Chapter 3

Plausible Reasoning

> Talking of those who denied the truth of Christianity, he [Dr. Johnson] said, "It is always easy to be on the negative side. If a man were to deny that there is salt on the table, you could not reduce him to an absurdity. Come, let us try this a little further. I deny that Canada is taken, and I can support my denial by pretty good arguments. The French are a much more numerous people than we; and it is not likely that they would allow us to take it. 'But the ministry have assured us, in all the formality of the Gazette, that it is taken.' — Very true. But the ministry have put us to an enormous expense by the war in America, and it is their interest to persuade us that we have got something for our money. 'But the fact is confirmed by thousands of men who were at the taking of it.' — Ay, but those men have still more interest in deceiving us. They don't want that you should think the French have beat them but that they have beat the French. Now suppose you should go over and find that it is really taken, that would only satisfy yourself; for when you come home we will not believe you. We will say, you have been bribed. — Yet, Sir, notwithstanding all these plausible objections, we have no doubt that Canada is really ours. Such is the weight of common testimony. How much stronger are the evidences of the Christian religion?"
>
> — Boswell's *Life of Johnson*, July 14, 1763

The logics outlined in the previous chapter describe how rules of universal validity can be applied to facts known with absolute certainty to deduce other facts known with absolute certainty. Unfortunately, few of the rules that we use in everyday life are universally true, and few, if any, of our beliefs are completely certain to the degree that we cannot imagine them being overthrown given sufficient contrary evidence. (Possible candidates for certain belief include mathematical theorems, statements true by definition, and our knowledge of our own current mental state.) Therefore, deduction does not adequately characterize commonsense inference; we need also a description of how plausible, provisional conclusion may be drawn from uncertain or partial evidence.

A great variety of issues arise in the study of plausible inference. There are many different functions that a theory of plausible inference may serve; there are many types of complications that it involves. The problems to be addressed, the centrality of the problems that are addressed, and the interrelations between problems vary widely from one theory to the next. Below, we enumerate

some prominent issues in plausible inference. The items on the list are not intended to be disjoint, independent, or exhaustive.

Representing degree of belief. I am entirely certain that there is train service between New York and New Haven, because I have taken this train dozens of times. I would guess that there is train service between Portland, Oregon and Seattle, Washington, because they are major cities fairly close together. It would be useful to be able to represent the difference in my strength of belief in these two statements, reflected in the phrases "entirely certain" and "would guess".

Evaluating the strength of arguments. I would guess that there is probably train service between Portland and Seattle, because they are large cities close together. This is a fairly weak argument. If I called train information, and they told me that there was no train, then that would stronger evidence, and I would take that argument as overriding the first. This evidence could be counteracted in various ways; for example, if I called the same information number, and was told that there was no train service between New York and New Haven, which I know to be false, then I would lose faith in the train information, and be in a state of doubt about whether there was a train. If I went to the Portland station and took the train to Seattle, then that would be overwhelming evidence that there was such a train; it would take very powerful counter-evidence, which would explain how I was deceived, to shake my faith then. We would like our theory to provide a calculus in which we can calculate and compare the strengths of such arguments.

Applying rules of general but not universal validity. Standard logic justifies the use of universally quantified rules; rules that are always true without exception. Much commonsense inference relies on applying *default* rules that hold in general but not always. For instance, in the above example, we use a rule like, "If two cities are large and close together, then they are likely to be connected by train." In reasoning that I can call my wife from the office, I use rules like, "My phone generally works," and "My wife is generally home at six o'clock." These rules sometimes fail, but they are usually valid.

Avoiding the enumeration of all the conditions on a rule. It often happens that a plausible commonsense rule, when examined closely, has an almost unlimited number of possible types of exceptions. The problem of dealing with all these potential exceptions is known as the *qualification* problem. For example, a rule like, "To find out how my stocks are doing, I can buy a newspaper at the news stand and read it," should really read, "To find out how my stocks are doing, I can buy a paper at the news stand and read it, as long as the newspaper is in English, and I can find some light, and no lunatic passerby tears the newspaper from my hands, and the newspaper is not a fake distributed so as to spread disinformation, and I don't go blind or mad, and the editor of the newspaper has not decided to stop printing the stock prices, and ..." It would probably be impossible to state all the necessary conditions; it would certainly be impossible to verify them each time you wanted to read a newspaper. What you would like is to state the rule, "To find out how my stock are doing, I can buy and read a newspaper, unless something strange happens," where the "unless" clause is not to be verified unless there is substantial reason to worry about some particular anomaly.

Inference from the absence of information: It is often reasonable to infer that a statement ϕ is false from the fact that one does not know ϕ to be true, or from the fact that it is not stated to be true in a problem statement. Such an inference can take a number of forms:

a. The statement ϕ may be unlikely given other facts that I know. For example, if I hear of an animal or infer the existence of one, and I do not know that it is an albino, I am probably safe in assuming that it is not one, since albinism is a rare condition.

b. I may have reason to believe that, if ϕ were true, I would have learned about ϕ. For example, I know that Mark Twain was never Governor of New York, not because I have ever read that he was not, nor because I know all the governors, nor because I know everything that Mark Twain ever did, but just because I may safely presume that I would have heard about it if he had been. This is called an *auto-epistemic* inference [Moore, 1985].

c. If another person is communicating facts to me, it may be a convention that, if ϕ were true, he should have communicated it. For example, if I find the statement "There are more than half a million people in Boston," in an article, I am probably safe in concluding that there are fewer than a million people, since otherwise the author of the article should have written "More than a million." Such conventions are known as Gricean conditions [Grice, 1957]

Limiting the extent of inference. Many intuitively appealing sets of axioms have the property that the first few inferences all seem to be reasonable and to have reasonable conclusions, but that, as the inferences get further and further from the starting axioms, the conclusions seem less and less sensible, and they eventually end up in pure nonsense. The "Sorites" paradox has this form: If you have a heap of sand then you will still have a heap of sand after taking away one grain; hence, by induction, if you take away all the sand, you still have a heap. No single application of the rule is very objectionable, but too many together lead to trouble. The "Consequential Closure" problem (section 8.2) is similar: It is often useful to postulate that, if a person knows that p is true and he knows that p implies q, then he knows q. Certainly it would be strange to encounter a man who knew that Socrates was a man and that all men are mortal but really did not know that Socrates was mortal. Unfortunately, if we accept the axiom, then we can show that everyone knows all mathematical theorems, since they can be proven in this way one step at a time from the axioms. Again, no single application of the rule is troublesome. A way around these problems is to cast them as plausible rather than certain inferences, and to construct a theory in which chaining plausible inferences together leads to rapidly less plausible conclusions.

Inference using vague concepts. Inferences that involve reasoning near the boundaries of a vague concept are often uncertain. For instance, given that a man is tall, it is a safe inference that he is more than four feet tall, but it is a questionable inference that he is more than six feet tall. The theory of fuzzy logic [Zadeh, 1987] addresses such issues.

Finding expected utility. Often, we are faced with choosing between actions whose consequences are uncertain. In such a case, we would like to guide our actions in terms of reasonable combination of the likelihoods of the various outcomes with their desirability. For example, we would like to be able to reason that in Britain it is generally worthwhile carrying an umbrella while in Arizona it is not generally worthwhile, using our knowledge of the cost of carrying the umbrella, the cost of getting caught without it in the rain, and the likelihood of rain.

Inferring an explanation. Commonsense reasoners try to explain the reasons underlying their observations. If I observe that the street is wet, I infer that it rained. On the presentation of other evidence, I may adopt another explanation. For instance, if I observe that the sidewalk is not wet, I may decide instead that the street-cleaners have been by. A distinctive characteristic of explanations

is that generally only one explanation at a time is needed. Wet streets can be evidence either for rain or for street-cleaners, but not for both; external evidence in favor of one, such as seeing overcast skies or the street cleaning truck, will generally be taken as reducing the likelihood of the other, even though they are not, in fact, incompatible. [Pearl, 1988]

Schema-Based Inference: Many useful commonsense concepts correspond to large systems of relations that are instantiated in many separate instances in the world. For example, the concept of a house involves both physical relationships between many different physical components (walls, windows, floors, doors ...) and functional relations, such as being inhabited. A visit to the doctor involves a structure of events and scenes (waiting in the waiting room, seeing the nurse, ...) Such concepts are called *schemas* or *frames*. A schema is represented as a collection of *slots*, which are characteristic variables, and the relations on these slots. The entity corresponding to a slot in an instantiation of a schema is called the *filler* of the slot. Fillers may be atomic entities or they may themselves be frames.

There are at least four different kinds of inferences involved in the use of a schema (called *schema application*):

a. Schema identification: From finding a number of the fillers of slots satisfying appropriate relations and properties, infer the presence of the frame. For example, from seeing an outer wall, windows, and roof in their customary arrangement, infer that these are part of a house. From seeing a man waiting in a doctor's waiting room, infer that he is a patient waiting to see the doctor.

b. Slot prediction: Predict that the instance of the schema contains fillers of the slots. For example, predict that there is a kitchen in the house; predict that the patient at the doctor's office will be asked to take off his clothes.

c. Filler identification: Identify a real-world entity as a filler of a slot. For example, identify the real kitchen as filling the "kitchen" slot; identify a particular act of writing a check as filling the "pay doctor" slot.

d. Relation prediction: Predict that the relations defined on the slots will hold on the specific fillers. For example, predict that the walls of the house support the roof; predict that the patient's taking off his clothes will come after he has been brought to an inner room.

In general, these inferences are plausible rather than certain inferences. An office building may look like a house from the outside; a house may lack a kitchen; a patient may write any number of other checks while at the doctor's; the roof of a house may be supported by some structure other than the walls. These are default inferences, generally valid but not always.

Analogy. If A is similar to B in some ways then it may be reasonable to assume that it is similar in other related ways as well. For example, if you meet a man who pronounces the words "car", "idea" and "fear" in the same way as your next-door neighbor, you may assume that he will also have the same pronunciation of "tire".

Inferring a general rule from examples. People are always on the lookout for general rules that encapsulate their observations. If I get sick to my stomach after eating a Brazil nut for the first time, I might construct the general rule, "Brazil nuts do not agree with me." If I meet four graduates of Pretentious U., and all four are unbearable, I might construct the general rule, "All graduates

of Pretentious U. are unbearable." No such inferences are sound; they can be overthrown either by bringing a counter-example (an experience of eating a Brazil nut without trouble; a charming graduate) or bringing a better alternative explanation. For example, if I observe that everyone else I know also had stomach trouble this week, whether or not they ate the nut, I may conclude that the trouble was due to a contagious stomach disease.

AI researchers and logicians have constructed a number of theories that address the above issues. Section 3.1 discusses various non-monotonic logics, including default logic and circumscription. Section 3.3 discusses probability theory. Analogy and generalization are not discussed here.

In subsequent chapters of this book, we will not, in fact, use any of these formal theories. Rather, on the rather rare occasions where we discuss plausible inference, we will either describe a narrow domain-specific inference rule, or we will use a bogus notation, "plausible(ϕ, ψ)", meaning, "Given ϕ, it is a plausible inference that ϕ is true, in the absence of contradictory evidence," and we will leave it open how this rule can, in fact, be reasonably defined and applied.

3.1 Non-monotonic Logic

3.1.1 Non-monotonicity

A major difference — arguably, the defining difference — between sound deductive inference and plausible inference is monotonicity. Deductive inference is monotonic in the following sense: If a sentence ϕ is a valid conclusion from a set of sentences Γ, and Γ is a subset of Δ, then ϕ is a valid conclusion from Δ. Symbolically, if $\Gamma \subseteq \Delta$ and $\Gamma \vdash \phi$ then $\Delta \vdash \phi$; if $\Gamma \models \phi$ then $\Delta \models \phi$. Plausible inference does not have this property; inferences are made provisionally, and they may be withdrawn if more evidence contravening them comes along.

(The statement is sometimes made that classical probability theory, such as we will discuss in section 3.3, is a monotonic theory. This is true at the meta-level; the statement, "Given Γ, ϕ has probability x," does not change its truth. But, by the same token, the statements of non-monotonic logic are monotonic at the meta-level; the statement "Given Γ, ϕ is a plausible inference," does not changes its truth. At the object level, probability theory is non-monotonic. The statement ϕ may become more or less likely as increasing evidence is accumulated.)

This non-monotonicity has profound consequences for theories of plausible inference. Monotonicity is deeply built into standard logic, both in the syntactic concept of a proof, and in the semantic concept of semantic consequence; a theory of plausible inference must therefore alter or ignore these basic concepts. In standard systems of proofs, the validity of a proof step depends only on particular steps being earlier in the proof or in the set of hypotheses. Adding additional axioms to the hypothesis set cannot invalidate a proof step. Therefore, a proof of statement ϕ from axioms Γ is also a proof of ϕ from any Δ containing Γ. The axioms of Γ are also axioms of Δ, and any inferences valid in proofs from Γ is also valid in proofs from Δ. In non-monotonic theories, an inference may require the absence as well as the presence of information; it may therefore be valid in Γ and invalid in Δ. Thus, an inference may involve the entire structure of the theory as a whole, not just some particular set of axioms in the theory.

Similarly, in the standard concept of semantic consequence, ϕ follows from Γ ($\Gamma \models \phi$) just if every model $\mathcal{M}$ that satisfies Γ also satisfies ϕ. This definition gives us the monotonicity property. Suppose that $\Gamma \models \phi$ and that $\Delta \supset \Gamma$. By the definition of consequence, if $\mathcal{M}$ is a model satisfying Γ, then $\mathcal{M}$ satisfies ϕ. Clearly, any model that satisfies Δ must also satisfy Γ, since Γ is a subset of Δ. Therefore any model satisfying Δ also satisfies ϕ, which is our definition of $\Delta \models \phi$. Therefore to describe plausible inference in terms of models, we will either have to change the definition of a model satisfying a statement, or our definition of semantic consequence, or both.

