maximal sets, then they will be associated with different collections of probability functions, since the probability functions are derived from weighting functions over $\mathcal{P}(U_M)$. In addition to semantic entailment, we could use our probability theory to define other standard notions from formal semantics. Thus classical conditional probability theory may be regarded as a very general metalanguage in terms of which the important notions from formal semantics may be elaborated.

It is also important to realize that theorems TLP.1 and TLP.2 are not totally satisfactory from the standpoint of formal semantics. The problem is that our definition of semantic entailment, definition DL.2, is not autonomous; that is, the definition of semantic entailment involves notions of proof theory in a crucial way in so far as it depends on the notion of maximal sets. For many interesting cases, that dependency may be removed, as we will now demonstrate. Our ultimate goal will be to develop a probabilistic account of entailment in which our probability functions are defined over ordered pairs of the form (E, Γ), where E is an expression of $\mathfrak{L}$ and Γ is a set of expressions of $\mathfrak{L}$; Γ may be thought of as the premise set and E may be thought of as the conclusion.

5. Probabilistic semantics

In this section, we will assume that our language contains at least the syntactic machinery of classical propositional logic. In particular, we assume the syntax contains a monadic negation operator designated by $\sim$, a dyadic conjunction operator designated by $\wedge$, and a dyadic disjunction operator designated by $\vee$. Further, we assume that our proof theory is sufficiently rich to assure that these connectives are Boolean. To fully explain this assumption, note that we may interpret negation as complement, conjunction as meet, and disjunction as join; further, we may treat $p \vee \sim p$ as 1 (identity with respect to meet), and treat $p \wedge \sim p$ as 0 (identity with respect to join). We assume that if $A = 1$ is a theorem of Boolean algebra, then $\Gamma \vdash_L A$ for all Γ; further, if $A = B$ is a theorem of Boolean algebra, then for all Γ, $\Gamma \vdash_L A$ if and only if $\Gamma \vdash_L B$. Under these assumptions, there is a very tight relationship between the Boolean connectives and set-theoretic operations on maximal sets, as indicated in the following:

B.1. $M(\{A\}) \cup M(\{B\}) = M(\{A \vee B\})$.
B.2. $M(\{A\}) \cap M(\{B\}) = M(\{A \wedge B\})$.
B.3. $M(\Gamma) \cap M(\Delta) = M(\Gamma \cup \Delta)$.
B.4. $M(\{A\})^c = M(\{\sim A\})$.
B.5. If $\Gamma \subseteq \Delta$ then $M(\Delta) \subseteq M(\Gamma)$.

Using these relationships, we can convert our conditional probability theory from a theory that applies to maximal sets of expressions to a theory that applies directly to the language itself. The following set of contraints defines what we will call neo-classical conditional probability theory.

NP.1. $0 \leq P(A, \Gamma) \leq 1$.
NP.2. If $A \in \Gamma$ then $P(A, \Gamma) = 1$.
NP.3. $P(A \vee B, \Gamma)$
$= P(A, \Gamma) + P(B, \Gamma) - P(A \wedge B, \Gamma)$.
NP.4. $P(A \wedge B, \Gamma) = P(A, \Gamma) \times P(B, \Gamma \cup \{A\})$.
NP.5. $P(-A, \Gamma) = 1 - P(A, \Gamma)$ provided Γ is P-normal.
NP.6. $P(A \wedge B, \Gamma) = P(B \wedge A, \Gamma)$.
NP.7. $P(C, \Gamma \cup \{A \wedge B\}) = P(C, \Gamma \cup \{A, B\})$.

Our definition of "normal" is just what would be expected in light of our previous discussion.

Definition DNP.1

A set Γ is said to be *P-abnormal* on neo-classical conditional probability distribution P if and only if for all expressions E of language $\mathcal{L}$, $P(E, \Gamma) = 1$. A set Γ is *P-normal* if and only if it is not P-abnormal.

A few comments about each of our constraints may prove to be illuminating. If we make the assumption that the functions defined by PP.1–PP.5 are defined over $\mathcal{P}(U_M)$, then each of the constraints NP.1–NP.7 may be derived from our "Popper" constraints PP.1–PP.5 and the relations given by B.1–B.5. Of course NP.1 corresponds directly to PP.2. Constraint NP.2 follows directly from B.5 and theorem TPP.5; recall that theorem TPP.5 is equivalent to condition PP.3. Given B.1, NP.3 follows from theorem TPP.3. Given B.2 and B.3, NP.4 follows from PP.5. Given B.4, NP.5 follows from theorem TPP.4. Given B.2, NP.6 follows from PP.1. Finally, NP.7 follows directly from B.2 and B.3. Thus neo-classical conditional probability theory does not deviate from our Popper classical conditional probability theory.

Popper's theory (1965) is strikingly similar to our own. However, he formulated his probability theory over pairs of expressions from a Boolean language, corresponding to functions on $\mathcal{L} \times \mathcal{L}$, while our theory corresponds to functions on $\mathcal{L} \times \mathcal{P}(\mathcal{L})$. Popper did not begin with the Komolgoroff theory KP.1–KP.4 and then conditionalize it as we have in PP.1–PP.5. Nor did Popper show how his conditions could be derived from an appropriately conditionalized form of the Komolgoroff constraints via the consideration of maximal sets of sentences, as we have done here.

We wish to emphasize at this point that our constraints NP.1–NP.7 do not depend on notions from proof theory, and hence they are autonomous. Thus they form an acceptable basis for a true probabilistic semantics. Given only our constraints NP.1–NP.7, we may prove that all of the standard Boolean identities in conjunction, disjunction, and negation hold for expressions in the conclusion position. In addition, the following three important theorems are not difficult to prove.

Theorem TNP.1

If $P(A, \Gamma \cup \Delta) = P(B, \Gamma \cup \Delta)$ for all Δ, then $P(C, \Gamma \cup \Delta \cup \{A\}) = P(C, \Gamma \cup \Delta \cup \{B\})$ for all C and all Δ.

Theorem TNP.2

If $P(C, \Gamma \cup \{A\}) = P(C, \Gamma \cup \{B\})$ for all C, then $P(A, \Gamma) = P(B, \Gamma)$.

Theorem TNP.3

If $P(C, \Gamma \cup \Delta \cup \{A\}) = P(C, \Gamma \cup \Delta \cup \{B\})$ for all C and all Δ, then $P(A, \Gamma \cup \Delta) = P(B, \Gamma \cup \Delta)$ for all Δ.

The first theorem tells us that if given background assumptions Γ there is no additional evidence Δ that can distinguish between A and B, then given the background assumptions Γ, A and B are indistinguishable when used as evidence, no matter what additional evidence Δ we may have. In short, statements indistinguishable as conclusions are indistinguishable as evidence. Our second theorem tells us that if given just the background assumptions Γ, the statements A and B function the same when used as additional evidence, then Γ does not distinguish between A and B as conclusions. Our third theorem is the converse of the first theorem and is an immediate consequence of our second theorem; it says that statements indistinguishable as evidence are indistinguishable as conclusions.

As we will soon show, our conditions NP.1–NP.7 are sufficient by themselves to exactly characterize classical propositional logic. A great many logics are just extensions of classical propositional logic. First-order predicate calculus with identity and second-order logic may both be regarded as extensions of classical propositional logic; other examples include standard modal, temporal, and deontic logics. These logics are all obtained by enriching the language and adding more axioms and (or) inference rules. To extend our probability theory to such logics, it is only necessary to add some simple restrictions to our basic neo-classical theory. For every axiom AX_i beyond those required for classical propositional logic, we must add the following restriction:

NP.8. $P(AX_i, \Gamma) = 1$.

