1

Plausible reasoning

> The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.
>
> *James Clerk Maxwell (1850)*

Suppose some dark night a policeman walks down a street, apparently deserted. Suddenly he hears a burglar alarm, looks across the street, and sees a jewelry store with a broken window. Then a gentleman wearing a mask comes crawling out through the broken window, carrying a bag which turns out to be full of expensive jewelry. The policeman doesn't hesitate at all in deciding that this gentleman is dishonest. But by what reasoning process does he arrive at this conclusion? Let us first take a leisurely look at the general nature of such problems.

1.1 Deductive and plausible reasoning

A moment's thought makes it clear that our policeman's conclusion was not a logical deduction from the evidence; for there may have been a perfectly innocent explanation for everything. It might be, for example, that this gentleman was the owner of the jewelry store and he was coming home from a masquerade party, and didn't have the key with him. However, just as he walked by his store, a passing truck threw a stone through the window, and he was only protecting his own property.

Now, while the policeman's reasoning process was not logical deduction, we will grant that it had a certain degree of validity. The evidence did not make the gentleman's dishonesty *certain*, but it did make it extremely *plausible*. This is an example of a kind of reasoning in which we have all become more or less proficient, necessarily, long before studying mathematical theories. We are hardly able to get through one waking hour without facing some situation (e.g. will it rain or won't it?) where we do not have enough information to permit deductive reasoning; but still we must decide immediately what to do.

In spite of its familiarity, the formation of plausible conclusions is a very subtle process. Although history records discussions of it extending over 24 centuries, probably nobody has ever produced an analysis of the process which anyone else finds completely satisfactory. In this work we will be able to report some useful and encouraging new progress, in which conflicting intuitive judgments are replaced by definite theorems, and *ad hoc* procedures are replaced by rules that are determined uniquely by some very elementary – and nearly inescapable – criteria of rationality.

All discussions of these questions start by giving examples of the contrast between deductive reasoning and plausible reasoning. As is generally credited to the *Organon* of Aristotle (fourth century BC)[1] deductive reasoning (*apodeixis*) can be analyzed ultimately into the repeated application of two strong syllogisms:

$$\begin{array}{c}\text{if } A \text{ is true, then } B \text{ is true}\\ A \text{ is true}\\ \hline \text{therefore, } B \text{ is true,}\end{array} \tag{1.1}$$

and its inverse:

$$\begin{array}{c}\text{if } A \text{ is true, then } B \text{ is true}\\ B \text{ is false}\\ \hline \text{therefore, } A \text{ is false.}\end{array} \tag{1.2}$$

This is the kind of reasoning we would like to use all the time; but, as noted, in almost all the situations confronting us we do not have the right kind of information to allow this kind of reasoning. We fall back on weaker syllogisms (*epagoge*):

$$\begin{array}{c}\text{if } A \text{ is true, then } B \text{ is true}\\ B \text{ is true}\\ \hline \text{therefore, } A \text{ becomes more plausible.}\end{array} \tag{1.3}$$

The evidence does not prove that A is true, but verification of one of its consequences does give us more confidence in A. For example, let

$$\begin{aligned} A &\equiv \text{it will start to rain by 10 AM at the latest;}\\ B &\equiv \text{the sky will become cloudy before 10 AM.}\end{aligned}$$

Observing clouds at 9:45 AM does not give us a logical certainty that the rain will follow; nevertheless our common sense, obeying the weak syllogism, may induce us to change our plans and behave *as if* we believed that it will, if those clouds are sufficiently dark.

This example shows also that the major premise, ‘if A then B’ expresses B only as a *logical* consequence of A; and not necessarily a causal physical consequence, which could be effective only at a later time. The rain at 10 AM is not the physical cause of the clouds at

[1] Today, several different views are held about the exact nature of Aristotle’s contribution. Such issues are irrelevant to our present purpose, but the interested reader may find an extensive discussion of them in Lukasiewicz (1957).

9:45 AM. Nevertheless, the proper logical connection is not in the uncertain causal direction (clouds $\Longrightarrow$ rain), but rather (rain $\Longrightarrow$ clouds), which is certain, although noncausal.

We emphasize at the outset that we are concerned here with *logical* connections, because some discussions and applications of inference have fallen into serious error through failure to see the distinction between logical implication and physical causation. The distinction is analyzed in some depth by Simon and Rescher (1966), who note that all attempts to interpret implication as expressing physical causation founder on the lack of contraposition expressed by the second syllogism (1.2). That is, if we tried to interpret the major premise as 'A is the physical cause of B', then we would hardly be able to accept that 'not-B is the physical cause of not-A'. In Chapter 3 we shall see that attempts to interpret plausible inferences in terms of physical causation fare no better.

Another weak syllogism, still using the same major premise, is

$$\begin{array}{c}\text{If } A \text{ is true, then } B \text{ is true}\\ A \text{ is false}\\ \hline \text{therefore, } B \text{ becomes less plausible.}\end{array} \tag{1.4}$$

In this case, the evidence does not prove that B is false; but one of the possible reasons for its being true has been eliminated, and so we feel less confident about B. The reasoning of a scientist, by which he accepts or rejects his theories, consists almost entirely of syllogisms of the second and third kind.

Now, the reasoning of our policeman was not even of the above types. It is best described by a still weaker syllogism:

$$\begin{array}{c}\text{If } A \text{ is true, then } B \text{ becomes more plausible}\\ B \text{ is true}\\ \hline \text{therefore, } A \text{ becomes more plausible.}\end{array} \tag{1.5}$$

But in spite of the apparent weakness of this argument, when stated abstractly in terms of A and B, we recognize that the policeman's conclusion has a very strong convincing power. There is something which makes us believe that, in this particular case, his argument had almost the power of deductive reasoning.

These examples show that the brain, in doing plausible reasoning, not only decides whether something becomes more plausible or less plausible, but that it evaluates the *degree* of plausibility in some way. The plausibility for rain by 10 AM depends very much on the darkness of those clouds at 9:45. And the brain also makes use of old information as well as the specific new data of the problem; in deciding what to do we try to recall our past experience with clouds and rain, and what the weatherman predicted last night.

To illustrate that the policeman was also making use of the past experience of policemen in general, we have only to change that experience. Suppose that events like these happened several times every night to every policeman – and that in every case the gentleman turned out to be completely innocent. Very soon, policemen would learn to ignore such trivial things.

Thus, in our reasoning we depend very much on *prior information* to help us in evaluating the degree of plausibility in a new problem. This reasoning process goes on unconsciously, almost instantaneously, and we conceal how complicated it really is by calling it *common sense*.

The mathematician George Pólya (1945, 1954) wrote three books about plausible reasoning, pointing out a wealth of interesting examples and showing that there are definite rules by which we do plausible reasoning (although in his work they remain in qualitative form). The above weak syllogisms appear in his third volume. The reader is strongly urged to consult Pólya's exposition, which was the original source of many of the ideas underlying the present work. We show below how Pólya's principles may be made quantitative, with resulting useful applications.

Evidently, the deductive reasoning described above has the property that we can go through long chains of reasoning of the type (1.1) and (1.2) and the conclusions have just as much certainty as the premises. With the other kinds of reasoning, (1.3)–(1.5), the reliability of the conclusion changes as we go through several stages. But in their quantitative form we shall find that in many cases our conclusions can still approach the certainty of deductive reasoning (as the example of the policeman leads us to expect). Pólya showed that even a pure mathematician actually uses these weaker forms of reasoning most of the time. Of course, on publishing a new theorem, the mathematician will try very hard to invent an argument which uses only the first kind; but the reasoning process which led to the theorem in the first place almost always involves one of the weaker forms (based, for example, on following up conjectures suggested by analogies). The same idea is expressed in a remark of S. Banach (quoted by S. Ulam, 1957):

> Good mathematicians see analogies between theorems; great mathematicians see analogies between analogies.

As a first orientation, then, let us note some very suggestive analogies to another field – which is itself based, in the last analysis, on plausible reasoning.

1.2 Analogies with physical theories

In physics, we learn quickly that the world is too complicated for us to analyze it all at once. We can make progress only if we dissect it into little pieces and study them separately. Sometimes, we can invent a mathematical model which reproduces several features of one of these pieces, and whenever this happens we feel that progress has been made. These models are called *physical theories*. As knowledge advances, we are able to invent better and better models, which reproduce more and more features of the real world, more and more accurately. Nobody knows whether there is some natural end to this process, or whether it will go on indefinitely.

In trying to understand common sense, we shall take a similar course. We won't try to understand it all at once, but we shall feel that progress has been made if we are able to construct idealized mathematical models which reproduce a few of its features. We expect that any model we are now able to construct will be replaced by more complete ones in the future, and we do not know whether there is any natural end to this process.

The analogy with physical theories is deeper than a mere analogy of method. Often, the things which are most familiar to us turn out to be the hardest to understand. Phenomena whose very existence is unknown to the vast majority of the human race (such as the difference in ultraviolet spectra of iron and nickel) can be explained in exhaustive mathematical detail – but all of modern science is practically helpless when faced with the complications of such a commonplace fact as growth of a blade of grass. Accordingly, we must not expect too much of our models; we must be prepared to find that some of the most familiar features of mental activity may be ones for which we have the greatest difficulty in constructing any adequate model.

There are many more analogies. In physics we are accustomed to finding that any advance in knowledge leads to consequences of great practical value, but of an unpredictable nature. Röntgen's discovery of X-rays led to important new possibilities of medical diagnosis; Maxwell's discovery of one more term in the equation for curl H led to practically instantaneous communication all over the earth.

Our mathematical models for common sense also exhibit this feature of practical usefulness. Any successful model, even though it may reproduce only a few features of common sense, will prove to be a powerful extension of common sense in some field of application. Within this field, it enables us to solve problems of inference which are so involved in complicated detail that we would never attempt to solve them without its help.

1.3 The thinking computer

Models have practical uses of a quite different type. Many people are fond of saying, 'They will never make a machine to replace the human mind – it does many things which no machine could ever do.' A beautiful answer to this was given by J. von Neumann in a talk on computers given in Princeton in 1948, which the writer was privileged to attend. In reply to the canonical question from the audience ('But of course, a mere machine can't really *think*, can it?'), he said:

> You insist that there is something a machine cannot do. If you will tell me precisely what it is that a machine cannot do, then I can always make a machine which will do just that!

In principle, the only operations which a machine cannot perform for us are those which we cannot describe in detail, or which could not be completed in a finite number of steps. Of course, some will conjure up images of Gödel incompleteness, undecidability, Turing machines which never stop, etc. But to answer all such doubts we need only point to the existence of the human brain, which *does* it. Just as von Neumann indicated, the only real limitations on making 'machines which think' are our own limitations in not knowing exactly what 'thinking' consists of.

But in our study of common sense we shall be led to some very explicit ideas about the mechanism of thinking. Every time we can construct a mathematical model which reproduces a part of common sense by prescribing a definite set of operations, this shows us how to 'build a machine', (i.e. write a computer program) which operates on incomplete information and, by applying quantitative versions of the above weak syllogisms, does plausible reasoning instead of deductive reasoning.

Indeed, the development of such computer software for certain specialized problems of inference is one of the most active and useful current trends in this field. One kind of problem thus dealt with might be: given a mass of data, comprising 10 000 separate observations, determine in the light of these data and whatever prior information is at hand, the relative plausibilities of 100 different possible hypotheses about the causes at work.

Our unaided common sense might be adequate for deciding between two hypotheses whose consequences are very different; but, in dealing with 100 hypotheses which are not very different, we would be helpless without a computer *and* a well-developed mathematical theory that shows us how to program it. That is, what determines, in the policeman's syllogism (1.5), whether the plausibility for A increases by a large amount, raising it almost to certainty; or only a negligibly small amount, making the data B almost irrelevant? The object of the present work is to develop the mathematical theory which answers such questions, in the greatest depth and generality now possible.

While we expect a mathematical theory to be useful in programming computers, the idea of a thinking computer is also helpful psychologically in developing the mathematical theory. The question of the reasoning process used by actual human brains is charged with emotion and grotesque misunderstandings. It is hardly possible to say anything about this without becoming involved in debates over issues that are not only undecidable in our present state of knowledge, but are irrelevant to our purpose here.

Obviously, the operation of real human brains is so complicated that we can make no pretense of explaining its mysteries; and in any event we are not trying to explain, much less reproduce, all the aberrations and inconsistencies of human brains. That is an interesting and important subject; but it is not the subject we are studying here. Our topic is the *normative principles of logic*, and not the principles of psychology or neurophysiology.

To emphasize this, instead of asking, 'How can we build a mathematical model of human common sense?', let us ask, 'How could we build a machine which would carry out useful plausible reasoning, following clearly defined principles expressing an idealized common sense?'