The non-monotonicity of plausible inference invalidates many types of inference familiar from ordinary logic; one must be careful about applying one's logical intuitions. For example, in deductive logic, if r follows from q and q follows from p then r follows from p. In plausible inference, this does not hold. Let p be the sentence "John is a naturalized U.S. citizen," let q be the sentence "John is a U.S. citizen," and r is the sentence, "John was born in the U.S." Then q is a plausible (indeed, a necessary) inference from p, and r is a plausible inference from q, but r is not a plausible inference from p. Similarly, in standard logic, if it possible to infer a statement "$\alpha(\gamma)$" from a hypothesis set that contains "$\beta(\gamma)$" but no other mention of the constant symbol γ, then it is possible to infer "$\forall_X \beta(X) \Rightarrow \alpha(X)$." In plausible inference this close connection between universal generalization and reasoning from partial knowledge does not hold. It may be plausible to infer "John was born in the U.S." from "John is a U.S. citizen," but that does not legitimate the inference, "Every U.S. citizen was born in the U.S."

A non-monotonic logic is an extension of standard logic that supports some types of plausible inferences. Non-monotonic logics address the problems of the non-monotonicity of plausible inference, and of inference from the absence of information; they do not, typically, allow the expression of the degree of belief, the comparison of strengths of arguments, or the weakening of an argument under many inference steps.

3.1.2 Domain-Independent Rules

Domain Closure

One basic type of inference often used in problem solving is to assume that the entities specified in the problem are the only entities, or the only entities of a given sort, that are relevant to its solution. For example, the initial conditions of a blocks world problem might be given by specifying that block A is on block B and blocks B and C are on the table. The problem solver should then assume that blocks A, B, and C are the only blocks relevant to the problem, and that, in particular, there are no other blocks on block A. The missionaries and cannibals problem may specify, "There are three missionaries, three cannibals, and a boat;" to address the problem one must assume that there are no other people, and no other ways to cross the river.

The ultimate justification of this inference depends on the source of the information. If the problem was given as a natural language communication, then it is an application of a Gricean rule: The person formulating the problem should give all the relevant information, so if there were another block, he should have mentioned it. If the table were seen, then it can be inferred that any other blocks would have been seen as well. However, in all these cases, the inference takes the same form, and can be handled in the same way.

This kind of inference is called the *domain closure assumption*. It can be formalized as follows: Given a theory $\mathcal{T}$, assume that the only objects that exist are those named by ground terms in the language of $\mathcal{T}$; that is, constant symbols and the application of function symbols to constant symbols. Thus, in a blocks world problem where the only constant symbols were "table", "a", "b", and "c", and where there are no function symbols, the domain closure assumption specifies that these were the only objects; i.e.

$$\forall_X \; X = \text{a} \; \vee \; X = \text{b} \; \vee \; X = \text{c} \; \vee \; X = \text{table}$$

In a problem in family relations, if the constants are "betty" and "mordred" and the function symbols are "father_of(X)", "mother_of(X)", and "common_ancestor(X, Y)", then the assumption specifies that the only objects are made by composing these functions with these constants, in terms such as "betty", "father_of(mordred)",
"common_ancestor(mother_of(betty), mordred)", and so on. This assumption cannot be written as a first-order formula.

The domain closure assumption is non-monotonic, since if $\mathcal{T}$ is augmented with axioms that use new constant or function symbols, then the assumption becomes weakened to allow terms made from these.

The assumption is often applied only to entities of certain sorts. For instance, in planning problems, it might be reasonable to suppose that all the relevant physical objects have been explicitly named, but not that all instants of time or points of space have been named. In this case, the domain closure assumption could be applied only to physical objects.

The assumption depends only on the symbols used in $\mathcal{T}$, not on the content of the theory $\mathcal{T}$. In fact, the assumption yields powerful consequences even when applied to a vacuous theory, such as one with the single axiom "f(a) = f(a)". Applying the domain closure assumption to this theory has as a consequence the assumption that the only entities are a, f(a), f(f(a)) and so on.

If there are no function symbols in the language and only finitely many constant symbols "a", "b", and "c", then the domain closure assumption is completely stated in the axiom

$$\forall_X \; X = \text{a} \; \vee \; X = \text{b} \; \vee \; X = \text{c}$$

Thus, a proof of ϕ from $\mathcal{T}$ with the domain closure assumption can take the following form:

1. Verify that no function symbols appear in $\mathcal{T}$ and that the only constant symbols are "a", "b", and "c". Increment the theory $\mathcal{T}$ by adding the axiom

$$\forall_X \; X = \text{a} \; \vee \; X = \text{b} \; \vee \; X = \text{c}$$

2. Verify that there is a proof of the standard sort of ϕ from the augmented version of $\mathcal{T}$.

Step (1) is not a proof step of an ordinary kind. It is a non-monotonic operation; if additional axioms are added to the starting version of $\mathcal{T}$, the domain closure axiom may be weakened by allowing the possibility of other constant values.

For example, let $\mathcal{T}$ contain the following axioms:

1. on(a,b).
2. on(b,table)
3. on(c,table)
4. $\text{on}(X, Y) \wedge Y \neq Z \Rightarrow \neg\text{on}(X, Z)$. ($X$ can only be on one object at a time.)
5. $\neg$on(table,X). (The table is not on anything.)
6. $\text{clear}(X) \Leftrightarrow \forall_Y \neg\text{on}(Y, X)$ (An object is clear if nothing is on it.)
7. distinct(a,b,c,table).

Monotonically, $\mathcal{T}$ does not support the inference "clear(a)", since these axioms do not rule out the possibility that there is some other block d on a. However, we can derive this result through the domain closure inference. Since the only constant symbols in $\mathcal{T}$ are "a", "b", "c", and "table", and since there are no function symbols, the domain closure axiom allows us to add the axiom

DC. $\forall_X\ X = \text{a} \vee X = \text{b} \vee X = \text{c} \vee X = \text{table}$

We can then prove the desired result as follows:

From [1], [4], and [7] infer

8. $\neg$on(a,a)

From [2], [4], and [7] infer

9. $\neg$on(b,a)

From [3], [4], and [7] infer

10. $\neg$on(c,a)

From [5] infer

11. $\neg$on(table,a)

From [DC], [8], [9], [10], and [11], infer

12. $\forall_X\ \neg$on(X,a)

From [12] and [6] infer

13. clear(a).

If there are function symbols in the language, then the effect of the domain closure assumption cannot be expressed in any first-order axiom, or, indeed, in any recursive axiom schema. (Proof: Adding the domain closure assumption to the standard axioms of the integers is sufficient to rule out non-standard models of those axioms, and restrict the model to the "standard" integers. But there is no complete recursive axiomatization of the integers.) Thus, there is no complete categorization of proof using the domain closure assumption. We can, however, categorize various limited classes of proofs. For example, suppose that the constant symbols of $\mathcal{T}$ are "a" and "b", and the function symbols are "f" and "g". Then, clearly, the domain closure assumption entails the following axiom:

$$\forall_X\ X = \mathrm{a} \ \vee\ X = \mathrm{b} \ \vee\ \exists_Y\ X = \mathrm{f}(Y) \ \vee\ X = \mathrm{g}(Y)$$

Thus, any proof from $\mathcal{T}$ together with the above axiom is a valid proof from $\mathcal{T}$ with the domain closure assumption.

If the starting theory has at least one constant symbol, and it is fully Skolemized — that is, all existentially quantified variables are replaced by Skolem constants or terms — then the set of terms is the Herbrand universe of the theory, and the domain closure assumption is guaranteed to be consistent. If the theory contains sentences with existential quantifiers, then the domain closure assumption may not be consistent.

For example, consider the theory $\mathcal{T}$ with the following axioms:

A. $\exists_X \mathrm{p}(X)$

B. $\neg \mathrm{p}(\mathrm{a})$

Since the only constant symbol is "a" and there are no function symbols, the domain closure assumption is

C. $\forall_X X = \mathrm{a}$

However, [C] is obviously inconsistent with [A] and [B].

Closed World Assumption

In the domain closure assumption, we postulate that all the entities relevant to a given problem are mentioned in the problem statement. Similarly, it is often reasonable to assume that all the significant relations between entities in the problem are given in the problem statement. If we identify a significant relation as one that has a predicate symbol, and a significant entity as one named by a ground term, then we can formalize this assumption as follows: Let $\mathcal{T}$ be a theory, and let ϕ be a ground atomic formula – that is, a formula without variables or boolean connectives, including negation. If ϕ cannot be inferred from $\mathcal{T}$, then non-monotonically infer $\neg\phi$.

For example, using the closed world assumption in the above blocks world problem enables us to simplify the starting theory by eliminating several axioms. Let $\mathcal{T}$ have the following axioms:

1. on(a,b)

2. on(b,table)

3. on(c,table)

We now apply the closed world assumption to $\mathcal{T}$. Since "on(a,a)" is clearly not a consequence of $\mathcal{T}$, by the closed world assumption, we can add the axiom "¬on(a,a)". Similarly, since "on(b,a)", "on(c,a)" and "on(table,a)" are not consequences of $\mathcal{T}$, we can add the axioms "¬on(b,a)", "¬on(c,a)" and "¬on(table,a)". We now make the domain closure assumption

DC. $\forall_X\ X = \mathrm{a} \vee X = \mathrm{b} \vee X = \mathrm{c} \vee X = \mathrm{table}$

From these axioms, we can infer directly that nothing is on block a.

14. $\forall_X \neg\mathrm{on}(X, \mathrm{a})$

This is not quite the conclusion we arrived at before. We would like use 14 to infer "clear(a)" from the definition of "clear". Here, however, there is a difficulty. Let $\mathcal{T}_2$ be a theory containing [1], [2], and [3] as above and also axiom [6] defining clear:

6. $\mathrm{clear}(X) \Leftrightarrow \forall_Y \neg\mathrm{on}(Y, X)$

On the one hand, the closed world assumption and domain closure justify [14] in the same way as before and by combining [14] with [6] we can infer "clear(a)". However, we can also apply the closed world assumption to $\mathcal{T}_2$ in a different way: Since $\mathcal{T}_2$ certainly does not by itself imply "clear(a)", therefore the closed world assumption justifies the inference "¬clear(a)".

In short, the closed world assumption, applied across the board, is inconsistent for the theory $\mathcal{T}_2$. The solution, therefore, is not to apply it across the board, but to specify that it applies to some predicates but not others. In this example, we would wish to apply it to the predicate "on(X, Y)", but not to the predicate "clear(X)". In reasoning about people at a party, it might be reasonable to apply the closed world assumption to predicates that describe who is at the party, but not to predicates that describe their personal relations.

The closed world assumption is non-computable in the general case, since the statement "p(a) cannot be derived from $\mathcal{T}$" is non-computable.

The closed world assumption restricted to the equality predicate is known as the (non-monotonic) *unique names assumption*. The unique names assumption asserts that two distinct terms t_1 and t_2 may be assumed to be unequal unless they are demonstrably equal. In our blocks world example, the unique names assumption is equivalent to the axiom "distinct(a,b,c,table)."

It is suggestive of the power of these non-monotonic inference rules that, if we start with the constant symbol 0 and the function symbol s(X) for the successor function, and no proper axioms at all, and we apply the domain closure assumption and the unique names assumption, the result is the complete theory of the integers. The unique names assumption asserts that the terms 0, s(0), s(s(0)), and so on, are all distinct. The domain closure assumption asserts that these terms comprise all the integers.

All domain-independent types of non-monotonic inference depend strongly on particularly syntactic features of the language used; if the language is changed to some logically equivalent form, the non-monotonic inferences may all be changed. Domain closure depends on the use of particular constant and function symbols. For example, we may formulate a blocks world theory, either using the predicate "on(X, Y)" or using the predicate "off(X, Y)". The content of the two theories will be exactly equivalent. However, the effect of applying the closed world assumption will be exactly opposite. Similarly, we may formulate a theory of families, either using a function symbol "father_of(X)" or with a predicate symbol "father(Y, X)". The content of the two theories will be equivalent, if, in the latter case, we include an axiom stating that each person has exactly one father: $\forall_X \exists^1_Y \text{father}(Y, X)$. However, the domain closure assumption can be effectively applied only to the first, while the closed world assumption can be applied only to the second.

It should be noted that it is possible to have an algorithm whose behavior depends on the absence of some data structure but which can nonetheless be justified purely in terms of a monotonic logic. It all depends on the conventions for interpreting the data structure in terms of the logic. For example, consider an algorithm that computes least-cost paths in a weighted graph. The algorithm is non-monotonic with respect to the data structure; if you add additional edges of lower cost, the algorithm will arrive at a different answer. Whether this algorithm represents non-monotonic inference depends on the meaning of the graph, and the meaning of the answer. If the graph is defined as meaning just that the edges it records exist, and the answer is intented to be the exact cost of the least cost path, then the algorithm is performing the non-monotonic inference that other edges do not exist. If, however, the graph is defined as containing all relevant edges — that is, if the input would be considered incorrect, not merely incomplete, if there turned out to be other edges — then the inferences of the algorithm are entirely sound. The data structure has merely encoded the completeness statement, "No other edges exist besides these," in an implicit rather than explicit form. The inference is truly non-monotonic only if there are circumstances under which this reading of the data structure can be overruled. Likewise, if the answer is taken to be the cost of an inexpensive path, rather than necessarily the cost of the least expensive path, then the inference is monotonic, however the graph is interpreted.