And for every inference rule IR.i beyond those required for classical propositional logic, we must add the following restriction:

NP.9. If for all Δ, $P(A_1, \Gamma_1 \cup \Delta) = 1$ and ... and $P(A_j, \Gamma_j \cup \Delta) = 1$, then for all Δ, $P(A, \Gamma \cup \Delta) = 1$, where Γ, A, the Γ_i, and the A_i satisfy conditions COND.

For inference rules with vacuous antecedents, the antecedent of NP.9 will simply be eliminated. Recall that any logic defines the corresponding set U_M of maximal sets of expressions. It is a simple matter to show that our weighting function construction still works for any extended version of neo-classical conditional probability theory.

Theorem TNP.4

Let $\{w_1, w_2, \ldots\}$ be a well-ordered sequence of elementary weighting functions, all defined on $\mathcal{P}(U_M)$, and let the function P be defined as follows:

$$P(A, \Gamma) = 1, \quad \text{if } w_i(M(\Gamma)) = 0 \text{ for all } w_i$$
$$= w_i(M(\{A\}) \cap M(\Gamma))/w_i(M(\Gamma)),$$
$$\text{for the first } w_i \text{ such that } w_i(M(\Gamma)) \neq 0$$

Then P so defined satisfies NP.1–NP.9.

We wish to reiterate that the probability theory appropriate for classical propositional logic is given by constraints NP.1–NP.7 alone. For each logic determined by adding axioms and (or) inference rules to classical propositional logic, we obtain an appropriate probability theory by adding versions of NP.8 and NP.9 corresponding to that particular

logic. Thus each distinct extension of classical logic will correspond to a distinct probability theory. Our weighting function construction is a quick way of assuring ourselves that our probability theory for some specific logic is not trivial in the sense of being limited to a certain finite number of distinct values. Note that by theorem TNP.4, the number of distinct values a probability function may take is limited only by the number of distinct maximal sets determined by the logic. Thus if the logic is not trivially limited, then the corresponding probability functions will not be trivially limited.

For classical propositional logic and each of its extensions, it is now possible to give an autonomous probabilistic definition of semantic entailment based on the probability theory appropriate for the logic. Note that the following definition does not depend on the notion of maximal sets.

Definition DNP.2

The set of expressions Γ *semantically implies* expression E with respect to logic L, symbolically $\Gamma \Vdash_L E$, if and only if for every probability function P satisfying NP.1–NP.9 (NP.1–NP.7 if the logic is classical propositional logic), $P(E, \Gamma \cup \Delta) = 1$ for every set of expressions Δ.

All proof theoretic notions have been eliminated from our definition of semantic entailment. For a given logic L, constraints NP.8 and NP.9 model the proof theory in a rather direct way. However, note that in any particular case, these conditions may be stated without any mention of concepts from proof theory. Our concern here has been to demonstrate the existence of such a theory, and we have not been concerned to formulate it in any particularly clever way. It should be remembered that there is nothing sacred about our specific formulation of the constraints, and there will be many equivalent sets of constraints, in some of which there will be no appearance of conditions like NP.8 and NP.9. In fact, for many logics (all?) there will be non-equivalent sets of constraints, any one of which could serve as a basis for a formal semantics. In any case, no matter what the logic, the appropriate probability theory as outlined here will be truly autonomous in the sense that the probability theory for any particular logic L can be stated without mention of any L-proof theoretic notions. Consequently, the following soundness and completeness results are of much greater interest than theorems TLP.1 and TLP.2.

Theorem TNP.5 (soundness)
If $\Gamma \vdash_L E$ then $\Gamma \Vdash_L E$.

Theorem TNP.6 (completeness)
If $\Gamma \Vdash_L E$ then $\Gamma \vdash_L E$.

These results on soundness and completeness are similar to those obtained in Morgan (1982*a*). However, there the probability functions were defined over $\mathfrak{L} \times \mathfrak{L}$ rather than over $\mathfrak{L} \times \mathcal{P}(\mathfrak{L})$. It is interesting to note that this simple change in the probability functions permits much simpler proofs of the desired results. Further, the theory used in Morgan (1982*a*) was not derived from a conditionalized Komolgoroff theory, as we have done here. These results will be considerably strengthened by the end of Sect. 7 by the development of an appropriate theory for logics weaker than classical logic.

There are a number of ways theorem TNP.5 could be proved, depending on the specifics of the logic. If the logic is purely classical, then an algebraic argument will do. However, the simplest and most general technique that works for all logics is an inductive argument on derivations. The axioms and inference rules for classical propositional logic may be easily handled by constraints NP.1–NP.7; the details will depend on the specific formulation of the logic. Constraint NP.8 guarantees that all additional axioms take probability 1 on any set of assumptions. Constraint NP.2 guarantees that any member of the set of assumptions takes probability 1 given those assumptions even when supplemented by others. Finally, NP.9 may be applied for the induction step over the additional inference rules.

The proof of theorem TNP.6 is not substantially different from that for theorem TLP.2, which we sketched above. We follow the Henkin pattern and assume that not $\Gamma \vdash_L E$; we must then construct a function satisfying NP.1–NP.9 such that for some set Δ, it is not the case that $P(E, \Gamma \cup \Delta) = 1$. We simply select an arbitrary superset of Γ which is maximal with respect to E. We assign a weight of 1 to any set containing the chosen maximal set and 0 to all other sets. The probability function determined by this weighting function will assign E the value 0 given $\Gamma \cup \emptyset$.

It is important to emphasize once again that, following our procedure, we can begin with any logic satisfying the minimal conditions R.1–R.4 and obtain a probability theory which serves as a formal semantics for that logic. Perhaps two simple examples will be instructive.

For the first example, let us consider classical first-order logic. Briefly, let us assume that our language contains predicate and function symbols of arbitrary adicity, as well as individual constants and variables; we will use the notation $(\forall x)$ for the universal quantifier with respect to the individual variable x. We assume standard definitions for well-formed formulas, for individual terms, and for free and bound occurrences of variables. The material conditional $A \supset B$ is defined as $\sim(A \wedge \sim B)$. We use the notation At/x to mean the formula obtained by replacing every free occurrence of x in A by term t, providing x is free for t in A. If we begin with some standard axiomatic account of classical propositional logic, we obtain classical first-order logic by adding the following additional inference rule and two axioms:

QR. If $\Gamma \vdash A$ then $\Gamma \vdash (\forall x)A$.
QAX1. $(\forall x)A \supset At/x$ for all terms t free for x in A.
QAX2. $(\forall x)(A \supset B) \supset (A \supset (\forall x)B)$ providing A has no free occurrence of x.

See Mendelson (1964) for details. Following the procedure outlined above, we need to add the following constraints to NP.1–NP.7:

NP.QR. If $P(A, \Gamma \cup \Delta) = 1$ for all Δ, then $P((\forall x)A, \Gamma \cup \Delta) = 1$ for all Δ.
NP.QAX1. $P((\forall x)A \supset At/x, \Gamma) = 1$ for all terms t free for x in A.
NP.QAX2. $P((\forall x)(A \supset B) \supset (A \supset (\forall x)B, \Gamma) = 1$ if A has no free occurrence of x.