1.4 Introducing the robot

In order to direct attention to constructive things and away from controversial irrelevancies, we shall invent an imaginary being. Its brain is to be designed *by us*, so that it reasons according to certain definite rules. These rules will be deduced from simple desiderata which, it appears to us, would be desirable in human brains; i.e. we think that a rational person, on discovering that they were violating one of these desiderata, would wish to revise their thinking.

In principle, we are free to adopt any rules we please; that is our way of *defining* which robot we shall study. Comparing its reasoning with yours, if you find no resemblance you are in turn free to reject our robot and design a different one more to your liking. But if you find a very strong resemblance, and decide that you want and trust this robot to help you in your own problems of inference, then that will be an accomplishment of the theory, not a premise.

Our robot is going to reason about propositions. As already indicated above, we shall denote various propositions by italicized capital letters, $\{A, B, C, \text{etc.}\}$, and for the time being we must require that any proposition used must have, to the robot, an unambiguous meaning and must be of the simple, definite logical type that must be either true or false. That is, until otherwise stated, we shall be concerned only with two-valued logic, or Aristotelian logic. We do not require that the truth or falsity of such an 'Aristotelian proposition' be ascertainable by any feasible investigation; indeed, our inability to do this is usually just the reason why we need the robot's help. For example, the writer personally considers both of the following propositions to be true:

$$
\begin{aligned}
A &\equiv \text{Beethoven and Berlioz never met.}\\
B &\equiv \text{Beethoven's music has a better sustained quality than that of}\\
&\ \text{Berlioz, although Berlioz at his best is the equal of anybody.}
\end{aligned}
$$

Proposition B is not a permissible one for our robot to think about at present, whereas proposition A is, although it is unlikely that its truth or falsity could be definitely established today.[2] After our theory is developed, it will be of interest to see whether the present restriction to Aristotelian propositions such as A can be relaxed, so that the robot might help us also with more vague propositions such as B (see Chapter 18 on the A_p-distribution).[3]

1.5 Boolean algebra

To state these ideas more formally, we introduce some notation of the usual symbolic logic, or Boolean algebra, so called because George Boole (1854) introduced a *notation* similar to the following. Of course, the principles of deductive logic itself were well understood centuries before Boole, and, as we shall see, all the results that follow from Boolean algebra were contained already as special cases in the rules of plausible inference given

[2] Their meeting is a chronological possibility, since their lives overlapped by 24 years; my reason for doubting it is the failure of Berlioz to mention any such meeting in his memoirs – on the other hand, neither does he come out and say definitely that they did *not* meet.

[3] The question of how one is to make a machine in some sense 'cognizant' of the conceptual meaning that a proposition like A has to humans, might seem very difficult, and much of the subject of artificial intelligence is devoted to inventing *ad hoc* devices to deal with this problem. However, we shall find in Chapter 4 that for us the problem is almost nonexistent; our rules for plausible reasoning automatically provide the means to do the mathematical equivalent of this.

by (1812). The symbol

$$AB, \tag{1.6}$$

called the *logical product* or the *conjunction*, denotes the proposition 'both A and B are true'. Obviously, the order in which we state them does not matter; AB and BA say the same thing. The expression

$$A + B, \tag{1.7}$$

called the *logical sum* or *disjunction*, stands for 'at least one of the propositions, A, B is true' and has the same meaning as $B + A$. These symbols are only a shorthand way of writing propositions, and do not stand for numerical values.

Given two propositions A, B, it may happen that one is true if and only if the other is true; we then say that they have the same *truth value*. This may be only a simple tautology (i.e. A and B are verbal statements which obviously say the same thing), or it may be that only after immense mathematical labor is it finally proved that A is the necessary and sufficient condition for B. From the standpoint of logic it does not matter; once it is established, by any means, that A and B have the same truth value, then they are logically equivalent propositions, in the sense that any evidence concerning the truth of one pertains equally well to the truth of the other, and they have the same implications for any further reasoning.

Evidently, then, it must be the most primitive axiom of plausible reasoning that two propositions with the same truth value are equally plausible. This might appear almost too trivial to mention, were it not for the fact that Boole himself (Boole, 1854, p. 286) fell into error on this point, by mistakenly identifying two propositions which were in fact different – and then failing to see any contradiction in their different plausibilities. Three years later, Boole (1857) gave a revised theory which supersedes that in his earlier book; for further comments on this incident, see Keynes (1921, pp. 167–168); Jaynes (1976, pp. 240–242).

In Boolean algebra, the equal sign is used to denote not equal numerical value, but equal truth value: $A = B$, and the 'equations' of Boolean algebra thus consist of assertions that the proposition on the left-hand side has the same truth value as the one on the right-hand side. The symbol '$\equiv$' means, as usual, 'equals by definition'.

In denoting complicated propositions we use parentheses in the same way as in ordinary algebra, i.e. to indicate the order in which propositions are to be combined (at times we shall use them also merely for clarity of expression although they are not strictly necessary). In their absence we observe the rules of algebraic hierarchy, familiar to those who use hand calculators: thus $AB + C$ denotes $(AB) + C$; and not $A(B + C)$.

The *denial* of a proposition is indicated by a bar:

$$\overline{A} \equiv A \text{ is false.} \tag{1.8}$$

The relation between A, $\overline{A}$ is a reciprocal one:

$$A = \overline{A} \text{ is false,} \tag{1.9}$$

and it does not matter which proposition we denote by the barred and which by the unbarred letter. Note that some care is needed in the unambiguous use of the bar. For example, according to the above conventions,

$$\overline{AB} = AB \text{ is false;} \tag{1.10}$$

$$\overline{A}\,\overline{B} = \text{both } A \text{ and } B \text{ are false.} \tag{1.11}$$

These are quite different propositions; in fact, $\overline{AB}$ is not the logical product $\overline{A}\,\overline{B}$, but the logical sum: $\overline{AB} = \overline{A} + \overline{B}$.

With these understandings, Boolean algebra is characterized by some rather trivial and obvious basic identities, which express the properties of:

$$\begin{array}{ll}
\text{Idempotence:} & \begin{cases} AA = A \\ A + A = A \end{cases} \\[2ex]
\text{Commutativity:} & \begin{cases} AB = BA \\ A + B = B + A \end{cases} \\[2ex]
\text{Associativity:} & \begin{cases} A(BC) = (AB)C = ABC \\ A + (B + C) = (A + B) + C = A + B + C \end{cases} \\[2ex]
\text{Distributivity:} & \begin{cases} A(B + C) = AB + AC \\ A + (BC) = (A + B)(A + C) \end{cases} \\[2ex]
\text{Duality:} & \begin{cases} \text{If } C = AB, \text{ then } \overline{C} = \overline{A} + \overline{B} \\ \text{If } D = A + B, \text{ then } \overline{D} = \overline{A}\,\overline{B} \end{cases}
\end{array} \tag{1.12}$$

but by their application one can prove any number of further relations, some highly nontrivial. For example, we shall presently have use for the rather elementary theorem:

$$\text{if } \overline{B} = AD \text{ then } A\overline{B} = \overline{B} \text{ and } B\overline{A} = \overline{A}. \tag{1.13}$$

Implication

The proposition

$$A \Rightarrow B \tag{1.14}$$

to be read as 'A implies B', does not assert that either A or B is true; it means only that $A\overline{B}$ is false, or, what is the same thing, $(\overline{A} + B)$ is true. This can be written also as the logical equation $A = AB$. That is, given (1.14), if A is true then B must be true; or, if B is false then A must be false. This is just what is stated in the strong syllogisms (1.1) and (1.2).

On the other hand, if A is false, (1.14) says nothing about B: and if B is true, (1.14) says nothing about A. But these are just the cases in which our weak syllogisms (1.3), (1.4) do say something. In one respect, then, the term 'weak syllogism' is misleading. The theory of plausible reasoning based on weak syllogisms is not a 'weakened' form of logic; it is an *extension* of logic with new content not present at all in conventional deductive logic. It will become clear in the next chapter (see (2.69) and (2.70)) that our rules include deductive logic as a special case.

A tricky point

Note carefully that in ordinary language one would take 'A implies B' to mean that B is logically deducible from A. But, in formal logic, 'A implies B' means only that the propositions A and AB have the same truth value. In general, whether B is logically deducible from A does not depend only on the propositions A and B; it depends on the totality of propositions $(A, A', A'', \ldots)$ that we accept as true and which are therefore available to use in the deduction. Devinatz (1968, p. 3) and Hamilton (1988, p. 5) give the truth table for the implication as a binary operation, illustrating that $A \Rightarrow B$ is false only if A is true and B is false; in all other cases $A \Rightarrow B$ is true!

This may seem startling at first glance; however, note that, indeed, if A and B are both true, then $A = AB$ and so $A \Rightarrow B$ is true; in formal logic every true statement implies every other true statement. On the other hand, if A is false, then AQ is also false for all Q, thus $A = AB$ and $A = A\overline{B}$ are both true, so $A \Rightarrow B$ and $A \Rightarrow \overline{B}$ are both true; a false proposition implies all propositions. If we tried to interpret this as logical deducibility (i.e. both B and $\overline{B}$ are deducible from A), it would follow that every false proposition is logically contradictory. Yet the proposition: 'Beethoven outlived Berlioz' is false but hardly logically contradictory (for Beethoven did outlive many people who were the same age as Berlioz).

Obviously, merely knowing that propositions A and B are both true does not provide enough information to decide whether either is logically deducible from the other, plus some unspecified 'toolbox' of other propositions. The question of logical deducibility of one proposition from a set of others arises in a crucial way in the Gödel theorem discussed at the end of Chapter 2. This great difference in the meaning of the word 'implies' in ordinary language and in formal logic is a tricky point that can lead to serious error if it is not properly understood; it appears to us that 'implication' is an unfortunate choice of word, and that this is not sufficiently emphasized in conventional expositions of logic.

1.6 Adequate sets of operations

We note some features of deductive logic which will be needed in the design of our robot. We have defined four operations, or 'connectives', by which, starting from two propositions A, B, other propositions may be defined: the logical product or conjunction AB, the logical sum or disjunction $A + B$, the implication $A \Rightarrow B$, and the negation $\overline{A}$. By combining these operations repeatedly in every possible way, one can generate any number of new propositions, such as

$$C \equiv (A + \overline{B})(\overline{A} + A\,\overline{B}) + \overline{A}\,B(A + B). \tag{1.15}$$

Many questions then occur to us: How large is the class of new propositions thus generated? Is it infinite, or is there a finite set that is closed under these operations? Can every proposition defined from $A,\ B$ be thus represented, or does this require further connectives beyond the above four? Or are these four already overcomplete so that some might be dispensed with? What is the smallest set of operations that is adequate to generate all such 'logic functions' of A and B? If instead of two starting propositions A, B we have an arbitrary number $\{A_1, \dots, A_n\}$, is this set of operations still adequate to generate all possible logic functions of $\{A_1, \dots, A_n\}$?

All these questions are answered easily, with results useful for logic, probability theory, and computer design. Broadly speaking, we are asking whether, starting from our present vantage point, we can (1) increase the number of functions, (2) decrease the number of operations. The first query is simplified by noting that two propositions, although they may appear entirely different when written out in the manner (1.15), are not different propositions from the standpoint of logic if they have the same truth value. For example, it is left for the reader to verify that C in (1.15) is logically the same statement as the implication $C = (B \Rightarrow \overline{A})$.

Since we are, at this stage, restricting our attention to Aristotelian propositions, any logic function $C = f(A, B)$ such as (1.15) has only two possible 'values', true and false; and likewise the 'independent variables' A and B can take on only those two values.

At this point, a logician might object to our notation, saying that the symbol A has been defined as standing for some fixed proposition, whose truth cannot change; so if we wish to consider logic functions, then instead of writing $C = f(A, B)$ we should introduce new symbols and write $z = f(x, y)$, where x, y, z, are 'statement variables' for which various specific statements A, B, C may be substituted. But if A stands for some fixed but unspecified proposition, then it can still be either true or false. We achieve the same flexibility merely by the understanding that equations like (1.15) which define logic functions are to be true for all ways of defining A, B ; i.e. instead of a statement variable we use a variable statement.

In relations of the form $C = f(A, B)$, we are concerned with logic functions defined on a discrete 'space' S consisting of only $2^2 = 4$ points; namely those at which A and B take on the 'values' {TT, TF, FT, FF}, respectively; and, at each point, the function $f(A, B)$ can take on independently either of two values {T, F}. There are, therefore, exactly $2^4 = 16$ different logic functions $f(A, B)$, and no more. An expression $B = f(A_1, \dots, A_n)$ involving n propositions is a logic function on a space S of $M = 2^n$ points; and there are exactly 2^M such functions.