3.1.3 Circumscription

Plausible inference often has the form "Assume that as few objects as possible have a particular property." The closed-world assumption, discussed above, is a simple form of this inference; we assume that the only atomic ground formulas that are true are those that can be proven within the theory. However, we often want to apply this assumption to non-ground formulas as well. For example, the default assumption, "Assume that eggs are fresh," can be phrased as, "Assume that the class of rotten eggs is as small as possible." Given this default assumption, we would like to be able to assume that all the eggs in the box we are carrying are fresh. However, unless we have a constant symbol for each egg in the box, the closed world assumption will not support this conclusion. Instead, we can use the method of *circumscription*.

The rotten eggs example above turns out to involve some subtleties, so we will start with a simpler example. Let us suppose that we know that kangaroos and opossums are marsupials, and we know that very few animals are marsupials. We know that cats are neither kangaroos nor opossums. We

would like to infer by default that cats are presumably not marsupials.

Our basic theory $\mathcal{T}$ consists of the following axioms:

1. $\forall_X$ kangaroo$(X) \Rightarrow$ marsupial(X)
2. $\forall_X$ opossum$(X) \Rightarrow$ marsupial(X)
3. $\forall_X$ cat$(X) \Rightarrow \neg$(kangaroo$(X) \vee$ opossum(X))

We want to force the set of marsupials to be as small as possible, consistent with the theory $\mathcal{T}$. We can formulate this requirement as follows: Let $\alpha(X)$ be any possible property of X which would satisfy the axioms on "marsupial" in $\mathcal{T}$. Then the real marsupials should not be a proper superset of α; otherwise, we could reduce the class of marsupials to α, and still satisfy the axioms of $\mathcal{T}$. That is to say, if α satisfies the conditions

$$\forall_X \text{ kangaroo}(X) \Rightarrow \alpha(X)$$
$$\forall_X \text{ opossum}(X) \Rightarrow \alpha(X)$$

then it cannot be the case that all α's are marsupials and that some marsupials are not α's. We can write this rule as the axiom schema

4. $[\forall_X \text{ kangaroo}(X) \Rightarrow \alpha(X) \wedge \forall_X \text{ opossum}(X) \Rightarrow \alpha(X)\ ] \Rightarrow$
   $\neg[\forall_X \alpha(X) \Rightarrow \text{ marsupial}(X) \wedge \exists_X \text{ marsupial}(X) \wedge \neg\alpha(X)\ ]$

or, equivalently,

5. $[\forall_X \text{ kangaroo}(X) \Rightarrow \alpha(X) \wedge$
   $\forall_X \text{ opossum}(X) \Rightarrow \alpha(X) \wedge$
   $\forall_X \alpha(X) \Rightarrow \text{ marsupial}(X)\ ] \Rightarrow$
   $\forall_X \text{ marsupial}(X) \Rightarrow \alpha(X)$

The above statement is true of every property α. In a second order logic, we can quantify over all α, and state the axiom

$$\forall_\alpha\ [\ \forall_X \text{ kangaroo}(X) \Rightarrow \alpha(X) \wedge$$
$$\forall_X \text{ opossum}(X) \Rightarrow \alpha(X) \wedge$$
$$\forall_X \alpha(X) \Rightarrow \text{ marsupial}(X)\ ] \Rightarrow$$
$$\forall_X \text{ marsupial}(X) \Rightarrow \alpha(X)$$

In a first order logic, we can assert formula (5) as an axiom schema, applying to each open formula α. Since we have not discussed higher-order logic, we will take this approach. (Using higher-order logic has some technical advantages. [Genesereth and Nilsson, 1987].)

Formula 5, above, is known as the *circumscriptive* axiom schema. The *circumscription* of theory $\mathcal{T}$ in the predicate "marsupial", written CIRC[$\mathcal{T}$;marsupial]) is the theory $\mathcal{T}$ with the circumscriptive axiom schema added.

We can easily show that the only marsupials are kangaroos or opossums in the circumscribed theory. Pick $\alpha(X)$ to be the formula "kangaroo$(X) \vee$ opossum(X)". Substituting this formula for α in the axiom schema gives the axiom

6. $[\forall_X$ kangaroo$(X) \Rightarrow$ [kangaroo$(X) \vee$ opossum$(X)] \wedge$
   $\forall_X$ opossum$(X) \Rightarrow$ [kangaroo$(X) \vee$ opossum$(X)] \wedge$
   $\forall_X$ [kangaroo$(X) \vee$ opossum(X)] $\Rightarrow$ marsupial(X)] $\Rightarrow$
  $\forall_X$ marsupial$(X) \Rightarrow$ kangaroo$(X) \vee$ opossum(X)

Of the three implications in the antecedent of axiom 6, the first two are trivial, and the third is just a restatement of axioms (i) and (ii) of $\mathcal{T}$. Therefore, the antecedent of the implication is true, which means that the consequent is also true. Thus we have the formula

7. $\forall_X$ marsupial$(X) \Rightarrow$ [kangaroo$(X) \vee$ opossum(X)]

i.e. all marsupials are kangaroos or opossums. Using this and axiom (iii) of $\mathcal{T}$, it follows directly that no cats are marsupials.

We have thus shown that formula (7) above, which states that all marsupials are either kangaroos or opossums, is a consequence of the circumscriptive axiom schema (5). In fact, (7) is equivalent to (5); that is, if $\mathcal{F}$ is a theory containing $\mathcal{T}$ and (7), then all instances of axiom schema (5) are true in $\mathcal{F}$. (Exercise 6: Prove this.) Thus, in this case, circumscribing the theory has the effect of adding a single first-order axiom. This is not true in all cases of circumscription, but it is true in many important cases.

If we extend $\mathcal{T}$ by telling it about some other marsupials, then we change the circumscriptive axiom. For example, let $\mathcal{W}$ be the theory containing axioms (1), (2), and (3) of $\mathcal{T}$, together with axiom (8), asserting that all koala bears are marsupials.

8. $\forall_X$ koala$(X) \Rightarrow$ marsupial(X)

i.e. all koalas are marsupials. Then the circumscription of $\mathcal{W}$ in "marsupial" adds the axiom schema,

9. [$\forall_X$ kangaroo$(X) \Rightarrow \alpha(X) \wedge$
   $\forall_X$ opossum$(X) \Rightarrow \alpha(X) \wedge$
   $\forall_X$ koala$(X) \Rightarrow \alpha(X) \wedge$
   $\forall_X \alpha(X) \Rightarrow$ marsupial(X)] $\Rightarrow$
  $\forall_X$ marsupial$(X) \Rightarrow \alpha(X)$

Substituting the formula "kangaroo$(X) \vee$ opossum$(X) \vee$ koala(X)" for $\alpha(X)$ in (9), we can infer the conclusion,

10. marsupial$(X) \Leftrightarrow$ [kangaroo$(X) \vee$ opossum$(X) \vee$ koala(X)]

Note that by strengthening the theory $\mathcal{T}$ to $\mathcal{W}$, we have weakened the circumscriptive axiom schema, and therefore have weakened the conclusions that can be drawn. We can no longer deduce that cats are not marsupials, unless we have an additional axiom that no cats are koala bears.

We can describe this technique formally as follows: Let $\mathcal{T}$ be a finite theory and let $\mu(X)$ be a predicate symbol appearing in $\mathcal{T}$. (X can be a single variable or a tuple of variables.) For any open formula $\alpha(X)$, let $\mathcal{T}[\mu/\alpha]$ be the theory formed from $\mathcal{T}$ by replacing every occurrence of μ by α. Then we define the circumscription of $\mathcal{T}$ in μ, CIRC$[\mathcal{T}, \mu]$ as the theory $\mathcal{T}$ together with the axiom schema

$$[\mathcal{T}[\mu/\alpha] \wedge [\forall_X\ \alpha(X) \Rightarrow \mu(X)]] \Rightarrow [\ \forall_X\ \mu(X) \Rightarrow \alpha(X)\ ]$$

where $\alpha(X)$ is any open formula. (In this formula, the antecedent $\mathcal{T}[\mu/\alpha]$ guarantees that α has all the properties that μ must have; the antecedent $\forall_X\ \alpha(X) \Rightarrow \mu(X)$ guarantees that α fits within μ; the consequent $\forall_X\ \mu(X) \Rightarrow \alpha(X)$ guarantees that μ fits inside α. Thus, the formula may be read, "If α is any property that satisfies the theory $\mathcal{T}$, and that fits inside μ, then α must just be μ." Thus, μ must be a minimal set satisfying $\mathcal{T}$.)

As mentioned above, as $\mathcal{T}$ is strengthened, the antecedent of the circumscription axiom schema is strengthened, and the axiom as a whole is weakened.

The original inference (7) that, if X is a marsupial then it is either a kangaroo or an opossum, can only be used in the backwards direction, to infer that something that is known not to be a kangaroo or opossum may be presumed not to be a marsupial. It cannot be used to deduced that something that is known to be a marsupial should be presumed to be either a kangaroo or an opossum, for, if the theory is extended to include other marsupials, the circumscriptive axiom will change. For example, if we add to the theory $\mathcal{T}$ the statement that Pete is a marsupial, the circumscriptive axiom will become

11. $[\forall_X\ \text{kangaroo}(X) \Rightarrow \alpha(X) \wedge$
$\forall_X\ \text{opossum}(X) \Rightarrow \alpha(X) \wedge$
$\alpha(\text{pete}) \wedge$
$\forall_X \alpha(X) \Rightarrow \text{marsupial}(X)\ ] \Rightarrow$
$\forall_X\ \text{marsupial}(X) \Rightarrow \alpha(X)$

or, equivalently,

12. $\forall_X\ \text{marsupial}(X) \Leftrightarrow [\text{kangaroo}(X) \vee \text{opossum}(X) \vee X = \text{pete}]$

Note, also, that, in order to apply the circumscriptive axiom [5] to deduce that a given animal is not a marsupial, it is necessary to know that it is not a kangaroo or opossum. One cannot deduce that it is not a marsupial if one is completely ignorant about the animal. We can get around this limitation by circumscribing over the three predicates "marsupial(X)", "kangaroo(X)" and "opossum(X)". Circumscription in parallel is a straightforward extension of circumscription over a single predicate. Consider the theory $\mathcal{T}_2$ containing the following axioms:

2.1. $\forall_X\ \text{kangaroo}(X) \Rightarrow \text{marsupial}(X)$

2.2. $\forall_X\ \text{opossum}(X) \Rightarrow \text{marsupial}(X)$

2.3. kangaroo(kanga).

2.4. opossum(george).

We require that the extension of the predicates "marsupial", "kangaroo", and "opossum" be as small as possible. To do this, we use three variable properties α, β, and γ, and require that, if α, β, and γ satisfy the theory $\mathcal{T}_2$ when substituted for "marsupial", "kangarooo", and "opossum", then they cannot be strict subsets of the desired predicates. The resultant axiom schema is

2.5. $[\ [\ \forall_X\ \beta(X) \Rightarrow \alpha(X)\ ] \wedge [\ \forall_X\ \gamma(X) \Rightarrow \alpha(X)\ ] \wedge \beta(\text{kanga}) \wedge \gamma(\text{george}) \wedge$
$[\ \forall_X\ \alpha(X) \Rightarrow \text{marsupial}(X)\ ] \wedge$
$[\ \forall_X\ \beta(X) \Rightarrow \text{kangaroo}(X)\ ] \wedge$
$[\ \forall_X\ \gamma(X) \Rightarrow \text{opossum}(X)\ ]\ ] \Rightarrow$
$[\ [\ \forall_X\ \text{marsupial}(X) \Rightarrow \alpha(X)\ ] \wedge$
$[\ \forall_X\ \text{kangaroo}(X) \Rightarrow \beta(X)\ ] \wedge$
$[\ \forall_X\ \text{opossum}(X) \Rightarrow \gamma(X)\ ]\ ]$

It is easily shown (exercise 7) that the circumscribed theory, CIRC[$\mathcal{T}_2$; marsupial,kangaroo,opossum], of $\mathcal{T}_2$ together with axiom schema (2.5) is equivalent to the single axiom (2.6).

2.6. $\forall_X\ [\ \text{marsupial}(X) \Leftrightarrow X = \text{kanga} \vee X = \text{george}\ ] \wedge$
$[\ \text{kangaroo}(X) \Leftrightarrow X = \text{kanga}\ ] \wedge$
$[\ \text{opossum}(X) \Leftrightarrow X = \text{george}\ ]$

Let us see what happens when we apply these techniques to the fresh eggs problem. We wish to infer that any given egg is fresh, unless they are known to be otherwise; that is, within the class of eggs, we wish to maximize the class of those that are fresh. To apply circumscription, we must pose the problem in terms of a predicate to be minimized. We can do this by introducing the new predicate "rotten_egg(X)". We give an axiom stating that an egg which is not a rotten egg is fresh; we can then achieve our goal of making as many eggs as possible fresh by minimizing the extension of "rotten_egg". (Such a predicate is often called an "abnormality" predicate, and given a names such as "ab(X)")

Thus we start with a single axiom

3.1 $\forall_X\ \text{egg}(X) \wedge \neg\text{rotten\_egg}(X) \Rightarrow \text{fresh}(X)$

and we circumscribe this theory in "rotten_egg". The result is the axiom schema

3.2. $[\ [\ \forall_X\ \text{egg}(X) \wedge \neg\alpha(X) \Rightarrow \text{fresh}(X)\ ] \wedge$
$[\ \forall_X\ \alpha(X) \Rightarrow \text{rotten\_egg}(X)\ ]\ ] \Rightarrow$
$\forall_X\ \text{rotten\_egg}(X) \Rightarrow \alpha(X)$

However, this circumscribed theory is equivalent to the axiom

3.3. $\forall_X\ \text{rotten\_egg}(X) \iff \text{egg}(X) \wedge \neg\ \text{fresh}(X)$

In other words, all that circumscription has done for us here is to restrict the extension of "rotten_egg" to eggs that are not fresh. Something has gone awry.