Theorems TNP.5 and TNP.6 guarantee that the probability theory obtained from NP.1–NP.7, NP.QR, NP.QAX1 and NP.QAX2 exactly captures classical first-order predicate calculus. Note that we do not require the usual notions of first-order model theory, such as a domain of objects and interpretation functions mapping variables and terms into the domain. There will of course be many alternative sets

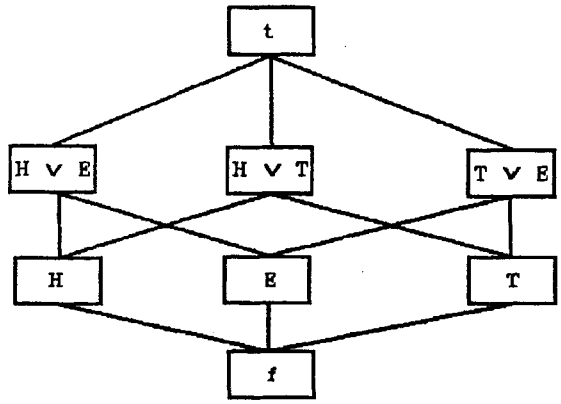

FIG. 1. Lattice of distinct propositions.

of constraints that yield the same result. For example, consider the following two constraints:

NP.Q1. $P((\forall x)A, \Gamma) \leq P(At/x, \Gamma)$ for all terms t free for x in A.

NP.Q2. If $P(A, \Gamma) \leq P(B, \Gamma)$ for all Γ and x does not occur free in A, then $P(A, \Gamma) \leq P((\forall x)B, \Gamma)$ for all Γ.

It can be shown by essentially the same proofs as given in Morgan (1984) that the probability theory given by constraints NP.1–NP.7 plus NP.Q1 and NP.Q2 also exactly captures classical first-order predicate calculus.

For our second example, let us consider the simple propositional modal logic T. We will use the symbol $\Box$ for the monadic necessity operator. One standard formulation for T is to add the following rule of necessitation and two axioms to a standard axiomatization of classical propositional logic:

T.NEC. If $\vdash A$ then $\vdash \Box A$.
T.AX1. $\Box(A \supset B) \supset (\Box A \supset \Box B)$.
T.AX2. $\Box A \supset A$.

Using the techniques of this section, the required additions to NP.1–NP.7 are the following:

NP.TNEC. If $P(A, \Delta) = 1$ for all Δ, then $P(\Box A, \Delta) = 1$ for all Δ.
NP.TAX1. $P(\Box(A \supset B) \supset (\Box A \supset \Box B), \Gamma) = 1$.
NP.TAX2. $P(\Box A \supset A, \Gamma) = 1$.

As before, theorems TNP.5 and TNP.6 guarantee that the probability theory obtained from NP.1–NP.7, NP.NEC, NP.TAX1, and NP.TAX2 exactly captures the propositional modal logic T. Note that we do not require any arcane notions like possible worlds or accessibility relations between possible worlds. And as usual, there will be many alternative formulations of the theory which will accomplish the same task. See Morgan (1982*b*) for constraints that are more intuitively appealing and psychologically better motivated.

It is interesting to note that we could have added even more constraints and still been able to obtain both soundness and completeness. For example, from the completeness proof, it should be obvious that we could have restricted the functions to be 0–1 valued. It is perhaps surprising to note that no matter what the logic, we never require more than 0–1 valued functions to invalidate a non-theorem. In this sense, the probability theory is incredibly rich; for most any logic, there will be non-equivalent sets of constraints that provide a theoretically adequate semantics. So, in any practical case, it may be possible to strengthen the set of constraints in order to make them more tractable from a computational point of view, and yet still retain soundness and completeness.

TABLE 1. Negations of distinct propositions

p	$\sim p$
t	f
T ∨ E	H
H ∨ T	E
H ∨ E	T
H	T ∨ E
E	H ∨ T
T	H ∨ E
f	t

Thus we have shown that rather than thinking of probability theory as being in some sense *opposed* to standard logics, it would be more correct to think of probability theory as incorporating standard logics as special cases. And since probability theory is so much richer than standard logics, it will be possible to explicate many more aspects of natural language and commonsense inference than can be handled with the more limited resources of any single standard logic.

6. A simple weighting function example

It might be useful at this point to give a concrete practical example of the construction of a probability distribution using weighting functions. To keep the example simple, we will consider only a classical propositional logic. Suppose I am considering a single flip of a coin, one side of which is heads and the other side of which is tails; if pressed, I may admit the *logical* possibility that the coin could come to rest on its edge, although I believe this to be a *practical* impossibility. Consider the language with sentence letters f, T, H, E, and t. Intuitively, we may think of f as some universal falsehood and t as some universal truth. The other sentence letters stand for English sentences as follows:

H: The coin will come to rest showing heads.
T: The coin will come to rest showing tails.
E: The coin will come to rest on its edge.

The lattice of logically distinct propositions is given in Fig. 1, while Table 1 specifies the negations of the distinct propositions.

Of course, disjunctions are least upper bounds and conjunctions are greatest lower bounds on the lattice diagram. The following are the only three maximal sets of expressions:

(*i*) $\{t, H \vee E, H \vee T, H\}$
(*ii*) $\{t, H \vee E, T \vee E, E\}$
(*iii*) $\{t, H \vee T, T \vee E, T\}$

Since we have only a finite number of maximal sets, a simple way of creating elementary weighting functions is to assign each maximal set a weight and then take the weight of a set of maximal sets to be the sum of the weights of the members. We will adopt this technique to construct two elementary weighting functions as follows:

TABLE 2. Probability values, $P(A, \Gamma)$

	Γ							
A	t	T ∨ E	H ∨ T	H ∨ E	H	E	T	f
t	1	1	1	1	1	1	1	1
T ∨ E	$n/(m+n)$	1	$n/(m+n)$	0	0	1	1	1
H ∨ T	1	1	1	1	1	0	1	1
H ∨ E	$m/(m+n)$	0	$m/(m+n)$	1	1	1	0	1
H	$m/(m+n)$	0	$m/(m+n)$	1	1	0	0	1
E	0	0	0	0	0	1	0	1
T	$n/(m+n)$	1	$n/(m+n)$	0	0	0	1	1
f	0	0	0	0	0	0	0	1

Maximal set	w_1	w_2
(*i*)	m	0
(*ii*)	0	1
(*iii*)	n	0

If I believe that heads and tails are equally likely, then I should choose $m = n$. On the other hand, if I believe the coin to be biased, then I may choose m and n to reflect my beliefs about the degree of bias. These weights give rise to the probability distribution given in Table 2. Note that since we are dealing with only finite sets of propositions, each set will be logically equivalent to its conjunction and hence to one of the propositions on the lattice given above. Hence we may represent sets of propositions for the assumption position by single expressions across the table. The single propositions on the left of the table represent the conclusion position.

As long as the assumption set is equivalent to neither E nor f, I use elementary weighting function w_1 to calculate the probability. If the assumption set is equivalent to E, then I must use w_2 to calculate the probability, since w_1 assigns the assumption set the value 0. Finally, if the assumption set is equivalent to f, then I just set all probabilities to 1, since all my weighting functions assign a value of 0 to the assumption set.

From the table it is easy to see that the only abnormal assumption sets would be those equivalent to f; examples of such sets are $\{f\}$ and $\{H, E\}$. Note that although E is assigned probability 0 on tautological evidence, evidence sets equivalent to E are not abnormal. I may well believe that it is practically impossible for the coin to come to rest on its edge, and yet I can still rationally entertain the assumption that it has done so. For example, on the assumption that the coin has come to rest on its edge, I should reject the proposal that it has come to rest with either heads or tails uppermost. This simple example is just one of many in which our normal response is to assign probability 0 to a statement that is not, technically speaking, inconsistent. Just because we are virtually certain some statement A can never occur (and are willing to wager accordingly), it does not follow that we cannot rationally entertain the possibility of the occurrence of A and reason appropriately from the hypothetical assumption that A has occurred.