In the case $n = 1$, there are four logic functions $\{f_1(A), \dots, f_4(A)\}$, which we can define by enumeration, listing all their possible values in a truth table:

A	T	F
$f_1(A)$	T	T
$f_2(A)$	T	F
$f_3(A)$	F	T
$f_4(A)$	F	F

But it is obvious by inspection that these are just

$$\begin{aligned} f_1(A) &= A + \overline{A} \\ f_2(A) &= A \\ f_3(A) &= \overline{A} \\ f_4(A) &= A\ \overline{A}, \end{aligned} \tag{1.16}$$

so we prove by enumeration that the three operations: conjunction, disjunction, and negation are adequate to generate all logic functions of a single proposition.

For the case of general n, consider first the special functions, each of which is true at one and only one point of S. For $n = 2$ there are $2^n = 4$ such functions,

A, B	TT	TF	FT	FF
$f_1(A, B)$	T	F	F	F
$f_2(A, B)$	F	T	F	F
$f_3(A, B)$	F	F	T	F
$f_4(A, B)$	F	F	F	T

It is clear by inspection that these are just the four basic conjunctions,

$$\begin{aligned} f_1(A, B) &= A\ B \\ f_2(A, B) &= A\ \overline{B} \\ f_3(A, B) &= \overline{A}\ B \\ f_4(A, B) &= \overline{A}\ \overline{B}. \end{aligned} \tag{1.17}$$

Consider now any logic function which is true on certain specified points of S; for example, $f_5(A, B)$ and $f_6(A, B)$, defined by

A, B	TT	TF	FT	FF
$f_5(A, B)$	F	T	F	T
$f_6(A, B)$	T	F	T	T

We assert that each of these functions is the logical sum of the conjunctions (1.17) that are true on the same points (this is not trivial; the reader should verify it in detail). Thus,

$$\begin{aligned} f_5(A, B) &= f_2(A, B) + f_4(A, B) \\ &= A\,\overline{B} + \overline{A}\,\overline{B} \\ &= (A + \overline{A})\,\overline{B} \\ &= \overline{B}, \end{aligned} \tag{1.18}$$

and, likewise,

$$\begin{aligned} f_6(A, B) &= f_1(A, B) + f_3(A, B) + f_4(A, B) \\ &= AB + \overline{A}\,B + \overline{A}\,\overline{B} \\ &= B + \overline{A}\,\overline{B} \\ &= \overline{A} + B. \end{aligned} \tag{1.19}$$

That is, $f_6(A, B)$ is the implication $f_6(A, B) = (A \Rightarrow B)$, with the truth table discussed above. Any logic function $f(A, B)$ that is true on at least one point of S can be constructed in this way as a logical sum of the basic conjunctions (1.17). There are $2^4 - 1 = 15$ such functions. For the remaining function, which is always false, it suffices to take the contradiction, $f_{16}(A, B) \equiv A\,\overline{A}$.

This method (called 'reduction to *disjunctive normal form*' in logic textbooks) will work for any n. For example, in the case $n = 5$ there are $2^5 = 32$ basic conjunctions,

$$\{ABCDE,\ ABCD\overline{E},\ ABC\overline{D}E, \ldots, \overline{A}\,\overline{B}\,\overline{C}\,\overline{D}\,\overline{E}\}, \tag{1.20}$$

and $2^{32} = 4\,294\,967\,296$ different logic functions $f_i(A, B, C, D, E)$; of which 4 294 967 295 can be written as logical sums of the basic conjunctions, leaving only the contradiction

$$f_{4294967296}(A, B, C, D, E) = A\,\overline{A}. \tag{1.21}$$

Thus one can verify by 'construction in thought' that the three operations

$$\{\text{conjunction, disjunction, negation}\}, \qquad \text{i.e.} \qquad \{\text{AND, OR, NOT}\}, \tag{1.22}$$

suffice to generate all possible logic functions; or, more concisely, they form an *adequate set*.

The duality property (1.12) shows that a smaller set will suffice; for disjunction of A, B is the same as denying that they are both false:

$$A + B = \overline{(\overline{A}\,\overline{B})}. \tag{1.23}$$

Therefore, the two operations (AND, NOT) already constitute an adequate set for deductive logic.[4] This fact will be essential in determining when we have an adequate set of rules for plausible reasoning; see Chapter 2.

[4] For you to ponder: Does it follow that these two commands are the only ones needed to write any computer program?

It is clear that we cannot now strike out either of these operations, leaving only the other; i.e. the operation 'AND' cannot be reduced to negations; and negation cannot be accomplished by any number of 'AND' operations. But this still leaves open the possibility that both conjunction and negation might be reducible to some third operation, not yet introduced, so that a single logic operation would constitute an adequate set.

It comes as a pleasant surprise to find that there is not only one but two such operations. The operation 'NAND' is defined as the negation of 'AND':

$$A \uparrow B \equiv \overline{AB} = \overline{A} + \overline{B} \tag{1.24}$$

which we can read as 'A NAND B'. But then we have at once

$$\begin{aligned} \overline{A} &= A \uparrow A \\ AB &= (A \uparrow B) \uparrow (A \uparrow B) \\ A + B &= (A \uparrow A) \uparrow (B \uparrow B). \end{aligned} \tag{1.25}$$

Therefore, every logic function can be constructed with NAND alone. Likewise, the operation NOR defined by

$$A \downarrow B \equiv \overline{A + B} = \overline{A}\,\overline{B} \tag{1.26}$$

is also powerful enough to generate all logic functions:

$$\begin{aligned} \overline{A} &= A \downarrow A \\ A + B &= (A \downarrow B) \downarrow (A \downarrow B) \\ AB &= (A \downarrow A) \downarrow (B \downarrow B). \end{aligned} \tag{1.27}$$

One can take advantage of this in designing computer and logic circuits. A 'logic gate' is a circuit having, besides a common ground, two input terminals and one output. The voltage relative to ground at any of these terminals can take on only two values; say +3 volts, or 'up', representing 'true'; and 0 volts or 'down', representing 'false'. A NAND gate is thus one whose output is up if and only if at least one of the inputs is down; or, what is the same thing, down if and only if both inputs are up; while for a NOR gate the output is up if and only if both inputs are down.

One of the standard components of logic circuits is the 'quad NAND gate', an integrated circuit containing four independent NAND gates on one semiconductor chip. Given a sufficient number of these and no other circuit components, it is possible to generate any required logic function by interconnecting them in various ways.

This short excursion into deductive logic is as far as we need go for our purposes. Further developments are given in many textbooks; for example, a modern treatment of Aristotelian logic is given by Copi (1994). For non-Aristotelian forms with special emphasis on Gödel incompleteness, computability, decidability, Turing machines, etc., see Hamilton (1988).

We turn now to our extension of logic, which is to follow from the conditions discussed next. We call them 'desiderata' rather than 'axioms' because they do not assert that anything is 'true' but only state what appear to be desirable goals. Whether these goals are attainable without contradictions, and whether they determine any unique extension of logic, are matters of mathematical analysis, given in Chapter 2.

1.7 The basic desiderata

To each proposition about which it reasons, our robot must assign some degree of plausibility, based on the evidence we have given it; and whenever it receives new evidence it must revise these assignments to take that new evidence into account. In order that these plausibility assignments can be stored and modified in the circuits of its brain, they must be associated with some definite physical quantity, such as voltage or pulse duration or a binary coded number, etc. – however our engineers want to design the details. For present purposes, this means that there will have to be some kind of association between degrees of plausibility and real numbers:

(I) *Degrees of plausibility are represented by real numbers.* (1.28)

Desideratum (I) is practically forced on us by the requirement that the robot's brain must operate by the carrying out of some definite physical process. However, it will appear (Appendix A) that it is also required theoretically; we do not see the possibility of any consistent theory without a property that is equivalent functionally to desideratum (I).

We adopt a natural but nonessential convention: that a greater plausibility shall correspond to a greater number. It will also be convenient to assume a continuity property, which is hard to state precisely at this stage; to say it intuitively: an infinitesimally greater plausibility ought to correspond only to an infinitesimally greater number.

The plausibility that the robot assigns to some proposition A will, in general, depend on whether we told it that some other proposition B is true. Following the notation of Keynes (1921) and Cox (1961), we indicate this by the symbol

$$A|B, \tag{1.29}$$

which we may call 'the conditional plausibility that A is true, given that B is true' or just 'A given B'. It stands for some real number. Thus, for example,

$$A|BC \tag{1.30}$$

(which we may read as 'A given BC') represents the plausibility that A is true, given that both B and C are true. Or,

$$A + B|CD \tag{1.31}$$

represents the plausibility that at least one of the propositions A and B is true, given that both C and D are true; and so on. We have decided to represent a greater plausibility by a greater number, so

$$(A|B) > (C|B) \tag{1.32}$$

says that, given B, A is more plausible than C. In this notation, while the symbol for plausibility is just of the form $A|B$ without parentheses, we often add parentheses for clarity of expression. Thus, (1.32) says the same thing as

$$A|B \; > \; C|B, \tag{1.33}$$

but its meaning is clearer to the eye.

In the interest of avoiding impossible problems, we are not going to ask our robot to undergo the agony of reasoning from impossible or mutually contradictory premises; there could be no 'correct' answer. Thus, we make no attempt to define $A|BC$ when B and C are mutually contradictory. Whenever such a symbol appears, it is understood that B and C are compatible propositions.

Also, we do not want this robot to think in a way that is directly opposed to the way you and I think. So we shall design it to reason in a way that is at least *qualitatively* like the way humans try to reason, as described by the above weak syllogisms and a number of other similar ones.

Thus, if it has old information C which gets updated to C' in such a way that the plausibility for A is increased:

$$(A|C') \; > \; (A|C); \tag{1.34}$$

but the plausibility for B given A is not changed:

$$(B|AC') \; = \; (B|AC). \tag{1.35}$$

This can, of course, produce only an increase, never a decrease, in the plausibility that both A and B are true:

$$(AB|C') \; \geq \; (AB|C); \tag{1.36}$$

and it must produce a decrease in the plausibility that A is false:

$$(\overline{A}|C') \; < \; (\overline{A}|C). \tag{1.37}$$

This qualitative requirement simply gives the 'sense of direction' in which the robot's reasoning is to go; it says nothing about *how much* the plausibilities change, except that our continuity assumption (which is also a condition for qualitative correspondence with common sense) now requires that if $A|C$ changes only infinitesimally, it can induce only an infinitesimal change in $AB|C$ and $\overline{A}|C$. The specific ways in which we use these qualitative requirements will be given in the next chapter, at the point where it is seen why we need them. For the present we summarize them simply as:

(II) *Qualitative correspondence with common sense.* (1.38)

Finally, we want to give our robot another desirable property for which honest people strive without always attaining: that it always reasons *consistently*. By this we mean just the three common colloquial meanings of the word ‘consistent’:

(IIIa) *If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.* (1.39a)

(IIIb) *The robot always takes into account all of the evidence it has relevant to a question. It does not arbitrarily ignore some of the information, basing its conclusions only on what remains. In other words, the robot is completely nonideological.* (1.39b)

(IIIc) *The robot always represents equivalent states of knowledge by equivalent plausibility assignments. That is, if in two problems the robot’s state of knowledge is the same (except perhaps for the labeling of the propositions), then it must assign the same plausibilities in both.* (1.39c)

Desiderata (I), (II), and (IIIa) are the basic ‘structural’ requirements on the inner workings of our robot’s brain, while (IIIb) and (IIIc) are ‘interface’ conditions which show how the robot’s behavior should relate to the outer world.

At this point, most students are surprised to learn that our search for desiderata is at an end. The above conditions, it turns out, uniquely determine the rules by which our robot must reason; i.e. there is only one set of mathematical operations for manipulating plausibilities which has all these properties. These rules are deduced in Chapter 2.

(At the end of most chapters, we insert a section of informal Comments in which are collected various side remarks, background material, etc. The reader may skip them without losing the main thread of the argument.)

1.8 Comments

As politicians, advertisers, salesmen, and propagandists for various political, economic, moral, religious, psychic, environmental, dietary, and artistic doctrinaire positions know only too well, fallible human minds are easily tricked, by clever verbiage, into committing violations of the above desiderata. We shall try to ensure that they do not succeed with our robot.

We emphasize another contrast between the robot and a human brain. By Desideratum I, the robot’s mental state about any proposition is to be represented by a real number. Now, it is clear that our attitude toward any given proposition may have more than one ‘coordinate’. You and I form simultaneous judgments about a proposition not only as to whether it is plausible, but also whether it is desirable, whether it is important, whether it is useful, whether it is interesting, whether it is amusing, whether it is morally right, etc. If we assume that each of these judgments might be represented by a number, then a fully adequate description of a human state of mind would be represented by a vector in a space of a rather large number of dimensions.