The problem is that our formalism treats "egg" and "fresh" symmetrically, while, in reality, we are thinking of them in quite different ways. We can rephrase axiom (3.1) in the form

3.1a. $\forall_X$ [egg(X) $\wedge$ $\neg$fresh(X)] $\Rightarrow$ rotten_egg(X)

There are three ways that we can make the extension of "rotten_egg" as small as possible, while preserving the truth of axiom (3.1a) above:

i. Disallow anything from satisfying "rotten_egg(X)" unless it is an egg that is not fresh.

ii. Make the extension of "egg" as small as possible within the class of unfresh objects.

iii. Make the extension of "fresh" as large as possible within the class of eggs.

Simple circumscription, as we have seen, accomplishes only (i). In reality, what we primarily want to accomplish is (iii). (i) is irrelevant and (ii) is misdirected. We can achieve our aim by using circumscription with variable predicates.

In circumscription with variable predicates, we minimize the extension of one predicate (in this case, "rotten_egg"), letting certain other, specified predicates (in this case, "fresh") range over all possible extensions. That is, we choose a value for "rotten_egg" and a value for "fresh" such that the extension of "rotten_egg" be as small as possible, subject to the constraint that the starting theory be true.

Formally, if α and β are predicates that satisfy the given theory when substituted for "rotten_egg", and "fresh" respectively, then the extension of α may not be a proper subset of the extension of "rotten_egg". We state this in the axiom schema

3.4. [[$\forall_X$ egg(X) $\wedge$ $\neg\beta(X)\Rightarrow\alpha(X)$] $\wedge$
[$\forall_X$ $\alpha(X)$ $\Rightarrow$ rotten_egg(X)]] $\Rightarrow$
$\forall_X$ rotten_egg(X) $\Rightarrow$ $\alpha(X)$

It is easily shown (exercise 8) that this axiom schema is satisfied only if all eggs are fresh and none are rotten. Thus, all the eggs in our box may assumed to be fresh. If we knew of some particular rotten eggs, a similar circumscription would allow us to assume that the only rotten eggs in the world were those known to be rotten.

3.1.4 Default Theory

One basic form of plausible reasoning involves the use of *default* inferences; inferring that an element of a given class has properties characteristic of the class. For example — the standard example in the field — given that Tweety is a bird and that birds can typically fly, infer provisionally that Tweety can fly. This inference is non-monotonic, since it can be overruled by additional information that Tweety is an ostrich, or is injured, or can't fly.

Reiter [1980a] describes a logic in which rules like "Birds typically can fly" can be stated as a tentative rule of inference. This particular rule is written[1] in the form

$$\frac{\text{bird(X)} : \text{can\_fly(X)}}{\text{can\_fly(X)}}$$

or, inline, as "bird(X) : can_fly(X) / can_fly(X)". The general form of a default inference rule is

$$\frac{\alpha : \beta}{\gamma}$$

The meaning of this rule is roughly, "If you believe α and you have no reason to doubt β then you are justified in believing γ." The formula α is called the *prerequisite* of the rule; β is the *justification*; γ is the *consequent.* Frequently, default rules take the form "$\alpha : \beta/\beta$", with identical justification and consequent. Such a rule is said to be *normal*; it can be read "α's are typically β's" or "If α then assume β."

Defining the consequences of a theory containing axioms and default rules is tricky, because the condition "no reason to doubt β" has a difficult circularity; whether one has reason to doubt β may depend, circularly, on whether one can apply the rule in which β is a justification. For example, consider a theory containing the two rules, "If someone speaks Turkish fluently, then assume that he is a Turkish citizen," and "If someone lives in the US, assume that he is a US citizen", and the three facts, "Kemal speaks Turkish fluently," "Kemal lives in the US," and "No one is both a Turkish and an American citizen." One way of thinking about this theory would be to deduce from the first rule that Kemal is Turkish and use that conclusion to establish doubt of the justification "Kemal is an American citizen" and thus block the use of the second rule. The use of the first rule is now legitimated, since there is no reason to doubt that Kemal is Turkish. Alternatively, one can do the reverse; use the second rule to deduce that Kemal is American and thus block the justification of the first rule. Another possibility is to take the disjunction of these two, and conclude that Kemal is either Turkish or American. Or perhaps the best thing under the circumstances is to abstain from concluding anything at all about Kemal's nationality.

In Reiter's logic, either of the first two arguments is valid, but they cannot be combined. Each argument gives a possible way of looking at the world, given these facts and default rules. The view is that the proper use of a default theory $\mathcal{T}$ is to find one way of choosing one's belief that are "consistent" with $\mathcal{T}$; if there is more than one such belief set, the reasoner can choose between them arbitrarily. Specifically, one looks for an *extension* $\mathcal{E}$ of $\mathcal{T}$, a set of sentences with the following properties: (i) $\mathcal{E}$ contains the sentences (not the default rules) of $\mathcal{T}$. (ii) $\mathcal{E}$ is closed under ordinary logical inference. (iii) $\mathcal{E}$ contains the conclusions of every applicable default rule in $\mathcal{T}$. A default rule is applicable if $\mathcal{E}$ contains its prerequisite, but does not contain the negation of any of its justifications. (iv) Every sentence in $\mathcal{E}$ can be justified by a finite-length derivation, where each step of the derivation is either an ordinary first-order inference or the application of an applicable default rule.

[1]See exercise 11 for an alternative representation of the rule "Birds can typically fly" within Reiter's theory with some technical advantages.

The catch here is that the definition of "applicable" depends on $\mathcal{E}$ itself. In formulating our definition of an extension, therefore, we start with a set $\mathcal{E}$, use that set to define applicability, and thus a set of conclusions. $\mathcal{E}$ is an extension if this operation brings us back where we started.

Formally, let $\mathcal{L}$ be a first-order language. If $\mathcal{S}$ is a set of sentences in $\mathcal{L}$, let $\mathrm{Th}(\mathcal{S})$ be the set of first-order consequences of $\mathcal{S}$. We define a default theory to be a pair $< \mathcal{D}, \mathcal{W} >$ where $\mathcal{W}$ is a set of sentences in $\mathcal{L}$ and where $\mathcal{D}$ is a set of inference rules of the form "$\alpha : \mathrm{M}\beta/\gamma$" where α, β, γ are formulas in $\mathcal{L}$. A default rule Φ' is an instance of rule Φ if Φ' is closed, and some substitution for the variables in Φ gives Φ'.

Let $< \mathcal{D}, \mathcal{W} >$ be a default theory and let $\mathcal{E}$ be any set of sentences in $\mathcal{L}$. For any subset $\mathcal{S} \subset \mathcal{E}$, define the function $\mathrm{concs}(\mathcal{S}, \mathcal{D}, \mathcal{E})$ (the conclusions from $\mathcal{S}$ using default rules $\mathcal{D}$ in extension $\mathcal{E}$) as follows:

$$\begin{aligned}\mathrm{concs}(\mathcal{S}, \mathcal{D}, \mathcal{E}) = \{\ \gamma \mid & \text{ there exists a rule } \Phi' = \alpha : \mathrm{M}\beta/\gamma \\ & \text{which is an instance of some rule } \Phi \in \mathcal{D} \\ & \text{such that } \alpha \in \mathcal{S} \text{ and } \neg\beta \notin \mathcal{E}\ \}\end{aligned}$$

Now, define the sequence $\mathcal{E}_0, \mathcal{E}_1, \ldots$ as follows

$$\begin{aligned}&\mathcal{E}_0 = \mathcal{W} \\ &\mathcal{E}_{i+1} = \mathrm{Th}(\mathcal{E}_i) \cup \mathrm{concs}(\mathcal{E}_i, \mathcal{D}, \mathcal{E})\end{aligned}$$

$\mathcal{E}$ is an *extension* of $< \mathcal{D}, \mathcal{W} >$ if

$$\mathcal{E} = \bigcup_{i=0}^{\infty} \mathcal{E}_i$$

Examples:

A. Let $\mathcal{D}$ = { bird(X) : can_fly(X) / can_fly(X) } and
$\mathcal{W}$ = { bird(tweety) }.
Then there is a unique extension $\mathcal{E}$ = Th({ bird(tweety), can_fly(tweety) }).
In this case $\mathcal{E}_0 = \mathcal{W}$; $\mathcal{E}_1$ = Th($\mathcal{W}$) $\cup$ { can_fly(tweety) }; $\mathcal{E}_2 = \mathcal{E}$.

B. Let $\mathcal{D}$ = { bird(X) : can_fly(X) / can_fly(X) } and
$\mathcal{W}$ = { ostrich(ozzie),
$\forall_X$ ostrich(X) $\Rightarrow$ (bird(X) $\wedge$ $\neg$can_fly(X)) }.
This has a single extension $\mathcal{E}$=Th($\mathcal{W}$).

C. Let $\mathcal{D}$ = { bird(X) : can_fly(X) / can_fly(X);
sings(X,bird_song) : bird(X) / bird(X) } and
$\mathcal{W}$ = { sings(tweety,bird_song) }.
This has a single extension $\mathcal{E}$ = Th($\mathcal{W}\cup$ { bird(tweety), can_fly(tweety) }).

D. Let $\mathcal{D}$ = { speak(X,turkish) : citizen(X,turkey) / citizen(X,turkey);
lives_in(X, C) : citizen(X, C) / citizen(X, C) }.
Let $\mathcal{W}$ = { speak(kemal,turkish),
lives_in(kemal,usa),
$\forall_X$ ¬(citizen(X,turkey) ∧ citizen(X,usa)) }.
This has two extensions: $\mathcal{E}_A$ = Th($\mathcal{W}$ ∪ { citizen(kemal,turkey) }) and
$\mathcal{E}_B$ = Th($\mathcal{W}$ ∪ { citizen(kemal,usa) }).

E. Let $\mathcal{D}$ = { :B / ¬B } and let $\mathcal{W} = \emptyset$.
This theory has no extensions.
Proof: Let $\mathcal{E}$ be any set of sentences, and construct the sequence $\mathcal{E}_0, \mathcal{E}_1 \ldots$
according to the above definition. If ¬B $\notin \mathcal{E}$, then the justification of the default is satisfied,
so that ¬B$\in \mathcal{E}_1$, and $\mathcal{E} \neq \cup\mathcal{E}_i$. If ¬B $\in \mathcal{E}$, then the justification of the rule is denied,
and ¬B$\notin \mathcal{E}_i$ for any i, so again $\mathcal{E} \neq \cup\mathcal{E}_i$.

3.1.5 Preferred Models

As we discussed above, the monotonicity of standard logic is inherent in its semantics. A theory $\mathcal{T}$ has a consequence ϕ if ϕ is true in all models satisfying $\mathcal{T}$. This is an inescapably monotonic idea of consequence, because when we increment $\mathcal{T}$ with a new fact ψ, we simply reduce the set of relevant models to a subset; ϕ must still be true in all of them.

How can we get out of this? One approach is to say that $\mathcal{T}$ has consequence ϕ if ϕ is true in *most* of the worlds satisfying $\mathcal{T}$. Once we add a new fact ψ, we may select an atypical subset of these models, and most of the models in this subset may not satisfy ϕ. This approach leads to the possible worlds semantics for probability, which we will discuss in section 3.2.2. Another approach, which we will study here, is to say that ϕ is a consequence of $\mathcal{T}$ if ϕ is true in the *best* models that satisfy $\mathcal{T}$. If $\mathcal{T}$ is augmented by a statement ψ, then the class of best models may change. ϕ may be true in the best models of $\mathcal{T} \cup \{\psi\}$, though false in the best models of $\mathcal{T}$.

Formally, we define a partial ordering $\mathcal{M}_1 < \mathcal{M}_2$ on models, read "$\mathcal{M}_1$ is preferred to $\mathcal{M}_2$." $\mathcal{M}$ is said to be a *minimal model* of theory $\mathcal{T}$ if $\mathcal{M}$ satisfies $\mathcal{T}$, and no other model satisfying $\mathcal{T}$ is preferred to $\mathcal{M}$. If the sentence ψ is true in all minimal models of $\mathcal{T}$, then ψ is called a *non-monotonic consequence* of $\mathcal{T}$ relative to the partial ordering "<".

For example, the default rule "Typical sonatas have three movements," can be describing by saying that model $\mathcal{M}_1$ is preferred to $\mathcal{M}_2$ if every sonata with three movements in $\mathcal{M}_2$ also has three in $\mathcal{M}_1$, but not vice-versa. Now, suppose we do not know how many movements are in Beethoven's Opus 111. Our theory thus allows models in which it has three movements and models in which it has other numbers of movements. The former will be preferred — to be precise, for each model $\mathcal{M}_2$ in which Opus 111 has some other number of movements, there will be another model $\mathcal{M}_1$ that agrees with $\mathcal{M}_2$ on all other sonatas but gives Opus 111 the standard three movements. Thus, the statement "Opus 111 has three movements" is true in all minimal models of our theory; it is a non-monotonic consequence of our theory. If we now find out that Opus 111 has only two movements, then models where it has three movements do not satisfy our new theory; hence, the conclusion is no longer valid.

A number of types of non-monotonic inference can be described in terms of preferences on models. For example, the effect of circumscription in a predicate μ can corresponds to the preference relation $\mathcal{M}_1 < \mathcal{M}_2$ if the extension of μ in $\mathcal{M}_1$ is a proper subset of its extension in $\mathcal{M}_2$, and all other predicate symbols have equal extensions.

3.2 Classical Probability Theory

The oldest and best understood formal theory of plausible reasoning is the classical theory of mathematical probability. Probability theory is a quantitative theory; the probability of a statement given a body of evidence is a real number between 0 and 1.[2] The theory thus provides a straightforward measure of strength of belief. Probability theory consists of two parts. The first part, the theory of probability proper, asserts weak, inviolable constraints relating the probabilities of different statements; for example, the probability of $E \wedge F$ is less than or equal to probability of F. The second part, the theory of statistical inference, gives non-monotonic suggestions for picking the actual values; for example, if there are n possibilities, and there is no reason to suppose one more likely than another, then assign each of them a probability of $1/n$.