Finally, as promised earlier, the table provides a simple example in which $P(A, \Gamma) = 1$, but $P(A, \Gamma \cup \Delta) = 0$. Note that $P(H \vee T, t) = 1$ but $P(H \vee T, E) = 0$. That is, at the start of the coin toss experiment, it may be the case that I am absolutely certain that the coin will come to rest with either heads or tails uppermost. However, if (by some wild miracle) the coin were to come to rest on edge, I would say there is no chance at all that either heads or tails would be uppermost. Thus $P(A, \Gamma) = 1$ is not sufficient for asserting that Γ logically entails A.

7. Core confirmation theory and weak logics

Our success with logics formulated as extensions of classical propositional logic naturally leads us to examine logics weaker than classical logic. Standard examples include intuitionistic logic and quantum logic. To incorporate all such logics, we have developed a minimal "core" probability for any language. Since our core theory is considerably weaker than classical probability, some may object to the use of the term "probability." Consequently, we will refer to the theory as a system of confirmation functions.

Since we want our core confirmation theory to serve as the basis of a formal semantics for *any* logic on *any* language, the core theory can make no assumptions about the syntactic apparatus available in the language. The only thing that can be assumed is that the language $\mathcal{L}$ consists of a (possibly infinite) set of well-formed expressions. Our confirmation functions may be thought of as specifying the degree of rational belief in one expression when given some set of expressions as evidence. Thus as before, we will assume the functions (here designated by c) are defined on $\mathcal{L} \times \mathcal{P}(\mathcal{L})$. The following simple constraints serve as our core confirmation theory:

C.1. $0 \leq c(A, \Gamma) \leq 1$.
C.2. If $A \in \Gamma$ then $c(A, \Gamma) = 1$.
C.3. $c(A, \Gamma) \times c(B, \Gamma \cup \{A\}) = c(B, \Gamma) \times c(A, \Gamma \cup \{B\})$.

Standard versions of these three constraints hold for virtually all accounts of conditional probability theory. In particular, note that C.1–C.3 may be derived directly from NP.1–NP.7. Thus adopting C.1–C.3 amounts to a relaxation of constraints and simply permits more functions to be included. Constraint C.1 merely sets out the range of the functions, corresponding directly to NP.1. Constraint C.2 simply says that if your set of assumptions includes A, then in light of those assumptions, the degree of confirmation of A must be maximal; it corresponds directly to NP.2.

Constraint C.3 requires a bit more discussion. It follows directly from NP.4 and NP.6, above. Since our language is completely arbitrary, we cannot be sure that there is a con-

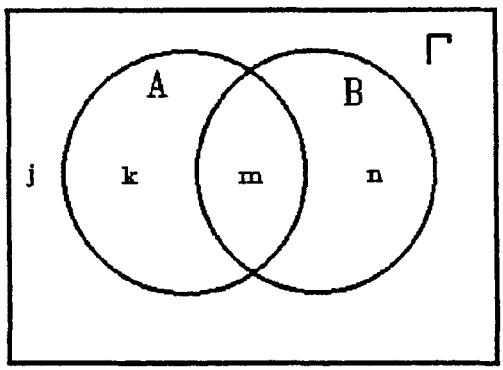

FIG. 2. Relative frequency example.

nective in the language corresponding to conjunction. However, by combining NP.4 and NP.6, we obtain an essential condition that makes no mention of any connective. Note that we use unions rather than intersections in stating constraint C.3, as is appropriate when talking about sets of expressions from the language rather than sets of events or sets of "universe descriptions." The semantic impact of C.3 is to require that our confirmation functions accord with elementary relative frequency considerations. For a diagrammatic representation, see Fig. 2. The box is presumed to contain all of the universe designs which are compatible with the statements in Γ; in this example, for simplicity we assume there are only a finite number of universe designs. We assume the circle with A inside it contains all of the Γ universe designs that are compatible with A; and similarly, we assume the circle with B inside it contains all of the Γ universe designs that are compatible with B. Let j be the number of Γ universe designs that are compatible with neither A nor B. Let k be the number of Γ universe designs compatible with A but not with B. Let n be the number of Γ universe designs compatible with B but not with A. And let m be the number of Γ universe designs compatible with both A and B. Then we may take the confirmation function $c(E, \Delta)$ as the relative frequency of E designs among the Δ designs. Accordingly, we obtain the following values:

$c(A, \Gamma) = (k + m)/(j + k + m + n)$
$c(B, \Gamma \cup \{A\}) = m/(k + m)$
product $= m/(j + k + m + n)$
$c(B, \Gamma) = (m + n)/(j + k + m + n)$
$c(A, \Gamma \cup \{B\}) = m/(m + n)$
product $= m/(j + k + m + n)$

Standard concrete examples abound including all the well-known games of chance involving dice and cards. In short, C.3 embodies the requirement that our confirmation functions must accord with elementary relative frequency considerations.

Although C.1–C.3 are quite simple, they have several important consequences worth mentioning at this point. First, note that these constraints are sufficient to ensure that statements indistinguishable by any conceivable evidence are indistinguishable in their role *as* evidence statements.

Theorem TC.1

If $c(A, \Gamma) = c(B, \Gamma)$, for all Γ, then for all statements D, $c(D, \Gamma \cup \{A\}) = c(D, \Gamma \cup \{B\})$, for all Γ.

The proof of this result is not difficult, but since it cannot be found elsewhere, we will sketch it here. Those not interested in the details should skip to the next paragraph. Assume the hypothesis of the theorem:

(TC1.1) $c(A, \Gamma) = c(B, \Gamma)$, for all Γ

Let Γ' be a completely arbitrary set. Then (TC1.1) and C.2 guarantee all of the following:

(TC1.2) $c(A, \Gamma' \cup \{B\}) = 1$

(TC1.3) $c(B, \Gamma' \cup \{A\}) = 1$

(TC1.4) $c(A, \Gamma' \cup \{B, D\}) = 1$

(TC1.5) $c(B, \Gamma' \cup \{A, D\}) = 1$

Using (TC1.2) and (TC1.3), we have

(TC1.6) $c(A, \Gamma' \cup \{B\}) \times c(D, \Gamma' \cup \{A, B\}) = c(B, \Gamma' \cup \{A\}) \times c(D, \Gamma' \cup \{A, B\})$

Applying constraint C.3 to both sides of (TC1.6) yields

(TC1.7) $c(D, \Gamma' \cup \{B\}) \times c(A, \Gamma' \cup \{B, D\}) = c(D, \Gamma' \cup \{A\}) \times c(B, \Gamma' \cup \{A, D\})$

Then using (TC1.4) on the left and (TC1.5) on the right of (TC1.7), we have

(TC1.8) $c(D, \Gamma' \cup \{B\}) = c(D, \Gamma' \cup \{A\})$

But Γ' was arbitrary, so the proof of the theorem is complete.

Note that theorem TC.1 has nothing whatever to do with any logic that might be defined on our language. It is solely concerned with the way in which evidence is assessed. The theorem is also independent of any internal syntactic structure present in the language. The result holds independently of any connectives or sentence forming operators.

The reader may feel that by not specifying any sentence forming operators, our core confirmation theory is too weak to be of any interest. It may therefore be surprising to note the bounds which our three simple constraints impose on functions for some of the standard sentential connectives. Let us first consider a general conjunctive operator, here designated by $\wedge$. We usually think of conjunction as a binary operator, but we can instead define it to apply to arbitrarily long finite sequences of formulas in the obvious way, thereby avoiding the need to state associativity principles. Because of general familiarity with the binary operator, I will write $A_1 \wedge \ldots \wedge A_n$ *instead of* $\wedge(A_1, \ldots, A_n)$. An operator is deemed to be conjunctive just in case it is idempotent, is permutative, and satisfies a greatest lower bound principle. For definiteness, we list these properties in terms of proof theory.