Not all propositions require this. For example, the proposition 'The refractive index of water is less than 1.3' generates no emotions; consequently the state of mind which it produces has very few coordinates. On the other hand, the proposition, 'Your mother-in-law just wrecked your new car' generates a state of mind with many coordinates. Quite generally, the situations of everyday life are those involving many coordinates. It is just for this reason, we suggest, that the most familiar examples of mental activity are often the most difficult to reproduce by a model. Perhaps we have here the reason why science and mathematics are the most successful of human activities: they deal with propositions which produce the simplest of all mental states. Such states would be the ones least perturbed by a given amount of imperfection in the human mind.

Of course, for many purposes we would not want our robot to adopt any of these more 'human' features arising from the other coordinates. It is just the fact that computers do *not* get confused by emotional factors, do *not* get bored with a lengthy problem, do *not* pursue hidden motives opposed to ours, that makes them safer agents than men for carrying out certain tasks.

These remarks are interjected to point out that there is a large unexplored area of possible generalizations and extensions of the theory to be developed here; perhaps this may inspire others to try their hand at developing 'multidimensional theories' of mental activity, which would more and more resemble the behavior of actual human brains – not all of which is undesirable. Such a theory, if successful, might have an importance beyond our present ability to imagine.[5]

For the present, however, we shall have to be content with a much more modest undertaking. Is it possible to develop a consistent 'one-dimensional' model of plausible reasoning? Evidently, our problem will be simplest if we can manage to represent a degree of plausibility uniquely by a single real number, and ignore the other 'coordinates' just mentioned.

We stress that we are in no way asserting that degrees of plausibility in actual human minds have a unique numerical measure. Our job is not to postulate – or indeed to conjecture about – any such thing; it is to *investigate* whether it is possible, in our robot, to set up such a correspondence without contradictions.

But to some it may appear that we have already assumed more than is necessary, thereby putting gratuitous restrictions on the generality of our theory. Why must we represent degrees of plausibility by real numbers? Would not a 'comparative' theory based on a system of qualitative ordering relations such as $(A|C) > (B|C)$ suffice? This point is discussed further in Appendix A, where we describe other approaches to probability theory and note that some attempts have been made to develop comparative theories which it was thought would be logically simpler, or more general. But this turned out not to be the case; so, although it is quite possible to develop the foundations in other ways than ours, the final results will not be different.

[5] Indeed, some psychologists think that as few as five dimensions might suffice to characterize a human personality; that is, that we all differ only in having different mixes of five basic personality traits which may be genetically determined. But it seems to us that this must be grossly oversimplified; identifiable chemical factors continuously varying in both space and time (such as the distribution of glucose metabolism in the brain) affect mental activity but cannot be represented faithfully in a space of only five dimensions. Yet it may be that five numbers can capture enough of the truth to be useful for many purposes.

1.8.1 Common language vs. formal logic

We should note the distinction between the statements of formal logic and those of ordinary language. It might be thought that the latter is only a less precise form of expression; but on examination of details the relation appears different. It appears to us that ordinary language, carefully used, need not be less precise than formal logic; but ordinary language is more complicated in its rules and has consequently richer possibilities of expression than we allow ourselves in formal logic.

In particular, common language, being in constant use for other purposes than logic, has developed subtle nuances – means of implying something without actually stating it – that are lost on formal logic. Mr A, to affirm his objectivity, says, ‘I believe what I see.’ Mr B retorts: ‘He doesn’t see what he doesn’t believe.’ From the standpoint of formal logic, it appears that they have said the same thing; yet from the standpoint of common language, those statements had the intent and effect of conveying opposite meanings.

Here is a less trivial example, taken from a mathematics textbook. Let L be a straight line in a plane, and S an infinite set of points in that plane, each of which is projected onto L. Now consider the following statements:

(I) The projection of the limit is the limit of the projections.
(II) The limit of the projections is the projection of the limit.

These have the grammatical structures ‘A is B’ and ‘B is A’, and so they might appear logically equivalent. Yet in that textbook, (I) was held to be true, and (II) not true in general, on the grounds that the limit of the projections may exist when the limit of the set does not.

As we see from this, in common language – even in mathematics textbooks – we have learned to read subtle nuances of meaning into the exact phrasing, probably without realizing it until an example like this is pointed out. We interpret ‘A is B’ as asserting first of all, as a kind of major premise, that A exists; and the rest of the statement is understood to be conditional on that premise. Put differently, in common grammar the verb ‘is’ implies a distinction between subject and object, which the symbol ‘$=$’ does not have in formal logic or in conventional mathematics. (However, in computer languages we encounter such statements as ‘$\mathrm{J} = \mathrm{J} + 1$’, which everybody seems to understand, but in which the ‘$=$’ sign has now acquired that implied distinction after all.)

Another amusing example is the old adage ‘knowledge is power’, which is a very cogent truth, both in human relations and in thermodynamics. An ad writer for a chemical trade journal[6] fouled this up into ‘power is knowledge’, an absurd – indeed, obscene – falsity.

These examples remind us that the verb ‘is’ has, like any other verb, a subject and a predicate; but it is seldom noted that this verb has two entirely different meanings. A person whose native language is English may require some effort to see the different meanings in the statements: ‘The room is noisy’ and ‘There is noise in the room’. But in Turkish these meanings are rendered by different words, which makes the distinction so clear that a visitor

[6] *LC-CG Magazine*, March 1988, p. 211.

who uses the wrong word will not be understood. The latter statement is ontological, asserting the physical existence of something, while the former is epistemological, expressing only the speaker's personal perception.

Common language – or, at least, the English language – has an almost universal tendency to disguise epistemological statements by putting them into a grammatical form which suggests to the unwary an ontological statement. A major source of error in current probability theory arises from an unthinking failure to perceive this. To interpret the first kind of statement in the ontological sense is to assert that one's own private thoughts and sensations are realities existing externally in Nature. We call this the 'mind projection fallacy', and note the trouble it causes many times in what follows. But this trouble is hardly confined to probability theory; as soon as it is pointed out, it becomes evident that much of the discourse of philosophers and Gestalt psychologists, and the attempts of physicists to explain quantum theory, are reduced to nonsense by the author falling repeatedly into the mind projection fallacy.

These examples illustrate the care that is needed when we try to translate the complex statements of common language into the simpler statements of formal logic. Of course, common language is often less precise than we should want in formal logic. But everybody expects this and is on the lookout for it, so it is less dangerous.

It is too much to expect that our robot will grasp all the subtle nuances of common language, which a human spends perhaps 20 years acquiring. In this respect, our robot will remain like a small child – it interprets all statements literally and blurts out the truth without thought of whom this may offend.

It is unclear to the writer how difficult – and even less clear how desirable – it would be to design a newer model robot with the ability to recognize these finer shades of meaning. Of course, the question of principle is disposed of at once by the existence of the human brain, which does this. But, in practice, von Neumann's principle applies; a robot designed by us cannot do it until someone develops a theory of 'nuance recognition', which reduces the process to a definitely prescribed set of operations. This we gladly leave to others.

In any event, our present model robot is quite literally real, because today it is almost universally true that any nontrivial probability evaluation is performed by a computer. The person who programmed that computer was necessarily, whether or not they thought of it that way, designing part of the brain of a robot according to some preconceived notion of how the robot should behave. But very few of the computer programs now in use satisfy all our desiderata; indeed, most are intuitive *ad hoc* procedures that were not chosen with any well-defined desiderata at all in mind.

Any such adhockery is presumably usable within some special area of application – that was the criterion for choosing it – but as the proofs of Chapter 2 will show, any adhockery which conflicts with the rules of probability theory must generate demonstrable inconsistencies when we try to apply it beyond some restricted area. Our aim is to avoid this by developing the general principles of inference once and for all, directly from the requirement of consistency, and in a form applicable to any problem of plausible inference that is formulated in a sufficiently unambiguous way.

1.8.2 Nitpicking

As is apparent from the above, in the present work we use the term 'Boolean algebra' in its long-established meaning as referring to two-valued logic in which symbols like 'A' stand for propositions. A compulsive nitpicker has complained to us that some mathematicians have used the term in a slightly different meaning, in which 'A' could refer to a class of propositions. But the two usages are not in conflict; we recognize the broader meaning, but just find no reason to avail ourselves of it.

The set of rules and symbols that we have called 'Boolean algebra' is sometimes called 'the propositional calculus'. The term seems to be used only for the purpose of adding that we need also another set of rules and symbols called 'the predicate calculus'. However, these new symbols prove to be only abbreviations for short and familiar phrases. The 'universal quantifier' is only an abbreviation for 'for all'; the 'existential quantifier' is an abbreviation for 'there is a'. If we merely write our statements in plain English, we are using automatically all of the predicate calculus that we need for our purposes, and doing it more intelligibly.

The validity of the second strong syllogism (in two-valued logic) is sometimes questioned. However, it appears that in current mathematics it is still considered valid reasoning to say that a supposed theorem is disproved by exhibiting a counterexample, that a set of statements is considered inconsistent if we can derive a contradiction from them, and that a proposition can be established by *reductio ad absurdum*, deriving a contradiction from its denial. This is enough for us; we are quite content to follow this long tradition. Our feeling of security in this stance comes from the conviction that, while logic may move forward in the future, it can hardly move backward. A new logic might lead to new results about which Aristotelian logic has nothing to say; indeed, that is just what we are trying to create here. But surely, if a new logic was found to conflict with Aristotelian logic in an area where Aristotelian logic is applicable, we would consider that a fatal objection to the new logic.

Therefore, to those who feel confined by two-valued deductive logic, we can say only: 'By all means, investigate other possibilities if you wish to; and please let us know about it as soon as you have found a new result that was not contained in two-valued logic or our extension of it, *and* is useful in scientific inference.' Actually, there are many different and mutually inconsistent multiple-valued logics already in the literature. But in Appendix A we adduce arguments which suggest that they can have no useful content that is not already in two-valued logic; that is, that an n-valued logic applied to one set of propositions is either equivalent to a two-valued logic applied to an enlarged set, or else it contains internal inconsistencies.

Our experience is consistent with this conjecture; in practice, multiple-valued logics seem to be used not to find new useful results, but rather in attempts to remove supposed difficulties with two-valued logic, particularly in quantum theory, fuzzy sets, and artificial intelligence. But on closer study, all such difficulties known to us have proved to be only examples of the mind projection fallacy, calling for direct revision of the concepts rather than a new logic.

2

The quantitative rules

Probability theory is nothing but common sense reduced to calculation.
Laplace, 1819

We have now formulated our problem, and it is a matter of straightforward mathematics to work out the consequences of our desiderata, which may be stated broadly as follows:

(I) Representation of degrees of plausibility by real numbers;
(II) Qualitative correspondence with common sense;
(III) Consistency.

The present chapter is devoted entirely to deduction of the quantitative rules for inference which follow from these desiderata. The resulting rules have a long, complicated, and astonishing history, full of lessons for scientific methodology in general (see the Comments sections at the end of several chapters).

2.1 The product rule

We first seek a consistent rule relating the plausibility of the logical product AB to the plausibilities of A and B separately. In particular, let us find $AB|C$. Since the reasoning is somewhat subtle, we examine this from several different viewpoints.

As a first orientation, note that the process of deciding that AB is true can be broken down into elementary decisions about A and B separately. The robot can

(1) decide that B is true; $(B|C)$
(2) having accepted B as true, decide that A is true. $(A|BC)$

Or, equally well,

(1′) decide that A is true; $(A|C)$
(2′) having accepted A as true, decide that B is true. $(B|AC)$

In each case we indicate above the plausibility corresponding to that step.

Now let us describe the first procedure in words. In order for AB to be a true proposition, it is necessary that B is true. Thus the plausibility $B|C$ should be involved. In addition, if B

is true, it is further necessary that A should be true; so the plausibility $A|BC$ is also needed. But if B is false, then of course AB is false independently of whatever one knows about A, as expressed by $A|\overline{B}\,C$; if the robot reasons first about B, then the plausibility of A will be relevant only if B is true. Thus, if the robot has $B|C$ and $A|BC$ it will not need $A|C$. That would tell it nothing about AB that it did not have already.

Similarly, $A|B$ and $B|A$ are not needed; whatever plausibility A or B might have in the absence of information C could not be relevant to judgments of a case in which the robot knows that C is true. For example, if the robot learns that the earth is round, then in judging questions about cosmology today, it does not need to take into account the opinions it might have (i.e. the extra possibilities that it would need to take into account) if it did not know that the earth is round.

Of course, since the logical product is commutative, $AB = BA$, we could interchange A and B in the above statements; i.e. knowledge of $A|C$ and $B|AC$ would serve equally well to determine $AB|C = BA|C$. That the robot must obtain the same value for $AB|C$ from either procedure is one of our conditions of consistency, desideratum (IIIa).