There are several different ways to interpret probability theory. All of them use the same formulas; they differ in what they connect probabilities to. In general, the simpler the interpretation of probabilities, the further the theory is from anything that can be applied to plausible reasoning. In the simplest formulation, probability is just a ratio of the measures of two sets [Kolmogorov, 1950] [Russell, 1948]; the probability of "Brown" given "Cow" is just the number of brown cows divided by the number of cows. In this interpretation, the axioms of probability theory are trivial consequences of the axioms of measures on sets, and the question of deriving the initial value of probabilities does not arise. In a more complex formulation [Mises, 1960], probabilities are based on the idea of infinite random sequences of repeatable events. However, plausible reasoning involves reasoning about events that occur only once; the probability that Sam has the measles, or the probability that Clyde is grey. The only obvious sets here are singletons, namely the set of Sam and the set of Clyde, and there is nothing repeatable about the events involved. In AI, therefore, we are obliged to use *subjective* probability theory, in which probabilities are interpreted as the likelihood of particular statements, given a body of evidence. This interpretation is much less concrete and its logical form less clear than the other interpretations; indeed, its legitimacy has been hotly debated back and forth over the last two hundred years (see the bibliography for references). Even in using a subjective interpretation of probabilities, it is useful to have the frequency interpretation to refer to, since any general formula of subjective probability can be justified in terms of frequencies, and frequencies are a more concrete object of thought than judgements of likelihood.

We will view a theory as having two kinds of statements: unqualified statements about the real world, such as "Clyde is an elephant," or "95% of all elephants are grey," and statements about probabilities, such as "The probability that Clyde is grey is greater than the probability that he is white," or "The probability that Wendy is white is greater if Wendy is a seagull than if she is an elephant." These probabilities represent a possible judgement of the likelihood of E by an rational

[2]This does not mean that a program or a theory that uses probability theory must assign floating point numbers as the probability of statements. It just means that probabilities are quantities (real numbers, in the classical theory) which can be described in the languages of quantities discussed in chapter 4.

agent who knows the given unqualified facts. Our definition of a "rational" agent is just one that assigns probabilities in a way conforming to the rules given below.

Formally, we can define a probabilistic language as follows: (This is one method among many.) Let $\mathcal{L}_0$ be an object language: this will be the language in which we write the object level sentences to which we will assign probability. These object level sentences are called *events*.[3] The unqualified part of a theory $\mathcal{T}_0$ is a set of sentences in $\mathcal{L}_0$. We introduce two probability functions: the prior probability of E, relative to $\mathcal{T}_0$, written $\mathrm{P}(E)$, and the conditional probability of E given F, relative to $\mathcal{T}_0$, written $\mathrm{P}(E \mid F)$. These are partial functions: $\mathrm{P}(E \mid F)$ is always undefined if F is known to be false, and it may be undefined if there is no relevant information whatever. (What is the probability that Clyde is grey given the parallel postulate?) E and F, the arguments to P, are events; the function P maps them to a real number in the interval [0,1]. We speak about these probabilities by using them as atomic terms in a first-order language $\mathcal{L}_R$ of real arithmetic. $\mathcal{L}_R$ will typically contain the arithmetic functions plus, minus, times and divide, the order relations $X > Y$, and any other useful arithmetic relations and functions. A *probabilistic sentence*, then, is some sentence in $\mathcal{L}_R$ containing terms of the form $\mathrm{P}(E)$ and $\mathrm{P}(E \mid F)$ as atomic terms. A *probabilistic theory* is a pair $< \mathcal{T}_0, \mathcal{T}_P >$ where $\mathcal{T}_0$ is a set of sentences in $\mathcal{L}_0$, and $\mathcal{T}_P$ is a set of probabilistic sentences.

Thus, in the above example, $\mathcal{L}_0$ would be language rich enough to express sentences such as "Clyde is an elephant," "Wendy is white", "95% of elephants are grey," and so on. The unqualified theory $\mathcal{T}_0$ would contain "Clyde is an elephant" and "95% of elephants are grey". The terms "P(grey(clyde))", the probability that Clyde is grey, and "P(white(wendy) | seagull(wendy))", the conditional probability that Wendy is white given that she is a seagull, are probabilistic terms. The sentences "P(grey(clyde)) = .95" and

$$\mathrm{P}(\text{white(wendy)} \mid \text{seagull(wendy)}) > \mathrm{P}(\text{white(wendy)} \mid \text{elephant(wendy)})$$

are probabilistic sentences, which may be part of the probabilistic component $\mathcal{T}_P$.

The above definition imposes two restrictions that should be noted:

- The definition does not allow quantification into the scope of a probability operator. For example, the definition rules out a sentence such as "$\forall_X$ P(grey(X) | elephant(X)) = 0.95", or "$\exists_X$ P(color(wendy)=X | elephant(wendy)) $\geq$ 0.95," (meaning "There is some particular color that is almost certainly Wendy's color.") Such quantification would require that P be viewed as a modal operator creating an opaque context; we should not want to substitute "the youngest white elephant," for X in the first sentence above, nor to derive the second sentence from the fact "P(color(wendy)=color(wendy) | elephant(wendy)) $\geq$ 0.95" by existential abstraction on "color(wendy)". To simplify our theory, we therefore rule out these sentences. Part of the effect of a universal quantifier can be gotten by using axiom schemas; for example, we might posit the rule "For any constant symbol γ, P(grey(γ) | elephant(γ)) = 0.95."
- The definition does not allow "higher-order" probabilities such as "P(P(grey(clyde)) > .95) = .5" meaning "There is a .5 probability that the probability that Clyde is grey is greater than .95." Such higher-order probabilities can generally be avoided in

[3] This is quite different from the use of the term *events* in temporal reasoning.

simple applications of probability theory to plausible reasoning, though they can be useful in a meta-level evaluation of a system that assigns object-level probabilities. [Gaifman, 1983,1986]

In any probabilistic theory $\mathcal{T} =< \mathcal{T}_0, \mathcal{T}_P >$ the following axioms must hold:

1. (Closure) If $\mathrm{P}(E)$ and $\mathrm{P}(F)$ are defined then $\mathrm{P}(E \wedge F)$, $\mathrm{P}(E \vee F)$, and $\mathrm{P}(\neg E)$ are all defined. If two of the quantities $\mathrm{P}(E)$, $\mathrm{P}(F)$, and $\mathrm{P}(E \mid F)$ are defined, then the third is also defined, unless $\mathrm{P}(F) = 0$.

2. (Probability of Known Facts) If $\mathcal{T}_0 \vdash E$ then $\mathrm{P}(E) = 1$ and $\mathrm{P}(\neg E) = 0$. Facts known without qualification have probability 1, and their negations have probability 0. Note that the converse is not the case; a sentence may be given probability 1 without its being provable from the known facts.

3. (Range) $0 \leq \mathrm{P}(E) \leq 1$. All probabilities are between 0 and 1.

4. (Invariance under equivalence) If $\mathcal{T}_0 \vdash (E \Leftrightarrow F)$ then $\mathrm{P}(E) = \mathrm{P}(F)$. If E and F are known to have the same truth value, then they have equal probabilities.

5. (Conjunction) $\mathrm{P}(E \wedge F) = \mathrm{P}(E) \cdot \mathrm{P}(F \mid E)$. The probability that E and F are both true is the probability that E is true times the probability that F is true given that E is true. This can function as a definition of the conditional probability $\mathrm{P}(E \mid F)$, except in cases where $\mathrm{P}(F) = 0$.

It should be noted that $\mathrm{P}(E \mid F)$ is not the same as $\mathrm{P}(F \Rightarrow E) = \mathrm{P}(\neg F \vee E)$. The latter term is, in fact, rather useless. If $\mathrm{P}(E \mid F)$ is high, then we can legitimately infer E given F. By contrast $\mathrm{P}(F \Rightarrow E)$ may be high just because F is very unlikely; if F is found out to be true, E may still be unlikely.

6. (Disjunction) If $\mathcal{T}_0 \vdash \neg(E \wedge F)$ then $\mathrm{P}(E \vee F) = \mathrm{P}(E) + \mathrm{P}(F)$. If it is known that E and F cannot both be true, then the probability of the disjunction $E \vee F$ is the sum of the probability of E plus the probability of F.

7. (Belief Update) A final meta-axiom does not constrain any particular probabilistic theory; rather, it describes how a rational agent should pass from one theory to another when he establishes some new information. An agent who finds out that event E is true should adopt the previous value of $\mathrm{P}(F \mid E)$ as his new value for the prior $\mathrm{P}(F)$, and should adopt the previous value of $\mathrm{P}(F|E \wedge G)$ as his new value of $\mathrm{P}(F|G)$, assuming that these quantities were previously defined. For instance, suppose you previously calculated that there was a .95 probability Clyde is grey given that Clyde is an elephant, and a .90 probability that Clyde can do tricks given that Clyde is an elephant and that he is a circus animal. You now find out that Clyde is, indeed, an elephant. You should now believe that there is a .95 prior probability that Clyde is grey, and that there is a .90 probability that he can do tricks given that he is a circus animal.

The expected value of a term is defined as the sum of its possible values times the probability of its having that value.

$$E(T) = \sum_{X_i} \mathrm{P}(T = X_i) \cdot X_i$$

A number of important consequences follow directly from these axioms:

8. (Implication) If $\mathcal{T}_0 \vdash E \Rightarrow F$ then $\mathrm{P}(F) \geq \mathrm{P}(E)$

9. (Relativising rules 3, 5, 6, and 8 to conditional probabilities.)

   a. $0 \leq \mathrm{P}(E \mid F) \leq 1$.

   b. $\mathrm{P}(E \wedge F|G) = \mathrm{P}(E \mid G) \cdot \mathrm{P}(F|G \wedge E)$

   c. If $\mathcal{T}_0 \vdash G \Rightarrow (\neg(E \wedge F))$ then $\mathrm{P}(E \vee F \mid G) = \mathrm{P}(E \mid G) + \mathrm{P}(F \mid G)$.

   d. If $\mathcal{T}_0 \vdash G \Rightarrow (E \Rightarrow F)$ then $\mathrm{P}(F \mid G) \geq \mathrm{P}(E \mid G)$

This result is important for establishing the coherence of the update rule (6), since it guarantees that the new probabilities satisfy the axioms of a probability theory.

10. Exhaustive Disjoint Possibilities. Let $E_1 \ldots E_k$ be a set of events of which it is known that exactly one is true. Such a set of events is known as a *frame of discernment*. The sum of the probabilities of such events must be 1. This applies both to prior probabilities and to probabilities conditioned on some fixed condition F. Formally,

If $\mathcal{T}_0 \vdash E_1 \dot{\vee} E_2 \dot{\vee} \ldots \dot{\vee} E_k$ then

$$\sum_{i=1}^{k} \mathrm{P}(E_i) = \sum_{i=1}^{k} \mathrm{P}(E_i \mid F) = 1$$

11. The probability of a disjunction can be calculated as follows:

$$\mathrm{P}(E \vee F) = \mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \wedge F)$$

12. Corresponding to Modus Ponens: $\mathrm{P}(F) \geq \mathrm{P}(E) \cdot \mathrm{P}(F \mid E)$

13. Evidence augmentation: Let $\mathrm{P}(X \mid E) = a$ and let $\mathrm{P}(F \mid E) = b$. Then

$$\frac{a + b - 1}{b} \leq \mathrm{P}(X \mid E \wedge F) \leq \frac{a}{b}$$

If b is close to 1, this allows us to compute useful bounds on the effect of learning F on the plausibility of X.

3.2.1 Bayes' Formula

Axiom 5 allows us to reverse the direction of conditional probabilities. If we know $\mathrm{P}(E)$, $\mathrm{P}(F)$, and $\mathrm{P}(E \mid F)$, then by axiom 5,

$$\mathrm{P}(F \mid E) = \frac{\mathrm{P}(E \wedge F)}{\mathrm{P}(E)} = \frac{\mathrm{P}(E \mid F) \cdot \mathrm{P}(F)}{\mathrm{P}(E)}$$

This is called Bayes' formula.

Bayes' formula is often used in reasoning from an observed effect E to an inferred cause F. If we know the prior probabilities of E and F and we know the likelihood that F will occur if E

has occurred, then Bayes' formula allows us to computer the likelihood that E occurred given that F occurred. Often, the probability of the effect F is not known directly, but is computed from considering all the possible causes. Let $E_1,\ E_2,\ ...E_k$ be a exhaustive set of mutually disjoint causes for F; that is, given F we know that exactly one of the E_i must have occurred. Formally,

$$\mathcal{T}_0 \vdash F \Rightarrow (E_1 \dot{\vee} E_2 \dot{\vee} \ldots \dot{\vee} E_k)$$

Using axiom 6 above

$\mathrm{P}(F) = \mathrm{P}(F \wedge (E_1 \dot{\vee} \ldots \dot{\vee} E_k)) = \mathrm{P}((F \wedge E_1) \dot{\vee} \ldots \dot{\vee} (F \wedge E_k)) =$
$\mathrm{P}(F \wedge E_1) + \ldots + \mathrm{P}(F \wedge E_k) = \mathrm{P}(F \mid E_1)\ \mathrm{P}(E_1) + \ldots + \mathrm{P}(F \mid E_k)\ \mathrm{P}(E_k) =$
$\sum_{j=1}^{k} \mathrm{P}(F \mid E_j)\mathrm{P}(E_j)$

Substituting this last expression for $\mathrm{P}(F)$ in Bayes' formula above yields the formula

$$\mathrm{P}(E_i \mid F) = \frac{\mathrm{P}(F \mid E_i)\mathrm{P}(E_i)}{\sum_{j=1}^{k} \mathrm{P}(F|E_j)\ \mathrm{P}(E_j)}$$

For example, consider the following situation: Edgar invites Karen to a movie Saturday night, and Karen answers that unfortunately she has a prior engagement. Edgar is now interested in the probability that Karen is politely lying. Let A be the event that Karen answers that she is busy; B, the event that Karen is actually busy; F, the event that Karen is actually free. Edgar now estimates the prior probabilities as follows:

$\mathrm{P}(B) = 0.3$ (She's popular, but he called her on Monday, which is plenty of time)

$\mathrm{P}(F) = 1 - \mathrm{P}(B) = 0.7$

$\mathrm{P}(A \mid B) = 0.95$ (There was a 1/20 chance that Karen would be so wild about him as to break a previous date.)