CON.1. Idempotency:

$\Gamma \cup \{A \wedge \ldots \wedge A\} \vdash A$ and $\Gamma \cup \{A\} \vdash A \wedge \ldots \wedge A$.

CON.2. Permutativity:

$\Gamma \cup \{A_1 \wedge \ldots \wedge A_i \wedge \ldots \wedge A_k \wedge \ldots \wedge A_n\} \vdash A_1 \wedge \ldots A_k \wedge \ldots \wedge A_i \wedge \ldots \wedge A_n$.

CON.3. Greatest lower bound principle:

(a) $\Gamma \cup \{A_1 \wedge \ldots \wedge A_n\} \vdash A_i$ for all i, $1 \leq i \leq n$.
(b) If $\Gamma \cup \{B\} \vdash A_i$ for all i, $1 \leq i \leq n$, then $\Gamma \cup \{B\} \vdash A_1 \wedge \ldots \wedge A_n$.

To see how these principles (and others to be discussed below) transform into our core confirmation theory, we will need the following two theorems. As with theorem TC.1, the proofs depend only on constraints C.1–C.3.

Theorem TC.2
$c(A, \Gamma \cup \{B\}) = c(B, \Gamma \cup \{A\}) = 1$ for all Γ if and only if $c(A, \Gamma) = c(B, \Gamma)$ for all Γ.

Theorem TC.3
$c(A, \Gamma \cup \{B\}) = 1$ for all Γ if and only if $c(B, \Gamma) \leq c(A, \Gamma)$ for all Γ.

We may now ask about the possible functions which could serve to define a conjunctive operator in our core confirmation theory. To this end, we offer the following definition of a "product" function:

Definition DC.1
$\text{prd}(\langle A \rangle, \Gamma) = c(A, \Gamma)$.
$\text{prd}(\langle A_1, \ldots, A_n \rangle, \Gamma) = c(A_1, \Gamma) \times \text{prd}(\langle A_2, \ldots, A_n \rangle, \Gamma \cup \{A_1\})$.

Using theorems TC.2 and TC.3, it is easy to see that the three desiderata of idempotency, permutativity, and greatest lower bound principle all hold of the function prd, as the following theorem indicates.

Theorem TC.4
(*i*) $\text{prd}(\langle A, \ldots, A \rangle, \Gamma) = \text{prd}(\langle A \rangle, \Gamma)$.
(*ii*) $\text{prd}(\langle A_1, \ldots, A_i, \ldots, A_k, \ldots, A_n \rangle, \Gamma) = \text{prd}(\langle A_1, \ldots, A_k, \ldots, A_i, \ldots, A_n \rangle, \Gamma)$.
(*iii.a*) $\text{prd}(\langle A_1, \ldots, A_n \rangle, \Gamma) \leq \text{prd}(\langle A_i \rangle, \Gamma)$ for all i, $1 \leq i \leq n$.
(*iii.b*) If $\text{prd}(\langle B \rangle, \Gamma) \leq \text{prd}(\langle A_i \rangle, \Gamma)$ for all Γ and for all i, $1 \leq i \leq n$, then $\text{prd}(\langle B \rangle, \Gamma) \leq \text{prd}(\langle A_1, \ldots, A_n \rangle, \Gamma)$.

Note that theorem TC.4 does not *guarantee* that the language $\mathcal{L}$ contains a conjunctive connective, nor does it *require* that $\mathcal{L}$ contains a conjunctive operator. What the theorem tells us is that if our syntax does contain a conjunctive operator, then the function prd is an appropriate semantic characterization of it. The proof of theorem TC.4 requires only the three constraints of our core confirmation theory.

Many languages also contain a disjunctive operator, here designated by $\vee$. As with the conjunctive operator discussed above, we will here presume a general disjunctive operator applying to arbitrary sequences of formulas rather than a simple binary disjunctive operator; as before, by using such a general operator, we avoid the need to worry about associative principles. An operator is deemed to be disjunctive just in case it is idempotent, is permutative, and satisfies a least upper bound principle. We can state these characteristics in terms of proof theory.

DIS.1. Idempotency:

$\Gamma \cup \{A \vee \ldots \vee A\} \vdash A$ and $\Gamma \cup \{A\} \vdash A \vee \ldots \vee A$.

DIS.2. Permutativity:

$\Gamma \cup \{A_1 \vee \ldots \vee A_i \vee \ldots \vee A_k \vee \ldots \vee A_n\} \vdash A_1 \vee \ldots \vee A_k \vee \ldots \vee A_i \vee \ldots \vee A_n$.

DIS.3. Least upper bound principle:

(a) $\Gamma \cup \{A_i\} \vdash A_1 \vee \ldots \vee A_n$ for all i, $1 \leq i \leq n$.

(b) If $\Gamma \cup \{A_i\} \vdash B$ for all i, $1 \leq i \leq n$, then $\Gamma \cup \{A_1 \vee \ldots \vee A_n\} \vdash B$.

We may now ask about the possible functions which could serve to define a disjunctive operator in our core confirmation theory. To this end, we offer the following definition of a "sum" function:

Definition DC.2
$\text{sum}(\langle A \rangle, \Gamma) = c(A, \Gamma)$.

$$\text{sum}(\langle A_1, \ldots, A_n \rangle, \Gamma) = \sum_{i=1}^{n} (-1)^{i-1} \text{prd}(\langle A_{k_1}, \ldots, A_{k_i} \rangle, \Gamma),$$

for $1 \leq k_j \leq n$ and $k_x \neq k_y$.

Again using theorems TC.2 and TC.3, it is easy to see that the three desiderata of idempotency, permutativity, and least upper bound principle all hold of the function sum, as the following theorem indicates.

Theorem TC.5
(*i*) $\text{sum}(\langle A, \ldots, A \rangle, \Gamma) = \text{sum}(\langle A \rangle, \Gamma)$.
(*ii*) $\text{sum}(\langle A_1, \ldots, A_i, \ldots, A_k, \ldots, A_n, \rangle, \Gamma) = \text{sum}(\langle A_1, \ldots, A_k, \ldots, A_i, \ldots, A_n \rangle, \Gamma)$.
(*iii.a*) $\text{sum}(\langle A_i \rangle, \Gamma) \leq \text{sum}(\langle A_1, \ldots, A_n \rangle, \Gamma)$ for all i, $1 \leq i \leq n$.
(*iii.b*) If $\text{sum}(\langle A_i \rangle, \Gamma) \leq \text{sum}(\langle B \rangle, \Gamma)$ for all Γ and for all i, $1 \leq i \leq n$, then $\text{sum}(\langle A_1, \ldots, A_n \rangle, \Gamma) \leq \text{sum}(\langle B \rangle, \Gamma)$.

Theorem TC.5 is similar to theorem TC.4. Theorem TC.5 does not *guarantee* that the language $\mathcal{L}$ contains a disjunctive connective, nor does it *require* that $\mathcal{L}$ contains a disjunctive operator. What the theorem tells us is that if our syntax does contain a disjunctive operator, then the function sum is an appropriate semantic characterization of it. The proof of theorem TC.5 requires only the three constraints of our core confirmation theory.

The function prd is just what classical probability theory requires for conjunctions; see NP.4 above. And the function sum is just what classical probability theory requires for disjunctions; see NP.3 above. Thus, while our core confirmation theory places no restrictions on the syntax of the language, the three simple constraints of the theory are sufficient to guarantee that the classical probability functions are suitable for the classical connectives. Thus these theorems serve as a partial justification for calling our theory a core confirmation theory.