We can state this in a more definite form. $(AB|C)$ will be some function of $B|C$ and $A|BC$:

$$(AB|C) = F[(B|C), (A|BC)]. \tag{2.1}$$

Now, if the reasoning we went through here is not completely obvious, let us examine some alternatives. We might suppose, for example, that

$$(AB|C) = F[(A|C), (B|C)] \tag{2.2}$$

might be a permissible form. But we can show easily that no relation of this form could satisfy our qualitative conditions of desideratum (II). Proposition A might be very plausible given C, and B might be very plausible given C; but AB could still be very plausible or very implausible.

For example, it is quite plausible that the next person you meet has blue eyes and also quite plausible that this person's hair is black; and it is reasonably plausible that both are true. On the other hand it is quite plausible that the left eye is blue, and quite plausible that the right eye is brown; but extremely implausible that both of those are true. We would have no way of taking such influences into account if we tried to use a formula of this kind. Our robot could not reason the way humans do, even qualitatively, with that kind of functional relation.

But other possibilities occur to us. The method of trying out all possibilities – a kind of 'proof by exhaustion' – can be organized as follows. Introduce the real numbers

$$u = (AB|C), \quad v = (A|C), \quad w = (B|AC), \quad x = (B|C), \quad y = (A|BC). \tag{2.3}$$

If u is to be expressed as a function of two or more of v, w, x, y, there are 11 possibilities. You can write out each of them, and subject each one to various extreme conditions, as in the brown and blue eyes (which was the abstract statement: A implies that B is false). Other extreme conditions are $A = B$, $A = C$, $C \Rightarrow \overline{A}$, etc. Carrying out this somewhat tedious analysis, Tribus (1969) finds that all but two of the possibilities can exhibit qualitative violations of common sense in some extreme case. The two which survive are $u = F(x, y)$ and $u = F(w, v)$, just the two functional forms already suggested by our previous reasoning.

We now apply the qualitative requirement discussed in Chapter 1. Given any change in the prior information $C \to C'$, such that B becomes more plausible but A does not change,

$$B|C' > B|C, \tag{2.4}$$

$$A|BC' = A|BC, \tag{2.5}$$

common sense demands that AB could only become more plausible, not less:

$$AB|C' \geq AB|C, \tag{2.6}$$

with equality if and only if $A|BC$ corresponds to impossibility. Likewise, given prior information C'' such that

$$B|C'' = B|C, \tag{2.7}$$

$$A|BC'' > A|BC, \tag{2.8}$$

we require that

$$AB|C'' \geq AB|C, \tag{2.9}$$

in which the equality can hold only if B is impossible, given C (for then AB might still be impossible given C'', although $A|BC$ is not defined). Furthermore, the function $F(x, y)$ must be continuous; for otherwise an arbitrarily small increase in one of the plausibilities on the right-hand side of (2.1) could result in a large increase in $AB|C$.

In summary, $F(x, y)$ must be a continuous monotonic increasing function of both x and y. If we assume it is differentiable (this is not necessary; see the discussion following (2.13)), then we have

$$F_1(x, y) \equiv \frac{\partial F}{\partial x} \geq 0 \tag{2.10a}$$

with equality if and only if y represents impossibility; and also

$$F_2(x, y) \equiv \frac{\partial F}{\partial y} \geq 0 \tag{2.10b}$$

with equality permitted only if x represents impossibility. Note for later purposes that, in this notation, F_i denotes differentiation with respect to the ith argument of F, whatever it may be.

Next we impose the desideratum (IIIa) of 'structural' consistency. Suppose we try to find the plausibility $(ABC|D)$ that three propositions would be true simultaneously. Because of the fact that Boolean algebra is associative: $ABC = (AB)C = A(BC)$, we can do this in two different ways. If the rule is to be consistent, we must get the same result for either order of carrying out the operations. We can say first that BC will be considered a single proposition, and then apply (2.1):

$$(ABC|D) = F[(BC|D), (A|BCD)], \tag{2.11}$$

and then in the plausibility $(BC|D)$ we can again apply (2.1) to give

$$(ABC|D) = F\{F[(C|D), (B|CD)], (A|BCD)\}. \tag{2.12a}$$

But we could equally well have said that AB shall be considered a single proposition at first. From this we can reason out in the other order to obtain a different expression:

$$(ABC|D) = F[(C|D), (AB|CD)] = F\{(C|D), F[(B|CD), (A|BCD)]\}. \tag{2.12b}$$

If this rule is to represent a consistent way of reasoning, the two expressions (2.12a) and (2.12b) must always be the same. A necessary condition that our robot will reason consistently in this case therefore takes the form of a functional equation,

$$F[F(x, y), z] = F[x, F(y, z)]. \tag{2.13}$$

This equation has a long history in mathematics, starting from the work of N. H. Abel (1826). Aczél (1966), in his monumental work on functional equations, calls it, very appropriately, 'The Associativity Equation', and lists a total of 98 references to works that discuss it or use it. Aczél derives the general solution (2.27), below, without assuming differentiability; unfortunately, the proof fills 11 pages (pp. 256–267) of his book (see also Aczél, 1987). We give here the shorter proof by R. T. Cox (1961), which assumes differentiability; see also the discussion in Appendix B.

It is evident that (2.13) has a trivial solution, $F(x, y) =$ const. But that violates our monotonicity requirement (2.10), and is in any event useless for our purposes. Unless (2.13) has a nontrivial solution, this approach will fail; so we seek the most general nontrivial solution. Using the abbreviations

$$u \equiv F(x, y), \qquad v \equiv F(y, z), \tag{2.14}$$

but still considering (x, y, z) the independent variables, the functional equation to be solved is

$$F(x, v) = F(u, z). \tag{2.15}$$

Differentiating with respect to x and y we obtain, in the notation of (2.10),

$$\begin{aligned} F_1(x, v) &= F_1(u, z)F_1(x, y) \\ F_2(x, v)F_1(y, z) &= F_1(u, z)F_2(x, y). \end{aligned} \tag{2.16}$$

Elimination of $F_1(u, z)$ from these equations yields

$$G(x, v)F_1(y, z) = G(x, y) \tag{2.17}$$

where we use the notation $G(x, y) \equiv F_2(x, y)/F_1(x, y)$. Evidently, the left-hand side of (2.17) must be independent of z. Now, (2.17) can be written equally well as

$$G(x, v)F_2(y, z) = G(x, y)G(y, z), \tag{2.18}$$

and denoting the left-hand sides of (2.17), (2.18) by U, V respectively, we verify that $\partial V/\partial y = \partial U/\partial z$. Thus, $G(x, y)G(y, z)$ must be independent of y. The most general function $G(x, y)$ with this property is

$$G(x, y) = r\frac{H(x)}{H(y)} \tag{2.19}$$

where r is a constant and the function $H(x)$ is arbitrary. In the present case, $G > 0$ by monotonicity of F, and so we require that $r > 0$, and $H(x)$ may not change sign in the region of interest. Using (2.19), (2.17) and (2.18) become

$$F_1(y, z) = \frac{H(v)}{H(y)} \tag{2.20}$$

$$F_2(y, z) = r\frac{H(v)}{H(z)} \tag{2.21}$$

and the relation $\mathrm{d}v = \mathrm{d}F(y, z) = F_1\mathrm{d}y + F_2\mathrm{d}z$ takes the form

$$\frac{\mathrm{d}v}{H(v)} = \frac{\mathrm{d}y}{H(y)} + r\frac{\mathrm{d}z}{H(z)} \tag{2.22}$$

or, on integration,

$$w[F(y, z)] = w(v) = w(y)w^r(z), \tag{2.23}$$

where

$$w(x) \equiv \exp\left\{\int^x \frac{\mathrm{d}x}{H(x)}\right\}. \tag{2.24}$$

The absence of a lower limit on the integral signifies an arbitrary multiplicative factor in w. But taking the function $w(\cdot)$ of (2.15) and applying (2.23), we obtain $w(x)w^r(v) = w(u)w^r(z)$; applying (2.23) again, our functional equation now reduces to

$$w(x)w^r(y)[w(z)]^{r^2} = w(x)w^r(y)w^r(z). \tag{2.25}$$

Thus we obtain a nontrivial solution only if $r = 1$, and our final result can be expressed in either of the two forms:

$$w[F(x, y)] = w(x)w(y) \tag{2.26}$$

or

$$F(x, y) = w^{-1}[w(x)w(y)]. \tag{2.27}$$

Associativity and commutativity of the logical product thus require that the relation sought must take the functional form

$$w(AB|C) = w(A|BC)w(B|C) = w(B|AC)w(A|C), \tag{2.28}$$

which we shall call henceforth the *product rule*. By its construction (2.24), $w(x)$ must be a positive continuous monotonic function, increasing or decreasing according to the sign of $H(x)$; at this stage it is otherwise arbitrary.

The result (2.28) has been derived as a necessary condition for consistency in the sense of desideratum (IIIa). Conversely, it is evident that (2.28) is also sufficient to ensure this consistency for any number of joint propositions. For example, there are an enormous number of different ways in which $(ABCDEFG|H)$ could be expanded by successive partitions in the manner of (2.12); but if (2.28) is satisfied, they will all yield the same result.

The requirements of qualitative correspondence with common sense impose further conditions on the function $w(x)$. For example, in the first given form of (2.28) suppose that A is certain, given C. Then in the 'logical environment' produced by knowledge of C, the propositions AB and B are the same, in the sense that one is true if and only if the other is true. By our most primitive axiom of all, discussed in Chapter 1, propositions with the same truth value must have equal plausibility:

$$AB|C = B|C, \tag{2.29}$$

and also we will have

$$A|BC = A|C \tag{2.30}$$

because if A is already certain given C (i.e. C implies A), then, given any other information B which does not contradict C, it is still certain. In this case, (2.28) reduces to

$$w(B|C) = w(A|C)w(B|C), \tag{2.31}$$

and this must hold no matter how plausible or implausible B is to the robot. So our function $w(x)$ must have the property that

$$\text{certainty is represented by } w(A|C) = 1. \tag{2.32}$$

Now suppose that A is impossible, given C. Then the proposition AB is also impossible given C:

$$AB|C = A|C, \tag{2.33}$$

and if A is already impossible given C (i.e. C implies $\overline{A}$), then, given any further information B which does not contradict C, A would still be impossible:

$$A|BC = A|C. \tag{2.34}$$

In this case, (2.28) reduces to

$$w(A|C) = w(A|C)w(B|C), \tag{2.35}$$

and again this equation must hold no matter what plausibility B might have. There are only two possible values of $w(A|C)$ that could satisfy this condition: it could be zero or $+\infty$ (the choice $-\infty$ is ruled out because then by continuity $w(B|C)$ would have to be capable of negative values; (2.35) would then be a contradiction).

In summary, qualitative correspondence with common sense requires that $w(x)$ be a positive continuous monotonic function. It may be either increasing or decreasing. If it is increasing, it must range from zero for impossibility up to one for certainty. If it is decreasing, it must range from ∞ for impossibility down to one for certainty. Thus far, our conditions say nothing at all about how it varies between these limits.

However, these two possibilities of representation are not different in content. Given any function $w_1(x)$ which is acceptable by the above criteria and represents impossibility by ∞, we can define a new function $w_2(x) \equiv 1/w_1(x)$, which will be equally acceptable and represents impossibility by zero. Therefore, there will be no loss of generality if we now adopt the choice $0 \leq w(x) \leq 1$ as a *convention*; that is, as far as content is concerned, all possibilities consistent with our desiderata are included in this form. (As the reader may check, we could just as well have chosen the opposite convention; and the entire development of the theory from this point on, including all its applications, would go through equally well, with equations of a less familiar form but exactly the same content.)