$\mathrm{P}(A \mid F) = 0.25$ (Edgar has a pretty good opinion of his own attractiveness.)

Applying Bayes' formula, we have

$$\mathrm{P}(F \mid A) = \frac{\mathrm{P}(A \mid F) \cdot \mathrm{P}(F)}{\mathrm{P}(A \mid F) \cdot \mathrm{P}(F) + \mathrm{P}(A \mid B) \cdot \mathrm{P}(B)} = \frac{0.25 \cdot 0.7}{0.25 \cdot 0.7 + 0.95 \cdot 0.3} = 0.38$$

Not so large as to discourage another call next week.

3.2.2 Possible Worlds Semantics

There is a natural possible-worlds semantics for probability judgements. A world is described by the events in $\mathcal{L}_0$. We define a measure μ on sets of possible worlds, satisfying the following axioms:

1. For any set X, $\mu(X) \geq 0$.

2. If W and X are disjoint sets of possible worlds, then $\mu(W \cup X) = \mu(W) + \mu(X)$.

We define the probability of an event E as the fraction of worlds satisfying E out of all the worlds satisfying the background theory. Formally, let $\mathcal{U}$ be the set of all worlds that satisfy the background theory, and let $\mathcal{W}_E$ be the subset of worlds that satisfy both the background theory and E. Then

$$P(E) = \frac{\mu(\mathcal{W}_E)}{\mu(\mathcal{U})}$$

The conditional probability $\mathrm{P}(E \mid F)$ is defined as the ratio

$$P(E \mid F) = \frac{\mu(\mathcal{W}_E \cap \mathcal{W}_F)}{\mu(\mathcal{W}_F)}$$

It is easily shown that all the above axioms follow from these axioms and definitions, and that, conversely, for any probability distribution satisfying the axioms, there is a corresponding measure μ (just pick the measure to be equal to the probability.) In this way, subjective probability can be formally reduced to frequencies.

3.3 Statistical Inference

The above formulas express the necessary properties of a probabilistic theory. They are, however, very weak constraints. In particular, for any body of deterministic knowledge, it is possible to assign all prior and conditional probabilities to be either 0 or 1. Moreover, the axioms above provide little calculus of evidence combination, beyond the weak rule of augmentation (13) given above. If we have n logically independent events X, A, B, C, ... then the 2^{n-1} quantities

$\mathrm{P(X)}$, $\mathrm{P(X \mid A)}$, $\mathrm{P(X \mid B)}$, $\mathrm{P(X \mid C)}$, ...
$\mathrm{P(X \mid A \wedge B)}$, $\mathrm{P(X \mid A \wedge C)}$ ... $\mathrm{P(X \mid B \wedge C)}$ ...
$\mathrm{P(X \mid A \wedge B \wedge C)}$ ...
...

can each be assigned an arbitrary value between 0 and 1, and the result will be perfectly consistent with the axioms. To remedy the situation, we have a variety of rules that can used to pick probabilities other than 0 or 1, and to relate conditional probabilities. These are all non-monotonic rules; they are deduced from a body of information and can be overruled if more information becomes available. For example, suppose an object-level theory $\mathcal{T}_0$ supports the probabilistic sentence $\mathrm{P}(E)$=1/2. The extension $\mathcal{T}_0 \cup \{E\}$ supports the sentence $\mathrm{P}(E)$=1, and does not support the sentence $\mathrm{P}(E)$=1/2.

3.3.1 Frequency

The most basic probabilistic heuristic is the principle of relative frequency: if there are N objects with property α and M of these have property β, then for any term t, $\mathrm{P}(\beta(t) \mid \alpha(t)) = M/N$, unless we have some additional information about the term t. For example, given that about 51.5% of all people are women, the probability can be taken to be .515 that the next person I see will be a woman. This inference must be changed if I have more information available. For instance, if I know that I am sitting in a class in which there are 17 people other than myself, of whom 3 are women, then the probability is $3/17 = .176$ that the next person I see will be a woman.

A important special case of the frequency principle is the principle of equidistribution. Let $\beta(X)$ be the property $X = s$ for some term s. Since there is only one element with this property, namely s itself, the probability $\mathrm{P}(t = s \mid \alpha(t)) = 1/N$. Again, the inference is non-monotonic and rests on the assumption that there is no other information relating t and s. For example, if t and s are the same term then the probability is 1. If t and s are different constants and the unique names assumption is operative, then the probability is 0.

A much stronger form of the principle of equidistribution is the *principle of indifference:* If the n events $E_1 \ldots E_n$ form a frame of discernment, then, in the absence of other information, each should be assigned probability $1/n$. For example, if it is known that a block is either black, red, blue, or white, then, absent other information, assign probability 1/4 to each possibility. The obvious problem with this principle is that, in general, a given event may belong to many different frames of discernment, of different sizes, and the principle does not indicate which one should be chosen. For example, suppose that we carefully examine the contents of a supermarket. We are then told, "Reginald bought one item at the market," and we are asked, "What is the probability that he bought a potato?" What is the proper frame of discernment here? Should we assign equal probability to each individual item in the store? or to each category? or to each foot of shelf space? If individual items form the frame of discernment, does a six-pack of beer count as one or six? If categories are the frame of discernment, are Maine and Idaho potatoes two categories or one? It would seem that there is no way to choose without more information; but once more information is available, the problem is changed.

3.3.2 Independence

A key observation in probability theory is that most pairs of events have *nothing whatever* to do with one another, and that knowing one has no effect on the probability of the other. Knowing that one coin toss came up heads does not change the probabilities that another will come up heads; knowing the price of tea in China does not affect the probability that the next President will be a woman. Two such events are said to be *independent.*

Formally, events E and F are independent if $\mathrm{P}(E \mid F) = \mathrm{P}(E)$. By axiom 5, this is equivalent to the symmetric constraint,

$$\mathrm{P}(E \wedge F) = \mathrm{P}(E) \cdot \mathrm{P}(F).$$

By formula 11, above, this implies the equation,

$$\mathrm{P}(E \vee F) = \mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E) \cdot \mathrm{P}(F).$$

Thus, given a collection of independent events $E_1, E_2, \ldots, E_k$, and their prior probabilities $\mathrm{P}(E_1) \ldots \mathrm{P}(E_k)$, we can compute uniquely the probability of any boolean combination of the E_i. Note that we must know, not only that the E_i are pairwise independent, but that every subset of the E_i is collectively independent; that is, for any set S of the E_i,

$$\mathrm{P}(S) = \prod_{E \in S} \mathrm{P}(E)$$

This condition of collective independence does not follow from the pairwise independence constraints

$$\mathrm{P}(E_i \wedge E_j) = \mathrm{P}(E_i) \cdot \mathrm{P}(E_j)$$

Thus, the independence assumption is a very powerful tool when it can be applied. Moreover, it is a default assumption; two events may be assumed to be independent unless there is some reason to believe that they are connected. (We can also justify this assumption on the basis of the Maximum Entropy principle; see section 3.3.4 below.) This assumption is enormously valuable in ignoring the masses of irrelevant information always present; if we want to compute the probability that a patient has a heart condition, we are justified in ignoring his name, zodiacal sign, and the color of his eyes as independent of the event of interest. It is less frequently applicable in computing the probability of boolean combinations, as described above. The problem is that, if the probability of the event $E \wedge F$ is worth computing, it is generally because there is some connection between E and F, and where there is a connection there is often some degree of dependence.

Suppose, for example, that you are expecting twins, and you wish to know the probability that both children will be boys. As a first guess, one might use the principle of indifference to estimate that the probability that a random child will be a boy is 1/2, and then, assuming independence, square this to get 1/4. Of course, this is wrong; fraternal twins are independent but identical twins are not. To get an accurate probability, it is necessary to know the probability that a pair of twins will be identical. If no information on the frequency of identical twins is available, one might use the principle of indifference to guess that the probability is one-half, which would lead to a probability of 3/8 that both children are boys.

In practice, establishing independence is not simply a default inference; it can be a very difficult enterprise. For example, much of the ongoing debate about the interpretation of the differences in standardized test scores across different segments of the population rests on such questions as whether, for fixed intelligence, test scores are independent of social background; or whether, for fixed conditions of social pressures and expectations, test scores are independent of sex.

3.3.3 Independent Evidence

The independence assumption can also be used in evidence combination. Suppose that we have two events E and F, both of which are evidence for another event X. If we can assume that E and F are independent given X, and that E and F are independent given $\neg X$ then it is easy to compute $\mathrm{P}(X \mid E \wedge F)$ from $\mathrm{P}(X \mid E)$ and $\mathrm{P}(X \mid F)$ as follows:

Let E and F be independent, both with respect to X and with respect to $\neg X$; thus $\mathrm{P}(E \wedge F \mid X) = \mathrm{P}(E \mid X)\ \mathrm{P}(F \mid X)$ and $\mathrm{P}(E \wedge F \mid \neg X) = \mathrm{P}(E \mid \neg X)\ \mathrm{P}(F \mid \neg X)$. By Bayes' formula

1.

$$\mathrm{P}(X \mid E \wedge F) = \frac{\mathrm{P}(E \wedge F \mid X)\mathrm{P}(X)}{\mathrm{P}(E \wedge F)}$$

Similarly,

2.

$$\mathrm{P}(\neg X \mid E \wedge F) = \frac{\mathrm{P}(E \wedge F \mid \neg X)\mathrm{P}(\neg X)}{\mathrm{P}(E \wedge F)}$$

Dividing formula (1) by (2) yields

3.

$$\frac{\mathrm{P}(X \mid E \wedge F)}{\mathrm{P}(\neg X \mid E \wedge F)} = \frac{\mathrm{P}(E \wedge F \mid X)\mathrm{P}(X)}{\mathrm{P}(E \wedge F \mid \neg X)\mathrm{P}(\neg X)}$$

Using our independence assumptions, we can rewrite this.

4.

$$\frac{\mathrm{P}(X \mid E \wedge F)}{\mathrm{P}(\neg X \mid E \wedge F)} = \frac{\mathrm{P}(E \mid X)\mathrm{P}(F \mid X)\mathrm{P}(X)}{\mathrm{P}(E \mid \neg X)\mathrm{P}(F \mid \neg X)\mathrm{P}(\neg X)}$$

Let us define the odds on A as the ratio $\mathrm{P}(A)/\mathrm{P}(\neg A)$. Thus, if $\mathrm{P}(A) = 1/4$, then $\mathrm{P}(\neg A) = 1 - \mathrm{P}(A) = 3/4$ so $\mathrm{Odds}(A) = (1/4)/(3/4) = 1/3$. Analogously, we define $\mathrm{Odds}(A \mid B)$ as the ratio $\mathrm{P}(A \mid B)/\mathrm{P}(\neg A \mid B)$. Using Bayes' rule

5.

$$\mathrm{Odds}(A \mid B) = \frac{\mathrm{P}(A \mid B)}{\mathrm{P}(\neg A \mid B)} = \frac{\mathrm{P}(B \mid A)\mathrm{P}(A)/\mathrm{P}(B)}{\mathrm{P}(B \mid \neg A)\mathrm{P}(\neg A)/\mathrm{P}(B)} = \frac{\mathrm{P}(B \mid A)}{\mathrm{P}(B|\neg A)} \cdot \mathrm{Odds}(A)$$

so 6.

$$\frac{\mathrm{P}(B \mid A)}{\mathrm{P}(B \mid \neg A)} = \frac{\mathrm{Odds}(A \mid B)}{\mathrm{Odds}(A)}$$

Let us now define the odds updating function $\mathrm{OU}(A \mid B)$ as the ratio $\mathrm{Odds}(A \mid B)/\mathrm{Odds}(A)$, the change that evidence B makes in the odds of A. From (6), we have

7.

$$\frac{\mathrm{P}(B \mid A)}{\mathrm{P}(B \mid \neg A)} = \mathrm{OU}(A \mid B)$$

We can therefore rewrite formula (4) as

8.

$$\mathrm{OU}(X \mid E \wedge F) = \mathrm{OU}(X \mid E) \cdot \mathrm{OU}(X \mid F)$$

For example, suppose that $\mathrm{P}(X) = 1/3$, $\mathrm{P}(X \mid E) = 2/3$, and $\mathrm{P}(X \mid F) = 3/4$. Then $\mathrm{Odds}(X) = 1/2$, $\mathrm{Odds}(X \mid E) = 2$, $\mathrm{Odds}(X \mid F) = 3$; so $\mathrm{OU}(X \mid E) = 4$ and $\mathrm{OU}(X \mid F) = 6$. Therefore $\mathrm{OU}(X \mid E \wedge F) = 24$, so $\mathrm{Odds}(X \mid E \wedge F) = \mathrm{OU}(X \mid E \wedge F) \cdot \mathrm{Odds}(X) = 12$, and

$$\mathrm{P}(X \mid E \wedge F) = \frac{\mathrm{Odds}(X \mid E \wedge F)}{1 + \mathrm{Odds}(X \mid E \wedge F)} = 12/13$$

(The assumption we have made above that E and F are independent with respect to both X and $\neg X$ is sometimes replaced by the assumption that E and F are independent with respect to both X and the background theory (e.g. [Charniak and McDermott, 1985].) This leads to the following combination formula:

$$\frac{\mathrm{P}(X \mid E \wedge F)}{\mathrm{P}(X)} = \frac{\mathrm{P}(X \mid E)}{\mathrm{P}(X)} \cdot \frac{\mathrm{P}(X \mid F)}{\mathrm{P}(X)}$$

There are a number of reasons to prefer the assumption that we have made, leading to formula 8 above. First, intuitively, if E and F are both positive evidence for X, then it seems likely that they are connected, and hence unlikely that they are independent relative to the background. Second, the first formula always gives meaningful results; by contrast, the second can yield values for $\mathrm{P}(X \mid E \wedge F)$ that are greater than 1. Third, the first formula is consistent with the maximum entropy assumption (Exercise 12), while the second is not.)