Of course, even if the proof theory does contain a conjunctive or a disjunctive operator, there is no requirement that we semantically define the operator using the prd or sum functions. We could simply add very weak constraints to the core confirmation theory corresponding to the appropriate axioms and inference rules. However, such a procedure may not be as desirable simply from a computational point of view. The sum and prd functions give us at least a partial guide to computing the probability values. In some cases, the sum and prd functions will be the only possible choices anyway. Cox (1961) gives a rather neat derivation of these functions from rather sparse assumptions, but his assumptions do include the presumption that the language is Boolean. Nevertheless, once again we wish to emphasize that the real importance of theorems TC.4 and TC.5 is that they indicate the strength of the apparently simple constraints C.1–C.3.

We will now turn our attention to results more directly concerned with the presumed logic. We could continue from this point with a discussion of general sequent logics, in a way parallel to the development given above. However, we will vary the presentation a bit in the hope of preventing complete boredom. We will now assume that a logic L is defined on the language by recursively specifying a set of axioms AX_i and a set of weak inference rules of the following form:

WIR.i. If $\Gamma \vdash A_1$ and ... and $\Gamma \vdash A_j$, then $\Gamma \vdash A$, where Γ, A, and the A_i satisfy conditions COND.

Note that by restricting ourselves to rules of the sort WIR.i, we do not have the full generality of rules of the form IR.i, since the assumption sets in the antecedent must always be the same as the assumption set in the consequent. Instead of allowing derivations using sequents, we define derivations to be sequences of formulas of language $\mathcal{L}$ in the usual way.

Definition DC.3

Formula A is derivable in logic L from the set of formulas Γ (symbolically, $\Gamma \vdash_L A$) if and only if there is a finite sequence of formulas $E_1, \ldots, E_n$, the last one of which is the formula A, such that for each member E_i of the sequence: (a) E_i is an axiom, or (b) E_i is a member of Γ, or (c) E_i follows from previous members of the sequence by an inference rule.

As before, we will generally write $\Gamma \vdash A$ without any subscript when it is clear which logic is meant. Given definition DC.3 and the general form WIR.i of our rules, it is an easy matter to *prove* that all of the restrictions R.1–R.4 hold, providing the conditions COND in the rules WIR.i do not mention the set Γ. In any case, we will assume that R.1–R.4 hold, as before.

To obtain the confirmation functions appropriate for a given logic, we must add the following constraints (corresponding to NP.8–NP.9 above) to our core theory for each axiom and inference rule:

C.4. $c(AX_i, \Gamma) = 1$.

C.5. If for all Δ, $c(A_1, \Gamma \cup \Delta) = 1$ and ... and $c(A_j, \Gamma \cup \Delta) = 1$, then for all Δ, $c(A, \Gamma \cup \Delta) = 1$, where Γ, A, and the A_i satisfy conditions COND.

For constraints of the form C.5 corresponding to rules with vacuous antecedents, the antecedent of the conditional is simply eliminated.

We will say that a function is appropriate for the logic in question just in case it satisfies C.1–C.5. Just as with classical probability theory, these constraints will in general pick out a large class of functions for a given logic. Which function should be used for a particular application depends on the extra-logical facts of the case (e.g., how the die is weighted or how many cards there are in the deck). That is, the correct theory of confirmation depends on the logical facts, and the correct function satisfying that theory depends on the extra-logical facts.

At this point it is important to note that the characterization of our functions does not in any way depend on proof-theoretic notions like "maximal sets." As was the case in our formulation of neo-classical probability theory, some of the confirmation constraints we have listed parallel the proof theory in a rather direct way. But even these constraints do not use proof-theoretic notions; in general, there will be other equivalent sets of constraints that do not directly parallel the inference rules. In this formulation, we are concerned with the mere existence of an appropriate set of constraints, not with their syntactic "cuteness."

Thus, no matter what the logic, our core confirmation theory can be used to formulate an autonomous formal semantic theory which is characteristic of that logic, as we will now proceed to show. First, we define a notion of semantic implication, parallel to definition DNP.2 above. The soundness theorem follows immediately.

Definition DC.4

The set of expressions Γ *semantically implies* expression E with respect to logic L, symbolically $\Gamma \Vdash_L E$, if and only if for every confirmation function c satisfying C.1–C.5, $c(E, \Gamma \cup \Delta) = 1$ for every set of expressions Δ.

Theorem TC.6 (soundness)

If $\Gamma \vdash_L E$, then $\Gamma \Vdash_L E$.

The proof of theorem TC.6 is not quite as trivial as was the proof of theorem TNP.5. Assume that $\Gamma \vdash_L E$. By definition, there is a finite sequence of formulas, say $E_1, \ldots, E_n$, the last one of which is A, such that for each member E_i of the sequence: (a) E_i is an axiom, or (b) E_i is a member of Γ, or (c) E_i follows from previous members of the sequence by an inference rule. Our proof is the usual induction on the members E_i of the sequence. For case (a) we may appeal to C.4. Similarly, for case (b) we may appeal to C.2. For case (c), suppose that the inference rule in question allows us to infer B from $A_1, \ldots, A_m$. That is, E_i is B, and each of the A_j is one of the E_k earlier in the sequence than E_i. For notational convenience, we will use Γ_j to stand for the set $\{A_1, \ldots, A_j\}$. Let Γ' be an arbitrary set of expressions, and let c be an arbitrary confirmation function. Then C.5 guarantees the following:

(TC6.1) $c(E_i, \Gamma \cup \Gamma' \cup \Gamma_m) = 1$

Now, consider A_m. Since it occurs earlier in the sequence than E_i, the induction hypothesis ensures that

(TC6.2) $\Gamma \Vdash_L A_m$

But by definition, (TC6.2) guarantees that

(TC6.3) $c(A_m, \Gamma \cup \Gamma' \cup \Gamma_{m-1}) = 1$

Multiplying the two sides of (TC6.1) and (TC6.3) together and applying C.3 yields

(TC6.4) $c(E_i, \Gamma \cup \Gamma' \cup \Gamma_{m-1}) \times c(A_m, \Gamma \cup \Gamma' \cup \Gamma_{m-1} \cup \{E_i\}) = 1$

From C.1 and (TC6.4) we immediately have

(TC6.5) $c(E_i, \Gamma \cup \Gamma' \cup \Gamma_{m-1}) = 1$

Going through a similar sequence of steps $m-1$ more times will yield

(TC6.6) $c(E_i, \Gamma \cup \Gamma') = 1$

Since Γ' and c were completely arbitrary, we know from (TC6.6) that

(TC6.7) $\Gamma \Vdash_L E_i$

Thus Γ semantically entails every member of the sequence. Since A is the last member of the sequence, it follows that Γ semantically implies A. Thus the proof of soundness is finished. Note that every one of the conditions C.1–C.5 was explicitly used in the proof.

We will now turn our attention to strong completeness. Weak completeness can be obtained as a trivial consequence by letting Γ be empty.

Theorem TC.7 (completeness)
If $\Gamma \models_L E$ then $\Gamma \vdash_L E$.