2.2 The sum rule

Since the propositions now being considered are of the Aristotelian logical type which must be either true or false, the logical product $A\overline{A}$ is always false, the logical sum $A + \overline{A}$ always true. The plausibility that A is false must depend in some way on the plausibility that it is true. If we define $u \equiv w(A|B), \quad v \equiv w(\overline{A}|B)$, there must exist some functional relation

$$v = S(u). \tag{2.36}$$

Evidently, qualitative correspondence with common sense requires that $S(u)$ be a continuous monotonic decreasing function in $0 \leq u \leq 1$, with extreme values $S(0) = 1, \quad S(1) = 0$. But it cannot be just any function with these properties, for it must be consistent with the fact that the product rule can be written for either AB or $A\overline{B}$:

$$w(AB|C) = w(A|C)w(B|AC) \tag{2.37}$$

$$w(A\overline{B}|C) = w(A|C)w(\overline{B}|AC). \tag{2.38}$$

Thus, using (2.36) and (2.38), Eq. (2.37) becomes

$$w(AB|C) = w(A|C)S[w(\overline{B}|AC)] = w(A|C)S\left[\frac{w(A\overline{B}|C)}{w(A|C)}\right]. \tag{2.39}$$

Again, we invoke commutativity: $w(AB|C)$ is symmetric in A, B, and so consistency requires that

$$w(A|C)S\left[\frac{w(A\overline{B}|C)}{w(A|C)}\right] = w(B|C)S\left[\frac{w(B\overline{A}|C)}{w(B|C)}\right]. \tag{2.40}$$

This must hold for all propositions A, B, C; in particular, (2.40) must hold when

$$\overline{B} = AD, \tag{2.41}$$

where D is any new proposition. But then we have the truth values noted before in (1.13):

$$A\overline{B} = \overline{B}, \qquad\qquad B\overline{A} = \overline{A}, \tag{2.42}$$

and in (2.40) we may write

$$\begin{aligned} w(A\overline{B}|C) &= w(\overline{B}|C) = S[w(B|C)] \\ w(B\overline{A}|C) &= w(\overline{A}|C) = S[w(A|C)]. \end{aligned} \tag{2.43}$$

Therefore, using the abbreviations

$$x \equiv w(A|C), \qquad\qquad y \equiv w(B|C), \tag{2.44}$$

(2.25) becomes a functional equation

$$xS\left[\frac{S(y)}{x}\right] = yS\left[\frac{S(x)}{y}\right], \quad \begin{aligned} 0 \le\, &S(y) \le x \\ 0 \le\, &x \le 1 \end{aligned} \tag{2.45}$$

which expresses a scaling property that $S(x)$ must have in order to be consistent with the product rule. In the special case $y = 1$, this reduces to

$$S[S(x)] = x, \tag{2.46}$$

which states that $S(x)$ is a self-reciprocal function; $S(x) = S^{-1}(x)$. Thus, from (2.36) it follows also that $u = S(v)$. But this expresses only the evident fact that the relationship between A and $\overline{A}$ is a reciprocal one; it does not matter which proposition we denote by the simple letter, which by the barred letter. We noted this before in (1.8); if it had not been obvious before, we should be obliged to recognize it at this point.

The domain of validity given in (2.45) is found as follows. The proposition D is arbitrary, and so by various choices of D we can achieve all values of $w(D|AC)$ in

$$0 \le w(D|AC) \le 1. \tag{2.47}$$

But $S(y) = w(AD|C) = w(A|C)w(D|AC)$, and so (2.47) is just $(0 \le S(y) \le x)$, as stated in (2.45). This domain is symmetric in x, y; it can be written equally well with them interchanged. Geometrically, it consists of all points in the xy plane lying in the unit square $(0 \le x, y \le 1)$ and on or above the curve $y = S(x)$.

Indeed, the shape of that curve is determined already by what (2.45) says for points lying infinitesimally above it. For if we set $y = S(x) + \epsilon$, then as $\epsilon \to 0^+$ two terms in (2.45) tend to $S(1) = 0$, but at different rates. Therefore everything depends on the exact way in which $S(1-\delta)$ tends to zero as $\delta \to 0$. To investigate this, we define a new variable $q(x, y)$ by

$$\frac{S(x)}{y} = 1 - \exp\{-q\}. \tag{2.48}$$

Then we may choose $\delta = \exp\{-q\}$, define the function $J(q)$ by

$$S(1-\delta) = S(1 - \exp\{-q\} = \exp\{-J(q)\}, \tag{2.49}$$

and find the asymptotic form of $J(q)$ as $q \to \infty$.

Considering now x, q as the independent variables, we have from (2.48)

$$S(y) = S[S(x)] + \exp\{-q\}S(x)S'[S(x)] + O(\exp\{-2q\}). \tag{2.50}$$

Using (2.46) and its derivative $S'[S(x)]S'(x) = 1$, this reduces to

$$\frac{S(y)}{x} = 1 - \exp\{-(\alpha + q)\} + O(\exp\{-2q\}), \tag{2.51}$$

where

$$\alpha(x) \equiv \log\left[\frac{-xS'(x)}{S(x)}\right] > 0. \tag{2.52}$$

With these substitutions, our functional equation (2.45) becomes

$$J(q+\alpha) - J(q) = \log\left[\frac{x}{S(x)}\right] + \log(1 - \exp\{-q\}) + O(\exp\{-2q\}), \qquad \begin{array}{l} 0 < q < \infty \\ 0 < x \le 1 \end{array} \tag{2.53}$$

As $q \to \infty$ the last two terms go to zero exponentially fast, so $J(q)$ must be asymptotically linear,

$$J(q) \sim a + bq + O(\exp\{-q\}), \tag{2.54}$$

with positive slope

$$b = \alpha^{-1} \log\left[\frac{x}{S(x)}\right]. \tag{2.55}$$

In (2.54) there is no periodic term with period α, because (2.53) must hold for a continuum of different values of x, and therefore for a continuum of values of $\alpha(x)$. But, by definition, J is a function of q only, so the right-hand side of (2.55) must be independent of x. This gives, using (2.52),

$$\frac{x}{S(x)} = \left[\frac{-xS'(x)}{S(x)}\right]^b, \qquad 0 < b < \infty, \tag{2.56}$$

or, rearranging, $S(x)$ must satisfy the differential equation

$$S^{m-1}\mathrm{d}S + x^{m-1}\mathrm{d}x = 0, \tag{2.57}$$

where $m \equiv 1/b$ is some positive constant. The only solution of this satisfying $S(0) = 1$ is

$$S(x) = (1 - x^m)^{1/m}, \qquad \begin{array}{l} 0 \le x \le 1 \\ 0 < m < \infty \end{array} \tag{2.58}$$

and, conversely, we verify at once that (2.58) is a solution of (2.45).

The result (2.58) was first derived by R. T. Cox (1946) by a different argument which assumed $S(x)$ twice differentiable. Again, Aczél (1966) derives the same result without assuming differentiability. (But to assume differentiability in the present application seems to us a very innocuous step, for if the functional equations had led us to nondifferentiable functions, we would have rejected this whole theory as a qualitative violation of common sense.) In any event, (2.58) is the most general function satisfying the functional equation (2.45) and the left boundary condition $S(0) = 1$; whereupon we are encouraged to find that it automatically satisfies the right boundary condition $S(1) = 0$.

Since our derivation of the functional equation (2.45) used the special choice (2.41) for B, we have shown thus far only that (2.58) is a necessary condition to satisfy the general consistency requirement (2.40). To check its sufficiency, substitute (2.58) into (2.40). We obtain

$$w^m(A|C) - w^m(A\overline{B}|C) = w^m(B|C) - w^m(B\overline{A}|C), \tag{2.59}$$

a trivial identity by virtue of (2.28) and (2.38). Therefore, (2.58) is the necessary and sufficient condition on $S(x)$ for consistency in the sense (2.40).

Our results up to this point can be summarized as follows. Associativity of the logical product requires that some monotonic function $w(x)$ of the plausibility $x = A|B$ must obey the product rule (2.28). Our result (2.58) states that this same function must also obey a sum rule:

$$w^m(A|B) + w^m(\overline{A}|B) = 1 \tag{2.60}$$

for some positive m. Of course, the product rule itself can be written equally well as

$$w^m(AB|C) = w^m(A|C)w^m(B|AC) = w^m(B|C)w^m(A|BC), \tag{2.61}$$

but then we see that the value of m is actually irrelevant; for whatever value is chosen, we can define a new function

$$p(x) \equiv w^m(x), \tag{2.62}$$

and our rules take the form

$$p(AB|C) = p(A|C)p(B|AC) = p(B|C)p(A|BC), \tag{2.63}$$

$$p(A|B) + p(\overline{A}|B) = 1. \tag{2.64}$$

In fact, this entails no loss of generality, for the only requirement we have imposed on the function $w(x)$ is that it is a continuous monotonic increasing function ranging from $w = 0$ for impossibility to $w = 1$ for certainty. But if $w(x)$ satisfies this, then so also does $w^m(x)$, $0 < m < \infty$. Therefore, to say that we could use different values of m does not give us any freedom that we did not have already in the arbitrariness of $w(x)$. All possibilities allowed by our desiderata are contained in (2.63) and (2.64), in which $p(x)$ is any continuous monotonic increasing function with the range $0 \leq p(x) \leq 1$.

Are further relations needed to yield a complete set of rules for plausible inference, adequate to determine the plausibility of any logic function $f(A_1, \ldots, A_n)$ from those of $\{A_1, \ldots, A_n\}$? We have, in the product rule (2.63) and sum rule (2.64), formulas for the plausibility of the conjunction AB and the negation $\overline{A}$. However, we have noted, in the discussion following (1.23), that conjunction and negation are an adequate set of operations, from which all logic functions can be constructed.

Therefore, one would conjecture that our search for basic rules should be finished; it ought to be possible, by repeated applications of the product rule and sum rule, to arrive at the plausibility of any proposition in the Boolean algebra generated by $\{A_1, \ldots, A_n\}$.

To verify this, we seek first a formula for the logical sum $A + B$. Applying the product rule and sum rule repeatedly, we have

$$\begin{aligned} p(A + B|C) &= 1 - p(\overline{A}\,\overline{B}|C) = 1 - p(\overline{A}|C)p(\overline{B}|\overline{A}C) \\ &= 1 - p(\overline{A}|C)[1 - p(B|\overline{A}C)] = p(A|C) + p(\overline{A}B|C) \\ &= p(A|C) + p(B|C)p(\overline{A}|BC) = p(A|C) + p(B|C)[1 - p(A|BC)] \end{aligned} \tag{2.65}$$

and finally

$$p(A + B|C) = p(A|C) + p(B|C) - p(AB|C). \tag{2.66}$$

This generalized sum rule is one of the most useful in applications. Evidently, the primitive sum rule (2.64) is a special case of (2.66), with the choice $B = \overline{A}$.

Exercise 2.1. Is it possible to find a general formula for $p(C|A + B)$, analogous to (2.66), from the product and sum rules? If so, derive it; if not, explain why this cannot be done.

Exercise 2.2. Now suppose we have a set of propositions $\{A_1, \ldots, A_n\}$ which on information X are mutually exclusive: $p(A_i A_j|X) = p(A_i|X)\,\delta_{ij}$. Show that $p(C|(A_1 + A_2 + \cdots + A_n X)$ is a weighted average of the separate plausibilities $p(C|A_i X)$:

$$p(C|(A_1 + \cdots + A_n X) = p(C|A_1 X + A_2 X + \cdots + A_n X) = \frac{\sum_i p(A_i|X)\, p(C|A_i X)}{\sum_i p(A_i|X)}. \tag{2.67}$$

To extend the result (2.66), we noted following (1.17) that any logic function other than the trivial contradiction can be expressed in disjunctive normal form, as a logical sum of the basic conjunctions such as (1.17). Now the plausibility of any one of the basic conjunctions

$\{Q_i,\ 1 \le i \le 2^n\}$ is determined by repeated applications of the product rule; and then repeated application of (2.66) will yield the plausibility of any logical sum of the Q_i. In fact, these conjunctions are mutually exclusive, so we shall find (see (2.85) below) that this reduces to a simple sum $\sum_i p(Q_i|C)$ of at most $(2^n - 1)$ terms.

So, just as conjunction and negation are an adequate set of operations for deductive logic, the above product and sum rules are an adequate set for plausible inference, in the following sense. Whenever the background information is enough to determine the plausibilities of the basic conjunctions, our rules are adequate to determine the plausibility of every proposition in the Boolean algebra generated by $\{A_1, \ldots, A_n\}$. Thus, in the case $n = 4$ we need the plausibilities of $2^4 = 16$ basic conjunctions, whereupon our rules will determine the plausibility of each of the $2^{16} = 65\,536$ propositions in the Boolean algebra.

But this is almost always more than we need in a real application; if the background information is enough to determine the plausibility of a few of the basic conjunctions, this may be adequate for the small part of the Boolean algebra that is of concern to us.

2.3 Qualitative properties

Now let us check to see how the theory based on (2.63) and (2.64) is related to the theory of deductive logic and the various qualitative syllogisms from which we started in Chapter 1. In the first place it is obvious that in the limit as $p(A|B) \to 0$ or $p(A|B) \to 1$, the sum rule (2.64) expresses the primitive postulate of Aristotelian logic: if A is true, then $\overline{A}$ must be false, etc.