3.3.4 Maximum Entropy

(Note: This section is somewhat difficult mathematically. It may be omitted without loss of continuity.)

Both the rules above, equidistribution and independence, can be derived as special cases of a very general principle called the principle of maximum entropy. There are a number of ways to justify this principle; our approach is relatively simple, though informal.[4] We divide the principle of maximum entropy into two parts:

(I) Let $S = \{E_1, E_2, \ldots, E_k\}$ be a set of events forming a frame of discernment. Suppose we know some partial constraints on their probabilities p_i. The values that we should assign to the probabilities p_i are those that give us the least information about which of the events occurs, subject to our constraints.

[4] [Shore and Johnson, 80] gives a derivation of the principle of maximum entropy from a set of invariance principles.

(II) One's ignorance about which event will occur is measured by the *entropy* function $H(S) = -\sum_{i=1}^{k} p_i \log(p_i)$ Therefore, principle (I) is carried out by picking the p_i so as to maximize the entropy function.

Principle (I) would seem to be inherently plausible. Since it is an informal statement of a desideratum, there is no way of arguing for it formally, though it is evidence for its reasonableness that it is consistent with equidistribution and independence. Principle (II), on the other hand, though formally a definition, is not at all intuitive; its justification will require a brief excursion into information theory. We will develop the formula above in three steps:

(1) Consider a process that randomly generates bits. Since there are 2^k different strings of k bits, the probability of any particular set of k bits having a particular value is $1/2^k$. For example, the probability that the first 6 bits will be 101001 is $1/2^6 = 1/64$. Thus, k bits of information have probability $1/2^k$. Inverting this, we may say that an event of probability p has information $\log(1/p) = -\log(p)$.

(2) Let $S = \{E_1 \ldots E_k\}$ be events with probabilities $p_1 \ldots p_k$ forming a frame of discernment. Consider the measure of the information which is gained by finding out which event occurs. The expected value of this information is

$$H(S) = -\sum p_i \log(p_i)$$

This quantity is called the entropy of the probability distribution.

A fundamental theorem of information theory [Shannon and Weaver, 1949] states that if you devise any unambiguous binary code for the events $E_1 \ldots E_k$, then the expected number of bits necessary to specify which event occurred is greater than or equal to $H(S)$; and that there exists a code where the expected number of bits is between $H(S)$ and $H(S) + 1$.

(3) Now, suppose that we have a frame of discernment $S = \{E_1 \ldots E_k\}$ and two candidate probability distributions $p_1 \ldots p_k$ and $q_1 \ldots q_k$. Once we find out which event occurred, then we clearly have the same information, under either distribution. Finding out which event occurred will, on average, provide information $H_p(S) = -\sum p_i \log(p_i)$ if we assign the first distribution, and will provide information $H_q(S) = -\sum q_i \log(q_i)$ if we assign the second. Suppose that $H_p(S) > H_q(S)$. Since we end up with the same information once we know the event, and we gained more information under p_i than under q_i, we must have started with more information under q_i. That is, the probabilities q_i give us more information for predicting the event before knowing the outcome than the probabilities p_i do. Therefore, if we want to assume we have the least possible information about the event before it occurs, distribution p_i will be better than q_i. To minimize the prior information, therefore, we should choose the values of p_i that maximize the sum $H_p(S) = -\sum p_i \log(p_i)$.

The principle of maximum entropy, like the principle of indifference, requires an appropriate choice of frame of discernment for its correct application.

We will now illustrate the power of the principle by using it to derive the principles of equidistribution and of independence. We leave it as an exercise (12) to show that the maximum entropy assumption justifies combining evidence using the independence assumptions described above.

Equidistribution. Let $E_1 \ldots E_k$ be a frame of discernment, with probabilities $p_1 \ldots p_k$ about which nothing whatever is known. The only constraint, therefore, is the universally applying constraint

$\sum p_i = 1$. We want to maximize the entropy function $H(S) = -\sum p_i \log(p_i)$ subject to this constraint. To simplify the calculations, we will use natural logs for the entropy function rather than base-2 logs; the only effect is to change the magnitude of the entropy (a quantity of no great interest) by a uniform factor of $\ln 2$. We find our maximum using the method of Lagrangian multipliers: To maximize $H_p(S)$ subject to the constraint $C(p) = \sum p_i - 1 = 0$, we construct the formula

$$f(p) = H_p(S) + \alpha C(p) = -\sum p_i \ln(p_i) + \alpha(\sum p_i - 1)$$

and solve the simultaneous equations

$$\frac{\partial f}{\partial p_i} = 0$$

$$C(p) = 0.$$

Performing the differentiations, we get

$$-1 - \ln(p_1) + \alpha = 0$$

$$-1 - \ln(p_2) + \alpha = 0$$

$$\ldots$$

$$-1 - \ln(p_k) + \alpha = 0$$

$$\sum p_i = 1$$

The first k equation imply that all the p_i are equal to $e^{\alpha-1}$, and therefore all equal to one another: $p_1 = p_2 = \ldots = p_k$. Applying the last constraint gives us $p_1 = p_2 = \ldots = p_k = 1/k$, which is the principle of equidistribution.

Independence: The principle of maximum entropy can be used to show that, given the probability of a collection of events, and lacking any information connecting them, the minimum information assumption is to assume that they are all independent. We will illustrate with two events; the analysis for k events is analogous, but messier. Suppose that $\mathrm{P}(A) = a$ and that $\mathrm{P}(B) = b$. We wish to show that $\mathrm{P}(A \wedge B) = \mathrm{P}(A) \cdot \mathrm{P}(B) = ab$.

We construct the frame of discernment $E_1 = A \wedge B$; $E_2 = A \wedge \neg B$; $E_3 = \neg A \wedge B$; $E_4 = \neg A \wedge \neg B$. Let these have probabilities $p_1 \ldots p_4$. Our constraints are as follows:

Frame of discernment:	$p_1 + p_2 + p_3 + p_4 = 1$
$\mathrm{P}(A) = a$:	$p_1 + p_2 = a$
$\mathrm{P}(B) = b$:	$p_1 + p_3 = b$

We therefore construct the function

$$\begin{aligned} f(p) &= -(p_1 \ln(p_1) + p_2 \ln(p_2) + p_3 \ln(p_3) + p_4 \ln(p_4) \\ &\quad +\alpha(p_1 + p_2 + p_3 + p_4 - 1) + \beta(p_1 + p_2 - a) + \gamma(p_1 + p_3 - b)) \end{aligned}$$

The equations to be solved are

1.
$$0 = \frac{\partial f}{\partial p_1} = -1 - \ln(p_1) + \alpha + \beta + \gamma$$

2.
$$0 = \frac{\partial f}{\partial p_2} = -1 - \ln(p_2) + \alpha + \beta$$

3.
$$0 = \frac{\partial f}{\partial p_3} = -1 - \ln(p_3) + \alpha + \gamma$$

4.
$$0 = \frac{\partial f}{\partial p_4} = -1 - \ln(p_4) + \alpha$$

5.
$$p_1 + p_2 + p_3 + p_4 = 1$$

6.
$$p_1 + p_2 = a$$

7.
$$p_1 + p_3 = b$$

Adding equation 2 to 3 and 1 to 4 gives us

$$0 = -2 - \ln(p_2) - \ln(p_3) + 2\alpha + \beta + \gamma = -2 - \ln(p_1) - \ln(p_4) + 2\alpha + \beta + \gamma$$

Cancelling common terms, we get, $\ln(p_2) + \ln(p_3) = \ln(p_1) + \ln(p_4)$, or equivalently

8.
$$p_2 p_3 = p_1 p_4$$

From (7), $p_3 = b - p_1$; from (6), $p_2 = a - p_1$; Combining these with (5) gives $p_4 = 1 + p_1 - a - b$. Substituting these in (8) gives

9.
$$(a - p_1)(b - p_1) = p_1(1 + p_1 - a - b)$$

or, simplifying,

10.

$$p_1 = ab$$

which was the desired result.

3.3.5 Sampling

Frequently, it is not feasible or worthwhile to determine the actual incidence of a property in a population. In such cases, it is often possible to estimate the frequency by sampling. A sample of set S is just a subset of S. A sample T of S is *representative* of S with respect to a property α to within a given tolerance if the fraction of elements of T with property α is within tolerance of the fraction of the elements of S with property α. Thus, if you can determine the frequency of α in a representative sample T, you can use it as an estimate of the frequency of α in S as a whole.

Of course, this statement is a useless tautology unless you have some way to find a representative sample without knowing in advance the frequency of α in S. This can be done in various ways. One way is to know that T is representative of S for properties similar to α. For example, if you know that the overall popular vote for President has always been within 2 percentage points of the vote in McMurdo County, then you can treat the county as a representative sample of the country for the next election. Another technique is to divide the population into subgroups of known size along lines believed to be relevant to α, to sample these subgroups, and to weight their significance in proportion to the size of the subgroups. For example, political poll-takers are careful to pick a proportional mix of sexes, ages, and income-levels in their samples.

The most basic method of getting a representative sample is to take a random sample. Almost all samples of a substantial size are representative to within a small tolerance; therefore, if all different samples have equal probability of being chosen, there is a very large probability of getting a representative sample. In particular, Tschebyscheff's inequality states the following. Let p be the frequency of α in S, let T be a random sample of S of size n, and let q be the frequency of α in T. Then, for any $\epsilon > 0$

$$\mathrm{P}(\mid p - q \mid > \epsilon) < \frac{p(1-p)}{n\epsilon^2}$$

Thus, by making n large enough, one can make arbitrarily certain that q lies arbitrarily close to p.

We need to reverse this inference; to find the probabilities of frequencies in the overall population given the frequency in the sample. We can do this using Bayes' rule, if we have some prior distribution on the probabilities of frequency in the population. This prior is generally taken to be a uniform distribution, in which all frequencies are equally likely.

The following can be proven: Let T be a random sample of size n, out of a population S whose size is very much greater than n. Assume a prior probability distribution in which all frequencies of property α in S are equally likely. Then, given the further fact that k elements of T have property α, the following statments hold:

1. The most probable frequency of α in S is k/n.

2. The expected frequency of α in S is $(k+1)/(n+2)$

3. The probability that a randomly chosen element X in S is α is $(k+1)/(n+2)$.

The above theorem rests on the assumption that all samples are equally probable, or at least that the event of an element being in your sample is independent of the event of its having property α. This assumption is an instance of the principle of equidistribution or of independence, and it is therefore the default in the absence of other information. In practice, however, it requires considerable work to attain this condition. If you have access to all the elements of S, and can use a truely random procedure, such as rolling dice, to pick the ones to include in your sample, then you are on safe ground. In most commonsense cases, however, you do not have access to all elements of S; indeed, often you are obliged to be purely passive, and use as your sample those elements of S that happen within your grasp. Whether such a sampling is independent of the property under study depends both on the set S being sampled and the property being sampled for. The students in your classes may be an adequate sample for determining the incidence of Communism among students at the University, or the incidence of left-handedness among the population of the world; they are unlikely to be a valid sample for determining the frequencies of various majors among university students, or the incidence of Communism in the world population, or the number of legs among members of the animal kingdom. A large body of statistical theory addresses the question of whether samples are being drawn randomly relative to a given property.

3.3.6 Domain Specific Knowledge

In many problems, domain specific knowledge will provide reasonable prior probability distributions, or constraints on such distributions. Examples:

1. Let X be a point in space, let $V1$ be a region in space, and let $V2$ be a subset of $V1$. Then the probability that X is in $V2$ given that X is in $V1$ may be taken to be volume($V2$)/volume($V1$).

2. If you establish some desired condition, and then leave it alone, then the probability that it will still hold after time T is a decreasing function of T. A good prior distribution for this probability would be $e^{-T/\lambda}$, where λ is a characteristic time constant.

3. Consider a set of objects which are designed to be as nearly as possible a given length L — for example, nails that are supposed to be one-half inch long. The probability that the actual length of a randomly chosen object is less than $L+\Delta$ is typically given by the integral

$$\int_{-\infty}^{\Delta} e^{-x^2/\epsilon^2} dx$$

where ϵ is a constant length characteristic of the manufacturing process.

4. When someone tells you something that you had no particular reason to believe true or false, your evaluation of the probability of its truth depends largely on your estimation of the reliability of the source. This estimation combines a number of factors including:

- Has he been correct in the past?

- Is he likely to know the truth of the matter?
- Does he have any reason to wish to deceive you?
- Are there reasons that he would be particularly careful in making this statement?