To establish completeness, we follow the usual tack and argue for the contrapositive. Let Γ^* be an arbitrary set of expressions and let A^* be an arbitrary expression such that it is not the case that $\Gamma^* \vdash A^*$. We must show that it is not the case that $\Gamma^* \models A^*$. That is, we must find an appropriate function c and a set Γ' such that $c(A^*, \Gamma^* \cup \Gamma') \neq 1$. It will be sufficient to take $\Gamma' = \Gamma^*$; thus we need only find an appropriate function c such that $c(A^*, \Gamma^*) \neq 1$. We can define the desired function (for all expressions A and sets of expressions Γ) as follows:

$$\text{(TC7.1)} \quad c(A, \Gamma) = 1, \quad \text{if and only if } \Gamma \vdash A$$
$$= 0, \quad \text{otherwise}$$

By assumption, it is not the case that $\Gamma^* \vdash A^*$; so for c defined by (TC7.1), we clearly have $c(A^*, \Gamma^*) \neq 1$, as desired. Thus we only need to show that (TC7.1) defines a function satisfying C.1–C.5. Conditions C.1, C.2, C.4, and C.5 trivially can be seen to be satisfied. Only condition C.3 requires comment. We must show that

$$\text{(TC7.2)} \quad c(A, \Gamma) \times c(B, \Gamma \cup \{A\}) = c(B, \Gamma) \times c(A, \Gamma \cup \{B\})$$

From (TC7.1), we know that each side of equation (TC7.2) must be either 0 or 1. Suppose the left side of the equation is 1. Then from (TC7.1) we know both of the following:

$$\text{(TC7.3)} \quad \Gamma \vdash A$$

$$\text{(TC7.4)} \quad \Gamma \cup \{A\} \vdash B$$

But applying R.3 to (TC7.3) and (TC7.4) yields

$$\text{(TC7.5)} \quad \Gamma \vdash B$$

And using (TC7.3) and R.4, we know that

$$\text{(TC7.6)} \quad \Gamma \cup \{B\} \vdash A$$

Our definition of the function c in (TC7.1) allows us to conclude from (TC7.5) and (TC7.6) that the right side of (TC7.2) must also be 1, as required. Now, interchanging A and B in the above argument shows that if the right side of (TC7.2) is 1 then the left side must be 1 as well. Hence condition C.3 is satisfied. Thus the completeness result is established.

We will now consider one more justification for the appellation "core confirmation theory." Our justification is the fact that all relative frequency schemes must satisfy the constraints of the core theory. In short, we will obtain theorems similar to theorem TPP.6, above. The reader new to this material may wish to review our discussion of the intuitive motivation behind the definition of elementary weighting function (definition DPP.2), as well as our discussion of relative frequency following theorem TPP.6.

Relative frequency considerations are conveniently phrased in terms of the relative frequency of "universe designs." A universe design is usually thought of as a maximal set of sentences. If no logic is imposed, we may be much more liberal; we will allow *any* set of sentences to be a universe design. Thus, if no logic is imposed on the language $\mathcal{L}$, the set UD of universe designs is just an arbitrary subset of $\mathcal{P}(\mathcal{L})$. We define a mapping ud from sets of expressions of $\mathcal{L}$ into sets of universe designs, as follows:

$$\text{ud}(\Gamma) = \{\Delta: \Delta \in \text{UD and } \Gamma \subseteq \Delta\}$$

Intuitively, the universe designs compatible with Γ are just the supersets of Γ that qualify as universe designs. The following theorem concerning the core confirmation theory is easily proved.

Theorem TC.8
Let UD be an arbitrary subset of $\mathcal{P}(\mathcal{L})$ for arbitrary language $\mathcal{L}$. Let ud be a function mapping $\mathcal{P}(\mathcal{L})$ into $\mathcal{P}(\text{UD})$ as follows:

$$\text{ud}(\Gamma) = \{\Delta: \Delta \in \text{UD and } \Gamma \subseteq \Delta\}$$

Let $\{w_1, w_2, \ldots\}$ be a well-ordered sequence of elementary weighting functions, all defined on $\mathcal{P}(\text{UD})$, and let the function c be defined as follows:

$$c(A, \Gamma) = 1, \quad \text{if } w_i(\text{ud}(\Gamma)) = 0 \text{ for all } w_i$$
$$= w_i(\text{ud}(\{A\}) \cap \text{ud}(\Gamma))/w_i(\text{ud}(\Gamma)),$$
$$\text{for the first } w_i \text{ such that } w_i(\text{ud}(\Gamma)) \neq 0$$

Then c so defined satisfies C.1–C.3.

From theorem TC.8, we may conclude that any relative frequency scheme must satisfy our constraints for core confirmation theory. Carefully note that constraints C.1–C.3 are common to all logics, and it is for that reason that we could allow any set of expressions to serve as a universe design. However, as soon as we impose any logical structure, it is natural to insist that only maximal sets may count as universe designs. (The reader is cautioned to note that for some quite weak logics, maximal sets may very well be inconsistent from the point of view of classical logic; even so, such maximal sets will not contain every sentence of the language.) Since conditions R.1–R.4 hold for our logic, we know that theorems TL.1 and TL.2 also hold. Consequently, it is easy to prove the following theorem, analogous to theorem TNP.4 above.

Theorem TC.9
Let L be an arbitrary logic on language $\mathcal{L}$. Let UD be the set of all maximal sets of sentences with respect to L. Let ud be a function mapping $\mathcal{P}(\mathcal{L})$ into $\mathcal{P}(\text{UD})$ as follows:

$$\text{ud}(\Gamma) = \{\Delta: \Delta \in \text{UD and } \Gamma \subseteq \Delta\}$$

Let $\{w_1, w_2 \ldots\}$ be a well-ordered sequence of elementary weighting functions, all defined on $\mathcal{P}(\text{UD})$, and let the function c be defined as follows:

$$c(A, \Gamma) = 1, \quad \text{if } w_i(\text{ud}(\Gamma)) = 0 \text{ for all } w_i$$
$$= w_i(\text{ud}(A) \cap \text{ud}(\Gamma))/w_i(\text{ud}(\Gamma)),$$
$$\text{for the first } w_i \text{ such that } w_i(\text{ud}(\Gamma)) \neq 0$$

Then c so defined satisfies C.1–C.5.

We have used neither theorem TC.8 nor theorem TC.9 in establishing our soundness and completeness results. So

our formal semantics based on core confirmation theory is indeed autonomous, making no surreptitious use of proof theory. Consequently, our core confirmation theory may be deemed to be a universal semantics. Virtually any logic on any language turns out to be just a special case of core confirmation theory. Theorem TC.9 allows us to easily establish non-triviality results. Unless the logic imposes an upper bound on the number of maximal sets, there will be no upper bound on the number of values confirmation functions may assign. In short, if the logic is not trivial, then the corresponding confirmation theory will not be trivial.

8. Non-numerical accounts of probability theory

Thus far the accounts of probability we have discussed, including the core confirmation theory of the previous section, have all presumed a mapping into the real interval [0, 1]. Any realistic account of even ideally rational agents cannot seriously represent systems of *conscious* belief in this way. Virtually no one consciously assigns precise numerical values to their beliefs. And it seems totally unreasonable to assume that ideally rational agents could even linearly order their beliefs. Computer inference systems which attempt to impose such structures are guaranteed to yield results which run counter to our intuitions in some cases.

We should also bear in mind the fact that many animals other than humans (and humans are part of the animal kingdom, after all) engage in problem solving behavior of a sort that very strongly suggests that they have belief systems not totally dissimilar to our own; it seems completely absurd to suggest that such animals *consciously* assign numerical weights to their beliefs. Since the cognitive capabilities of many animals greatly exceed the capabilities of present computer technology, it would seem that we may not have to use all of the determinateness of classical probability theory to improve computer inference systems.