Indeed, all of that logic consists of the two strong syllogisms (1.1), (1.2) and all that follows from them; using now the implication sign (1.14) to state the major premise:

$$\begin{array}{cc} A \Rightarrow B & A \Rightarrow B \\ A \text{ is true} & B \text{ is false} \\ \hline B \text{ is true} & A \text{ is false} \end{array} \tag{2.68}$$

and the endless stream of their consequences. If we let C stand for their major premise:

$$C \equiv A \Rightarrow B \tag{2.69}$$

then these syllogisms correspond to our product rule (2.63) in the forms

$$p(B|AC) = \frac{p(AB|C)}{p(A|C)}, \qquad p(A|\overline{B}C) = \frac{p(A\overline{B}|C)}{p(\overline{B}|C)}, \tag{2.70}$$

respectively. But from (2.68) we have $p(AB|C) = p(A|C)$ and $p(A\overline{B}|C) = 0$, and so (2.70) reduces to

$$p(B|AC) = 1, \qquad p(A|\overline{B}C) = 0, \tag{2.71}$$

as stated in the syllogisms (2.68). Thus the relation is simply: *Aristotelian deductive logic is the limiting form of our rules for plausible reasoning, as the robot becomes more and more certain of its conclusions.*

But our rules have also what is not contained in deductive logic: a quantitative form of the weak syllogisms (1.3) and (1.4). To show that those original qualitative statements always follow from the present rules, note that the first weak syllogism

$$\begin{array}{c} A \Rightarrow B \\[6pt] B \text{ is true} \\ \hline \text{therefore, } A \text{ becomes more plausible} \end{array} \tag{2.72}$$

corresponds to the product rule (2.63) in the form

$$p(A|BC) = p(A|C)\frac{p(B|AC)}{p(B|C)}. \tag{2.73}$$

But from (2.68), $p(B|AC) = 1$, and since $p(B|C) \le 1$, (2.73) gives

$$p(A|BC) \ge p(A|C), \tag{2.74}$$

as stated in the syllogism. Likewise, the syllogism (1.4)

$$\begin{array}{c} A \Rightarrow B \\[6pt] A \text{ is false} \\ \hline \text{therefore, } B \text{ becomes less plausible} \end{array} \tag{2.75}$$

corresponds to the product rule in the form

$$p(B|\overline{A}C) = p(B|C)\frac{p(\overline{A}|BC)}{p(\overline{A}|C)}. \tag{2.76}$$

But from (2.74) it follows that $p(\overline{A}|BC) \le p(\overline{A}|C)$; and so (2.76) gives

$$p(B|\overline{A}C) \le p(B|C), \tag{2.77}$$

as stated in the syllogism.

Finally, the policeman's syllogism (1.5), which seemed very weak when stated abstractly, is also contained in our product rule, stated in the form (2.73). Letting C now stand for the background information (not noted explicitly in (1.5) because the need for it was not yet apparent), the major premise, 'If A is true, then B becomes more plausible', now takes the form

$$p(B|AC) > p(B|C), \tag{2.78}$$

and (2.73) gives at once

$$p(A|BC) > p(A|C), \tag{2.79}$$

as stated in the syllogism.

Now we have more than the mere qualitative statement (2.79). In Chapter 1 we wondered, without answering: What determines whether the evidence B elevates A almost to certainty, or has a negligible effect on its plausibility? The answer from (2.73) is that, since $p(B|AC)$

cannot be greater than unity, a large increase in the plausibility of A can occur only when $p(B|C)$ is very small. Observing the gentleman's behavior (B) makes his guilt (A) seem virtually certain, because that behavior is otherwise so very unlikely on the background information; no policeman has ever seen an innocent person behaving that way. On the other hand, if knowing that A is true can make only a negligible increase in the plausibility of B, then observing B can in turn make only a negligible increase in the plausibility of A.

We could give many more comparisons of this type; indeed, the complete qualitative correspondence of these rules with common sense has been noted and demonstrated by many writers, including Keynes (1921), Jeffreys (1939), Pólya (1945, 1954), R. T. Cox (1961), Tribus (1969), de Finetti (1974a,b), and Rosenkrantz (1977). The treatment of Pólya was described briefly in our Preface and Chapter 1, and we have just recounted that of Cox more fully. However, our aim now is to push ahead to quantitative applications; so we return to the basic development of the theory.

2.4 Numerical values

We have found so far the most general consistent rules by which our robot can manipulate plausibilities, granted that it must associate them with real numbers, so that its brain can operate by the carrying out of some definite physical process. While we are encouraged by the familiar formal appearance of these rules and their qualitative properties just noted, two evident circumstances show that our job of designing the robot's brain is not yet finished.

In the first place, while the rules (2.63), (2.64) place some limitations on how plausibilities of different propositions must be related to each other, it would appear that we have not yet found any *unique* rules, but rather an infinite number of possible rules by which our robot can do plausible reasoning. Corresponding to every different choice of a monotonic function $p(x)$, there seems to be a different set of rules, with different content.

Secondly, nothing given so far tells us what actual numerical values of plausibility should be assigned at the beginning of a problem, so that the robot can get started on its calculations. How is the robot to make its initial encoding of the background information into definite numerical values of plausibilities? For this we must invoke the 'interface' desiderata (IIIb), (IIIc) of (1.39), not yet used.

The following analysis answers both of these questions, in a way both interesting and unexpected. Let us ask for the plausibility $(A_1 + A_2 + A_3|B)$ that at least one of three propositions $\{A_1, A_2, A_3\}$ is true. We can find this by two applications of the extended sum rule (2.66), as follows. The first application gives

$$p(A_1 + A_2 + A_3|B) = p(A_1 + A_2|B) + p(A_3|B) - p(A_1A_3 + A_2A_3|B) \qquad (2.80)$$

where we first considered $(A_1 + A_2)$ as a single proposition, and used the logical relation

$$(A_1 + A_2)A_3 = A_1A_3 + A_2A_3. \qquad (2.81)$$

Applying (2.66) again, we obtain seven terms which can be grouped as follows:

$$\begin{aligned} p(A_1 + A_2 + A_3|B) = {} & p(A_1|B) + p(A_2|B) + p(A_3|B) \\ & - p(A_1A_2|B) - p(A_2A_3|B) - p(A_3A_1|B) \\ & + p(A_1A_2A_3|B). \end{aligned} \tag{2.82}$$

Now suppose these propositions are mutually exclusive; i.e. the evidence B implies that no two of them can be true simultaneously:

$$p(A_iA_j|B) = p(A_i|B)\delta_{ij}. \tag{2.83}$$

Then the last four terms of (2.82) vanish, and we have

$$p(A_1 + A_2 + A_3|B) = p(A_1|B) + P(A_2|B) + P(A_3|B). \tag{2.84}$$

Adding more propositions A_4, A_5, etc., it is easy to show by induction that if we have n mutually exclusive propositions $\{A_1, \ldots, A_n\}$, (2.84) generalizes to

$$p(A_1 + \cdots + A_m|B) = \sum_{i=1}^{m} p(A_i|B), \qquad 1 \le m \le n, \tag{2.85}$$

a rule which we will be using constantly from now on.

In conventional expositions, Eq. (2.85) is usually introduced first as the basic but, as far as one can see, arbitrary axiom of the theory. The present approach shows that this rule is deducible from simple qualitative conditions of consistency. The viewpoint which sees (2.85) as the primitive, fundamental relation is one which we are particularly anxious to avoid (see Comments section at the end of this chapter).

Now suppose that the propositions $\{A_1, \ldots, A_n\}$ are not only mutually exclusive but also exhaustive; i.e. the background information B stipulates that one and only one of them must be true. In that case, the sum (2.85) for $m = n$ must be unity:

$$\sum_{i=1}^{n} p(A_i|B) = 1. \tag{2.86}$$

This alone is not enough to determine the individual numerical values $p(A_i|B)$. Depending on further details of the information B, many different choices might be appropriate, and in general finding the $p(A_i|B)$ by logical analysis of B can be a difficult problem. It is, in fact, an open-ended problem, since there is no end to the variety of complicated information that might be contained in B; and therefore no end to the complicated mathematical problems of translating that information into numerical values of $p(A_i|B)$. As we shall see, this is one of the most important current research problems; every new principle we can discover for translating information B into numerical values of $p(A_i|B)$ will open up a new class of useful applications of this theory.

There is, however, one case in which the answer is particularly simple, requiring only direct application of principles already given. But we are entering now into a very delicate area, a cause of confusion and controversy for over a century. In the early stages of this theory, as in elementary geometry, our intuition runs so far ahead of logical analysis that the point of the logical analysis is often missed. The trouble is that intuition leads us to the same final conclusions far more quickly, but without any correct appreciation of their range of validity. The result has been that the development of this theory has been retarded for some 150 years because various workers have insisted on debating these issues on the basis, not of demonstrative arguments, but of their conflicting intuitions.

At this point, therefore, we must ask the reader to suppress all intuitive feelings you may have, and allow yourself to be guided solely by the following logical analysis. The point we are about to make cannot be developed too carefully; and, unless it is clearly understood, we will be faced with tremendous conceptual difficulties from here on.

Consider two different problems. Problem I is the one just formulated: we have a given set of mutually exclusive and exhaustive propositions $\{A_1, \dots, A_n\}$ and we seek to evaluate $p(A_i|B)_I$. Problem II differs in that the labels A_1, A_2 of the first two propositions have been interchanged. These labels are, of course, entirely arbitrary; it makes no difference which proposition we choose to call A_1 and which A_2. In Problem II, therefore, we also have a set of mutually exclusive and exhaustive propositions $\{A'_1, \dots, A'_n\}$, given by

$$\begin{aligned} A'_1 &\equiv A_2, \\ A'_2 &\equiv A_1, \\ A'_k &\equiv A_k, \quad 3 \le k \le n, \end{aligned} \tag{2.87}$$

and we seek to evaluate the quantities $p(A'_i|B)_{II}, \; i = 1, 2, \dots, n$.

In interchanging the labels, we have generated a different but closely related problem. It is clear that, whatever state of knowledge the robot had about A_1 in Problem I, it must have the same state of knowledge about A'_2 in Problem II, for they are the same proposition, the given information B is the same in both problems, and it is contemplating the same totality of propositions $\{A_1, \dots, A_n\}$ in both problems. Therefore we must have

$$p(A_1|B)_I = p(A'_2|B)_{II}, \tag{2.88}$$

and similarly

$$p(A_2|B)_I = p(A'_1|B)_{II}. \tag{2.89}$$

We will call these the *transformation equations*. They describe only how the two problems are related to each other, and therefore they must hold whatever the information B might be; in particular, however plausible or implausible the propositions A_1, A_2 might seem to the robot in Problem I.

Now suppose that information B is indifferent between propositions A_1 and A_2; i.e. if it says something about one, it says the same thing about the other, and so it contains nothing that would give the robot any reason to prefer either one over the other. In this case, Problems I and II are not merely related, but entirely equivalent; i.e. the robot is in exactly the same state of knowledge about the set of propositions $\{A'_1, \dots, A'_n\}$ in Problem II, *including their labeling*, as it is about the set $\{A_1, \dots, A_n\}$ in Problem I.

Now we invoke our desideratum of consistency in the sense (IIIc) in (1.39). This stated that equivalent states of knowledge must be represented by equivalent plausibility assignments.

In equations, this statement is

$$p(A_i|B)_I = p(A'_i|B)_{II}, \qquad i = 1, 2, \ldots, n, \tag{2.90}$$

which we shall call the *symmetry equations*. But now, combining (2.88), (2.89), and (2.90), we obtain

$$p(A_1|B)_I = p(A_2|B)_I. \tag{2.91}$$

In other words, propositions A_1 and A_2 must be assigned equal plausibilities in Problem I (and, of course, also in Problem II).

At this point, depending on your personality and background in this subject, you will be either greatly impressed or greatly disappointed by the result (2.91). The argument we have just given is the first 'baby' version of the group invariance principle for assigning plausibilities; it will be extended greatly in Chapter 6, when we consider the general problem of assigning 'noninformative priors'.

More generally, let $\{A''_1, \ldots, A''_n\}$ be any permutation of $\{A_1, \ldots, A_n\}$ and let Problem III be that of determining the $p(A''_i|B)$. If the permutation is such that $A''_k \equiv A_i$, there will be n transformation equations of the form

$$p(A_i|B)_I = p(A''_k|B)_{III} \tag{2.92}$$

which show how Problems I and III are related to each other; these relations will hold whatever the given information B.