3.3.7 Conclusion

The axioms of probability theory are a solidly justified set of constraints for quantitative evaluations of likelihood. They are, however, rather weak constraints. To get useful results, it is necessary to have additional values or constraints on the prior or conditional probabilities of events. We have seen a number of heuristics for computing such additional information: equidistribution, frequency, independence, maximum entropy, sampling, and domain dependent knowledge. These heuristics are more delicate and difficult than the axioms; they are non-monotonic inferences, and the relation on world knowledge is not wholly understood. Nonetheless, when they can be correctly applied, they give us powerful and well-founded techniques for evaluating likelihoods, combining evidence, and calculating expected values. Probability theory has been successfully applied in many different aspects of AI, including expert systems, robotics, planning, and natural language understanding.

Nonetheless, the problem of finding sufficient constraints or values for probabilistic computations is still difficult and poorly understood for many types of commonsense reasoning. Let us return, for example, to our example in section 3.2.1 of Edgar inviting Karen to a movie. Suppose that Edgar has asked Karen out three weeks in a row and been turned down all three times. Edgar is getting discouraged, and wants to compute the probability that, if he asks again this week, he will again be rejected. Let B_i be the event that Karen is actually busy when Edgar asks here on the ith week; F_i, the event that she is actually free; and A_i, the event that she rejects the invitation. Then we are interested in the probability of A_4 given A_1, A_2, and A_3.

If Karen were actually a random process, the estimates of the probabilities were firmly established, and the different invitations were independent trials, then there would be no reason to be discouraged, any more than there is reason to bet on white after a run of red on the roulette wheel. The probability of being rejected this week would be

$$\mathrm{P}(A_4 \mid A_1 \wedge A_2 \wedge A_3) = \mathrm{P}(A_4) = \mathrm{P}(A_4 \mid B_4) \cdot \mathrm{P}(B_4) + \mathrm{P}(A_4 \mid F_4) \cdot \mathrm{P}(F_4) = 0.95 \cdot 0.3 + 0.25 \cdot 0.7 = 0.46$$

But these assumptions do not hold. First, there is good reason to think that the events $A_1 \ldots A_4$ are not independent, relative to fixed B_i and F_i. They are likely to be positively correlated for any of a number of reasons: Karen may be involved with someone else, or she may have some habitual Saturday night activity, or she may dislike Edgar. Second, there is reason to suspect that the action of asking may itself influence the outcome of later trials. Karen may get tired of being bugged; she may deliberately arrange to be busy Saturday night so that she won't have to lie. Contrariwise, she may be impressed by Edgar's persistence, and she may deliberately arrange to leave her Saturday night free, so that she can accept his next invitation. Third, however the original prior probabilities were derived, the experiences with Karen suggest that they may be due for revision. The original priors were based on some kinds of estimates (based on sampling, or heuristics, or whatever) about

how often, in general, young women have booked their Saturdays by Monday; how often they will cancel a previous engagement for a date with Edgar; and how often they will invent a previous engagement in order to avoid a date with Edgar. Perhaps these estimates were wrong to start with.

Edgar must now consider all of these circumstances in trying now to evaluate his prospects for success if he calls again. Where is he to get any kind of estimate of the correlations and dependencies involved here?

Moreover, even supposing that we have found rules for assigning subjective probabilities to events, how are we to determine whether these rules are reasonable? Since each event is a unique occurrence and involves a different prior knowledge state, it is not possible to ask whether the prediction corresponds to the actual frequency. (See, however, [Gaifman, 1983].) We could try matching the assigned probabilities to likelihoods as estimated by humans. However, many studies have shown that human subjects judge likelihoods in bizarre ways, which violate the most fundamental laws of probability theory. For example, Daniel Kahneman and Amos Tversky (1982) carried out an experiment in which undergraduate subjects were presented with the following personality sketch: "Linda is 31 years old, single, outspoken, and very bright. As a student she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations." The subjects were then asked which of the following statements was more likely: (A) Linda is a bank teller; or (B) Linda is a bank teller who is active in the feminist movement. 86% of the subjects judged the second statement more likely, despite the fact that this violates the basic rule that a more general state (being a bank teller) is never less likely than a more specific state (being a feminist bank teller). (Perhaps more surprisingly, 43% of the psychology graduate students who were given the same question made the same mistake.) It is hard to conceive of a coherent theory of likelihood which would justify this answer.[5] Nonetheless, people continuously carry out plausible reasoning, and they do it well enough to work their way through a very uncertain world. They must be doing something right, but these experiments suggest that it will not be easy to explain what they are doing and why it is adequate.

3.4 References

General: [Turner, 1984] is a good introduction to many theories of plausible reasoning.

Non-monotonic logic: A good starting point is chapter 6 of [Genesereth and Nilsson, 1987], which gives a more extensive treatment of the issues covered in this chapter. Many of the important early papers on non-monotonic logic were published together in an special issue of *Artificial Intelligence* [Winograd, 1980]. Most of these are reprinted, together with other important papers, in [Ginsberg, 1987]. The closed world assumption was first analyzed in [Reiter, 1978]. Domain closure

[5]There are similar results that indicate that people often perform deductive inference poorly [Johnson-Laird and Wason 1970]. In my opinion, however, the importance of these results to the theory of deductive reasoning is less than the importance of Kahneman and Tversky's result to the theory of plausible reasoning for a number of reasons:
(i) We have a complete formal characterization of valid deduction. We do not have a complete theory of the determination of prior and conditional probabilities. It has been often suggested that these ultimately rest on subjective evaluations; these results show that this will be difficult.
(ii) The negative results on deductive reasoning appear only in rather artificial settings, while the results on likelihood evaluation appear with quite natural questions, such as the experiment cited above.
(iii) The experiments with likelihood evaluation have the bizarre feature that the wrong answer still seems to have something right about it even after the correct answer has been explained and understood.

was presented in [Reiter,19 80b]. Circumscription was presented in terms of an axiom schema in [McCarthy, 1980]; it was re-analyzed in terms of a second order axiom in [Lifschitz, 1985]. [McCarthy, 86] is a general discussion of the scope and limits of circumscription. Reiter's default logic was introduced in [Reiter, 1980a]. [Etherington and Reiter,19 83] studies the connection between default logic and hierarchical inheritance. The minimal model theory was presented in [Shoham, 1987]. Other theories of non-monotonic logic include the NML of [McDermott and Doyle, 1980] and [McDermott, 1982] and the autoepistemic logic of [Moore, 1985b]. Also closely related are the techniques for carrying out nonmonotonic inference in actual programs, such as the "negation as failure" rule of logic programming (analyzed formally in [Lifschitz, 1987b]) and various types of data-dependency maintenance [Doyle, 1979].

Probability theory: The deepest and most extensive study of the applicability and implementation of probabilistic reasoning in AI systems is [Pearl, 1988a]. [Kanal and Lemmer, 1986] and [Lemmer and Kanal, 1988] are collections of papers on the application of probability and other theories of plausible reasoning to AI. The applicability of probability to AI is a matter of heated debate; see, for example, [Cheeseman, 1985], [Charniak, 1983]. (This debate is closedly connected with a larger debate about the correct interpretation of probability theory, and the legitimacy of subjective probability [Fine, 1973].) The particular probabilistic logic described here is close to that of [Grosof, 1988]. The possible-worlds semantics for probability is discussed in [Nilsson, 1986]. For the principle of maximum entropy, see [Jaynes, 1979]. [Grosof, 1988] is a pleasing study of the use of non-monotonic inference in probabilistic reasoning, particular in deriving independence conditions. [Dawes, 1988] discusses how probability theory ought to be used by humans in commonsensical situations, and how far humans are from actually using it.

Schemas: The *locus classicus* for reasoning using schemas is [Minsky, 1975]. See also [Kuipers, 1975], [Hayes, 1979b], [Schank, 1982], [Kolodner, 1984], and [Charniak, 1988].

Other: Other theories of plausible reasoning include Dempster-Shafer theory [Shafer, 1976], [Hummel and Landy, 1986]; the theory of endorsements [Cohen, 1985]; and fuzzy logic [Zadeh, 1987], which deals with vagueness as well as uncertainty. Good overviews of these theories may be found in [Turner, 1984] and in [Bonissone, 1987].

3.5 Exercises

(Starred problems are more difficult.)

1. Let $\mathcal{T}$ be contain the following axioms:

i. father(joe,sam).

ii. father(sam,agnes).

iii. $\text{sibling}(X, Y) \Leftrightarrow \exists_Z \text{ father}(Z, X) \wedge \text{father}(Z, Y)$.

iv. $\text{only\_child}(Z) \Leftrightarrow (\text{sibling}(X, Z) \Rightarrow X = Z)$.

v. $\neg\text{father}(X, X)$.

vi. $\text{father}(X, Y) \wedge \text{father}(Z, Y) \Rightarrow X = Z$.

Show how the conclusion "only_child(sam)" can be inferred from $\mathcal{T}$ by applying domain closure.

2. Let $\mathcal{T}_2$ contain only axioms (i) - (iv) of $\mathcal{T}$. Show how we can infer "only_child(sam)" from $\mathcal{T}_2$ by applying the closed world assumption to the predicate "father" together with the domain closure assumption.

3. Let $\mathcal{T}_3$ contain the axioms in $\mathcal{T}$ together with the seventh axiom

vii. $\forall_X \exists_Y \text{ father}(Y, X)$

What erroneous conclusion can be drawn if the domain closure assumption is applied to $\mathcal{T}_3$?

4. Explain the difference between applying the closed world assumption to a predicate α, and circumscribing in α. Give an example where the two operations give different results.

5*. An important type of plausible inference is default inheritance of properties. A standard example is "If X is a bird, assume that X can fly.". Let $\mathcal{T}$ contain the following axioms:

a. $\text{creature}(X) \wedge \neg\text{ab1}(X) \Rightarrow \neg\text{can\_fly}(X)$.
(Typically creatures do not fly. The predicate $\text{ab1}(X)$ is an abnormality predicate. Peculiar creatures can fly.)

b. $\text{bird}(X) \wedge \neg\text{ab2}(X) \Rightarrow \text{can\_fly}(X)$. (Typically, birds can_fly.)

c. $\text{penguin}(X) \Rightarrow \neg\text{can\_fly}(X)$. (Penguins cannot fly.)

d. $\text{penguin}(X) \Rightarrow \text{bird}(X) \wedge \text{ab2}(X)$.

e. $\text{bird}(X) \Rightarrow \text{creature}(X) \wedge \text{ab1}(X)$.

f. creature(fred).

g. bird(wilma).

h. penguin(barney).

i. distinct(fred,wilma,barney).

Show that we can infer that Wilma can fly but Fred and Barney cannot, by circumscribing over ab1 and ab2 in parallel, letting can_fly vary.

6*. Show that formula (5) in section 3.2.2 is a consequence of formula (7), for all formulas α.

7*. Show that 2.6 in section 3.2.2 is equivalent to formula 2.5.

8. Show that formulas 3.1 and 3.4 in section 3.2.2 together imply the formula

$$\forall_X \text{egg}(X) \Rightarrow \text{fresh}(X)$$

9. Consider the deduction theorem "If $\phi \vdash \psi$ then $\vdash \phi \Rightarrow \psi$."

a. Prove that this theorem holds in a theory with circumscription.

b. Prove that it holds in a minimal models theory.

c. Prove that it does not hold in Reiter's default logic.

10.* Prove formulas 8, 9b, 10, 11, 13 in section 3.3 from axioms 1-5.

11. In first-order logic, one can make inferences from disjunction using the rule of *Splitting:* if $P \vdash R$ and $Q \vdash R$ then $P \vee Q \vdash R$.

a. Show that if the default rules "Birds can typically fly," and "Bats can typically fly," are written in the forms "bird(X) : M(can_fly(X)) / can_fly(X)" and "bat(X) : M(can_fly(X)) / can_fly(X)", then splitting fails. The conclusion "can_fly(tweety)" cannot be inferred from "bat(tweety) $\vee$ bird(tweety)", though it can be inferred from "bat(tweety)" and from "bird(tweety)".

b. Show that the problem in (a) goes away if these default rules are written using the axioms "bird(X) $\wedge$ $\neg$ab(X) $\Rightarrow$ can_fly(X)" and "bat(X) $\wedge$ $\neg$ab(X) $\Rightarrow$ can_fly(X)" and the default rule ": M($\neg$(ab(X)) / $\neg$ab(X)".

c.* Given that P(X | E)=a and P(X | F)=b, show that P(X | E $\vee$ F) $\geq$ ab / (a+b$-$ab).

12.* Show that the maximum entropy assumption justifies the independence assumptions used for evidence combination in section 3.3.3. Specifically, show that given values for P(X), P($X \mid E$), and P($X \mid F$), the maximum entropy solution satisfies the equations P($E \wedge F \mid X$) = P($E \mid X$) P($F \mid X$) and P($E \wedge F \mid \neg X$) = P($E \mid \neg X$) P($F \mid \neg X$).

You may use the following outline for the proof:

a. Use the frame of discernment

$$\begin{array}{ll} E_1 = X \wedge A \wedge B. & E_2 = X \wedge A \wedge \neg B. \\ E_3 = X \wedge \neg A \wedge B. & E_4 = X \wedge \neg A \wedge \neg B. \\ E_5 = \neg X \wedge A \wedge B. & E_6 = \neg X \wedge A \wedge \neg B. \\ E_7 = X \wedge \neg A \wedge B. & E_8 = \neg X \wedge \neg A \wedge \neg B. \end{array}$$

b. Show that the desired independence conditions are equivalent to the equations $p_1 p_4 = p_2 p_3$ and $p_5 p_8 = p_6 p_7$.

c. Set up the constraints P(X) = x, P($\neg X$) = $1 - x$, P($X \mid E$) = e, P($X \mid F$) = f, in terms of the p_i.

d. Apply the method of Lagrangian multipliers to set up constraints on the value of the p_i that maximize the entropy $-\sum p_i \ln p_i$ subject to the constraints derived in (c). Derive the equations in (b) from these constraints.

Note: This is not a complete proof, since it does not show that this solution is a maximum rather than a minimum or other point of zero derivative. Do not worry about it.