It might be objected that while we are not consciously aware of all the details, it may be that unknown to us, as it were, our neurophysiology does assign precise numerical weights to the items in our belief structure. Several replies may be made to this objection. First, in addition to the fact that we are not consciously aware of assigning numerical weights to all our beliefs, most people have the subjective impression that their rational inferences are the result of *conscious* deliberations of a sort that are frequently verbalized when those inferences are challenged. It may be that the vast majority of humans are systematically deluded about the true nature of rational inference, but a good first working hypothesis is that they are not. That is, it may well be that any numerical details embedded in our neurophysiology are merely accidental features of our particular embodiment and not essential components of rational inference. Second, even if the belief system of each of us corresponds to some numerical weighting of our beliefs, it is certain that the vast majority do not share exactly the same set of beliefs to the same degree. Yet, there is a tremendous commonality in the accepted modes of rational inference which cuts across specific belief systems. Not only do we frequently reach common conclusions about actual situations, but we also frequently reach common conclusions about hypothetical cases. These observations strongly suggest that it is not the actual weights that really matter in rational inference, but at most their relative values. So, even if it turns out that neurophysiologically our belief systems do correspond to some incredibly precise and detailed numerical weighting, there is good reason to believe that a calculus of rational inference may be based on less detailed comparative information.

In short, assigning precise numerical weights to beliefs seems to be neither necessary nor sufficient for modelling the human inference engine. Standard logics are excellent examples of this point. While not adequate for everything, standard logics are remarkably good for modelling many kinds of inference. On the basis of such logics (even so-called multiple-valued logics), we can always class all statements into two categories: those that are provable from the assumption set and those that are not. Such a classification is too crude for most practical purposes, and that is just the appeal of multiple-valued logics. For multiple-valued logics, even if it is not possible to derive A from Γ, it may be possible to derive something like $J_r(A)$, which intuitively says that A holds to degree r. The usual multiple-valued logics are truth functional, and it is well known that no truth functional semantics can accord with the full complexity of relative frequency. And if we are going to opt for a semantics that accords with relative frequency, it is natural to think of the probability theory in one of its usual formulations. But there is a great range of possibilities open to us between the binary classification "provable or not provable" and the precise specification of the degree to which a sentence holds, where that degree must fall on the linear continuum of the interval [0, 1].

There are many accounts of what may loosely be called probability or confirmation theory that are not fundamentally numerical; for examples, see Fine (1973). However, most such theories have not been shown to serve as a general foundation for logics of the traditional sort. Given the limited but very real successes of standard logics in modelling human reasoning, any proposed probabilistic account of inference must treat standard logics as special cases before it can be regarded as theoretically reasonable. I do not mean to suggest that the systems in Fine (1973) could not in principle serve such a foundational role, but rather to indicate an open area of research. Another very promising system which warrants further attention can be found in Aleliunas (1990).

To my knowledge, there is only one non-numerical system of probability that has been treated as a semantic theory. The weak conditional comparative theory treated by Morgan (1984) has been used as the basis for a semantic account of both classical propositional logic and classical first-order predicate calculus. No doubt the system could be used as the basis for any extension of classical logic by following a development similar to that above, but the details have not yet been worked out. Further, no real attention has been paid to logics weaker than classical propositional logic.

But all non-numeric probabilistic accounts of inference seem to suffer severe problems with machine implementation. It seems that we must buy theoretical flexibility at the expense of computational tractability. Perhaps the most sophisticated practical approach to the problem is due to Pearl (1986), who discusses techniques for updating a probabilistic belief network in light of new evidence in a computationally reasonable way. Pearl explicitly discusses the computational complexity of his procedures as a function of the size of the belief network. At first glance, I would suggest that strongly algebraic systems of the sort developed

by Aleliunas (1990) may be better suited to overcome the computational problems than numerically based systems.

9. Conclusions

There are various stages in the development of any science. At some stages, a proliferation of seemingly independent theories and techniques seems to be the best way to proceed. However, as the number of independent approaches increases, the pressure for a single, more unified theory also increases. There comes a point when significant further progress can be made only by trying to view the forest instead of the individual trees. Perhaps the time has come in the field of AI to attempt such a unifying view. There seems little doubt that classical logic is not an adequate theoretical underpinning for all of AI. It is my belief that probabilistic considerations provide our best hope for such a unifying approach.

It should be clear on the basis of this outline that probability theory in its broadest sense encompasses all of the standard logics as special cases. Thus probability plays the role of a unifying theory, helping us to understand why the more specialized logics are adequate for various specialized tasks. More importantly, our review should lay to rest any serious antagonism between the advocates of various logics and the advocates of probability theory; those who use the various logics are in fact using probability theory, albeit a very specialized form of probability theory. Further, the fact that probability can serve as the basis for the characterization of the key formal metalogical concepts provides a good ground for believing that probability theory may well be adequate for the analysis of many other linguistic components that seem to be metatheoretical in character. Finally, I have written this outline from the standpoint of a theoretical logician, not that of an applied mathematician. Too often probability theory is thought of as simply another computational device in the toolbox of the hack mathematician. Such a view is extremely shortsighted and narrow minded and is often the basis for an unwarranted rejection of probabilistic considerations.

There is a great deal more to probability theory than the simple computation of inverse probability values using Bayes' Theorem! For just one example, when we view probability theory in this very abstract way, it is easier to isolate those assumptions about rational belief structures which make a given logic appropriate for certain applications but not for others. Further, when faced with inference patterns which do not seem to be sanctioned by known systems of logic, it then makes sense to ask if there are general constraints on rational belief structures which must hold in order to make those inference patterns legitimate.

I hope that advocates of a probabilistic approach will see that the theory has broader scope than they may have at first realized. And I hope that those who have rejected a probabilistic approach will be persuaded to give it another look.

Acknowledgments

Parts of the material in this paper have been presented at various colloquia and conferences. I would like to thank those who listened, questioned, and offered advice. In particular, I would like to thank Romas Aleliunas, David Etherington, and Robert Hadley.

ALELIUNAS, R. 1990. A new normative theory of probabilistic logic. *In* Studies in cognitive systems: knowledge and defeasible reasoning. *Edited by* H. Kyburg *et al.* Kluwer Academic Publishers, Dordrecht, The Netherlands. pp. 387–403.

CARNAP, R. 1950. Logical foundations of probability. University of Chicago Press, Chicago, IL.

CHANG, C., and KEISLER, H. 1973. Model theory. North-Holland Publishing Company, Amsterdam, The Netherlands.

CHEESEMAN, P. 1988. An inquiry into computer understanding. Computational Intelligence, **4**: 58–66.

COX, R. 1961. The algebra of probable inference. The Johns Hopkins Press, Baltimore, MD.

FINE, T. 1973. Theories of probability. Academic Press, New York, NY.

FINE, K. 1974. An incomplete logic containing S4. Theoria, **40**: 23–29.

KOMOLGOROFF, A. 1950. Foundations of the theory of probability. Chelsea Publishing Company, New York, NY.

MCDERMOTT, D. 1987. A critique of pure reason. Computational Intelligence, **3**: 151–160.

MORGAN, C. 1982*a*. There is a probabilistic semantics for every extension of classical sentence logic. Journal of Philosophical Logic, **11**: 431–442.

_____ 1982*b*. Simple probabilistic semantics for propositional K, T, B, S4, and S5. Journal of Philosophical Logic, **11**: 443–458.

_____ 1984. Weak conditional comparative probability as a formal semantic theory. Zeitschrift für mathematische Logik, **30**: 199–212.

_____ 1988. Probability theory versus procedural pessimism. Computational Intelligence, **4**: 94–97.

PEARL, J. 1986. Fusion, propagation, and structuring in belief networks. Artificial Intelligence, **29**: 241–288.

POPPER, K. 1965. The logic of scientific discovery. Harper Torchbook edition, Harper and Row, Publishers, Inc., New York, NY.

ROBINSON, A. 1974. Introduction to model theory and to the metamathematics of algebra. Second revised printing. North-Holland Publishing Company, Amsterdam, The Netherlands.

THOMASON, S. 1974. An incompleteness theorem in modal logic. Theoria, **40**: 30–34.

WOODS, W. 1987. Don't blame the tool. Computational Intelligence, **3**: 228–237.