But if information B is now indifferent between all the propositions A_i, then the robot is in exactly the same state of knowledge about the set of propositions $\{A''_1, \ldots, A''_n\}$ in Problem III as it was about the set $\{A_1, \ldots, A_n\}$ in Problem I; and again our desideratum of consistency demands that it assign equivalent plausibilities in equivalent states of knowledge, leading to the n symmetry conditions

$$p(A_k|B)_I = p(A''_k|B)_{III}, \qquad k = 1, 2, \ldots, n. \tag{2.93}$$

From (2.92) and (2.93) we obtain n equations of the form

$$p(A_i|B)_I = p(A_k|B)_I. \tag{2.94}$$

Now, these relations must hold whatever the particular permutation we used to define Problem III. There are $n!$ such permutations, and so there are actually $n!$ equivalent problems among which, for given i, the index k will range over all of the $(n - 1)$ others in (2.94). Therefore, the only possibility is that all of the $p(A_i|B)_I$ be equal (indeed, this is required already by consideration of a single permutation if it is cyclic of order n). Since the $\{A_1, \ldots, A_n\}$ are exhaustive, (2.86) will hold, and the only possibility is therefore

$$p(A_i|B)_I = \frac{1}{n}, \qquad (1 \le i \le n), \tag{2.95}$$

and we have finally arrived at a set of definite numerical values! Following Keynes (1921), we shall call this result the *principle of indifference*.

Perhaps, in spite of our admonitions, the reader's intuition had already led to just this conclusion, without any need for the rather tortuous reasoning we have just been through. If so, then at least that intuition is consistent with our desiderata. But merely writing down (2.95) intuitively gives one no appreciation of the importance and uniqueness of this result. To see the uniqueness, note that if the robot were to assign any values different from (2.95), then by a mere permutation of labels we could exhibit a second problem in which the robot's state of knowledge is the same, but in which it is assigning different plausibilities.

To see the importance, note that (2.95) actually answers both of the questions posed at the beginning of this section. It shows – in one particular case which can be greatly generalized – how the information given the robot can lead to definite numerical values, so that a calculation can start. But it also shows something even more important because it is not at all obvious intuitively; the information given the robot determines the numerical values of the quantities $p(x) = p(A_i|B)$, and not the numerical values of the plausibilities $x = A_i|B$ from which we started. This, also, will be found to be true in general.

Recognizing this gives us a beautiful answer to the first question posed at the beginning of this section; after having found the product and sum rules, it still appeared that we had not found any unique rules of reasoning, because every different choice of a monotonic function $p(x)$ would lead to a different set of rules (i.e. a set with different content). But now we see that no matter what function $p(x)$ we choose, we shall be led to the same result (2.95), and the same numerical value of p. Furthermore, the robot's reasoning processes can be carried out entirely by manipulation of the quantities p, as the product and sum rules show; and the robot's final conclusions can be stated equally well in terms of the p's instead of the x's.

So, we now see that different choices of the function $p(x)$ correspond only to different ways we could design the robot's internal memory circuits. For each proposition A_i about which it is to reason, it will need a memory address in which it stores some number representing the degree of plausibility of A_i, on the basis of all the data it has been given. Of course, instead of storing the number p_i it could equally well store any strict monotonic function of p_i. But no matter what function it used internally, the externally observable behavior of the robot would be just the same.

As soon as we recognize this, it is clear that, instead of saying that $p(x)$ is an arbitrary monotonic function of x, it is much more to the point to turn this around and say that:

The plausibility $x \equiv A|B$ is an arbitrary monotonic function of p,
defined in $(0 \leq p \leq 1)$.

It is p that is rigidly fixed by the data, not x.

The question of uniqueness is therefore disposed of automatically by the result (2.95); in spite of first appearances, there is actually only one consistent set of rules by which our robot can do plausible reasoning, and, for all practical purposes, the plausibilities $x \equiv A|B$ from which we started have faded entirely out of the picture! We will just have no further use for them.

Having seen that our theory of plausible reasoning can be carried out entirely in terms of the quantities p, we finally introduce their technical names; from now on, we will call these quantities *probabilities*. The word 'probability' has been studiously avoided up to this point, because, whereas the word does have a colloquial meaning to the proverbial 'man on the street', it is for us a technical term, which ought to have a precise meaning. But until it had been demonstrated that these quantities are uniquely determined by the data of a problem, we had no grounds for supposing that the quantities p were possessed of any precise meaning.

We now see that the quantities p define a particular scale on which degrees of plausibility can be measured. Out of all possible monotonic functions which could, in principle, serve this purpose equally well, we choose this particular one, not because it is more 'correct', but because it is more convenient; i.e. it is the quantities p that obey the simplest rules of combination, the product and sum rules. Because of this, numerical values of p are directly determined by our information.

This situation is analogous to that in thermodynamics, where out of all possible empirical temperature scales t, which are monotonic functions of each other, we finally decide to use the Kelvin scale T; not because it is more 'correct' than others but because it is more convenient; i.e. the laws of thermodynamics take their simplest form [$\mathrm{d}U = T\mathrm{d}S - P\mathrm{d}V$, $\mathrm{d}G = -S\mathrm{d}T + V\mathrm{d}P$, etc.] in terms of this particular scale. Because of this, numerical values of temperatures on the kelvin scale are 'rigidly fixed' in the sense of being directly measurable in experiments, independently of the properties of any particular substance like water or mercury.

Another rule, equally appealing to our intuition, follows at once from (2.95). Consider the traditional 'Bernoulli urn' of probability theory; ours is known to contain ten balls of identical size and weight, labeled $\{1, 2, \ldots, 10\}$. Three balls (numbers 4, 6, 7) are black, the other seven are white. We are to shake the urn and draw one ball blindfolded. The background information B in (2.95) consists of the statements in the last two sentences. What is the probability that we draw a black one?

Define the propositions: $A_i \equiv$ 'the ith ball is drawn', $(1 \leq i \leq 10)$. Since the background information is indifferent to these ten possibilities, (2.95) applies, and the robot assigns

$$p(A_i|B) = \frac{1}{10}, \quad 1 \leq i \leq 10. \tag{2.96}$$

The statement that we draw a black ball is that we draw number 4, 6, or 7;

$$p(\text{black}|B) = p(A_4 + A_6 + A_7|B). \tag{2.97}$$

But these are mutually exclusive propositions (i.e. they assert mutually exclusive events), so (2.85) applies, and the robot's conclusion is

$$p(\text{black}|B) = \frac{3}{10}, \tag{2.98}$$

as intuition had told us already. More generally, if there are N such balls, and the proposition A is defined to be true on any specified subset of M of them, $(0 \leq M \leq N)$, false on the rest, we have

$$p(A|B) = \frac{M}{N}. \tag{2.99}$$

This was the original mathematical *definition* of probability, as given by James Bernoulli (1713) and used by most writers for the next 150 years. For example, Laplace's great *Théorie Analytique des Probabilités* (1812) opens with this sentence:

> The Probability for an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.

Exercise 2.3. As soon as we have the numerical values $a = P(A|C)$ and $b = P(B|C)$, the product and sum rules place some limits on the possible numerical values for their conjunction and disjunction. Supposing that $a \leq b$, show that the probability for the conjunction cannot exceed that of the least probable proposition: $0 \leq P(AB|C) \leq a$, and the probability for the disjunction cannot be less than that of the most probable proposition: $b \leq P(A + B|C) \leq 1$. Then show that, if $a + b > 1$, there is a stronger inequality for the conjunction; and if $a + b < 1$ there is a stronger one for the disjunction. These necessary general inequalities are helpful in detecting errors in calculations.

2.5 Notation and finite-sets policy

Now we can introduce the notation to be used in the remainder of this work (discussed more fully in Appendix B). Henceforth, our formal probability Symbols, will use the capital P:

$$P(A|B), \tag{2.100}$$

which signifies that the arguments are *propositions*. Probabilities whose arguments are numerical values are generally denoted by other functional symbols, such as

$$f(r|np), \tag{2.101}$$

which denote ordinary mathematical functions. The reason for making this distinction is to avoid ambiguity in the meaning of our symbols, which has been a recent problem in this field. However, in agreement with the customary loose notation in the existing literature, we sometimes relax our standards enough to allow the probability symbols with small p: $p(x|y)$ or $p(A|B)$ or $p(x|B)$ to have arguments which can be either propositions or numerical values, in any mix. Thus the meaning of expressions with small p can be judged only from the surrounding context.

It is very important to note that our consistency theorems have been established only for probabilities assigned on *finite sets* of propositions. In principle, every problem must start with such finite-set probabilities; extension to infinite sets is permitted only when this is the result of a well-defined and well-behaved limiting process from a finite set. More generally, in any mathematical operations involving infinite sets, the safe procedure is the finite-sets policy:

> *Apply the ordinary processes of arithmetic and analysis only to expressions with a finite number of terms. Then, after the calculation is done, observe how the resulting finite expressions behave as the number of terms increases indefinitely.*

In laying down this rule of conduct, we are only following the policy that mathematicians from Archimedes to Gauss have considered clearly necessary for nonsense avoidance in all of mathematics. But, more recently, the popularity of infinite-set theory and measure theory have led some to disregard it and seek shortcuts which purport to use measure theory directly. Note, however, that this rule of conduct is consistent with the original Lebesgue definition of measure, and *when a well-behaved limit exists* it leads us automatically to correct 'measure theoretic' results. Indeed, this is how Lebesgue found his first results.

The danger is that the present measure theory notation presupposes the infinite limit already accomplished, but contains no symbol indicating which limiting process was used. Yet, as noted in our Preface, different limiting Processes – equally well-behaved – lead in general to different results. When there is no well-behaved limit, any attempt to go directly to the limit can result in nonsense, *the cause of which cannot be seen as long as one looks only at the limit, and not at the limiting process.*

This little 'sermon' is an introduction to Chapter 15 on infinite-set paradoxes, where we shall see some of the results that have been produced by those who ignored this rule of conduct, and tried to calculate probabilities directly on an infinite set without considering any limit from a finite set. The results are at best ambiguous, at worst nonsensical.

2.6 Comments

It has taken us two chapters of close reasoning to get back to the point (2.99) from which Laplace started some 180 years ago. We shall try to understand the intervening period, as a weird episode of history, throughout the rest of the present work. The story is so complicated that we can unfold it only gradually, over the next ten chapters. To make a start on this, let us consider some of the questions often raised about the use of probability theory as an extension of logic.

2.6.1 'Subjective' vs. 'objective'

These words are abused so much in probability theory that we try to clarify our use of them. In the theory we are developing, any probability assignment is necessarily 'subjective' in the sense that it describes only a state of knowledge, and not anything that could be measured in a physical experiment. Inevitably, someone will demand to know: '*Whose* state of knowledge?' The answer is always: 'That of the robot – or of anyone else who is given the same information and reasons according to the desiderata used in our derivations in this chapter.'

Anyone who has the same information, but comes to a different conclusion than our robot, is necessarily violating one of those desiderata. While nobody has the authority to forbid such violations, it appears to us that a rational person, should he discover that he was violating one of them, would wish to revise his thinking (in any event, he would surely have difficulty in persuading anyone else, who was aware of that violation, to accept his conclusions).

Now, it was just the function of our interface desiderata (IIIb), (IIIc) to make these probability assignments completely 'objective' in the sense that they are independent of the personality of the user. They are a means of describing (or, what is the same thing, of encoding) the *information* given in the statement of a problem, independently of whatever personal feelings (hopes, fears, value judgments, etc.) you or I might have about the propositions involved. It is 'objectivity' in this sense that is needed for a scientifically respectable theory of inference.

2.6.2 Gödel's theorem

To answer another inevitable question, we recapitulate just what has and what has not been proved in this chapter. The main constructive requirement which determined our product and sum rules was the desideratum (IIIa) of 'structural consistency'. Of course, this does not mean that our rules have been proved consistent; it means only that any other rules which represent degrees of plausibility by real numbers, but which differ in content from ours, will lead necessarily either to inconsistencies or violations of our other desiderata.

A famous theorem of Kurt Gödel (1931) states that no mathematical system can provide a proof of its own consistency. Does this prevent us from ever proving the consistency of probability theory as logic? We are not prepared to answer this fully, but perhaps we can clarify the situation a little.

Firstly, let us be sure that 'inconsistency' means the same thing to us and to a logician. What we had in mind was that if our rules were inconsistent, then it would be possible to derive contradictory results from valid application of them; for example, by applying the rules in two equally valid ways, one might be able to derive both $P(A|BC) = 1/3$ and $P(A|BC) = 2/3$. Cox's functional equations sought to guard against this. Now, when a logician says that a system of axioms $\{A_1, A_2, \dots, A_n\}$ is inconsistent, he means that a contradiction can be deduced from them; i.e. some proposition Q and its denial $\overline{Q}$ are both deducible. Indeed, this is not really different from our meaning.

To understand the above Gödel result, the essential point is the principle of elementary logic that a contradiction $\overline{A}\,A$ implies all propositions, true and false. (Given any two propositions A and B, we have $A \Rightarrow (A + B)$, therefore $\overline{A}\,A \Rightarrow \overline{A}(A + B) = \overline{A}\,A + \overline{A}\,B \Rightarrow B$.) Then let $A = \{A_1, A_2, \dots, A_n\}$ be the system of axioms underlying a mathematical theory and T any proposition, or theorem, deducible from them:[1]

$$A \Rightarrow T. \tag{2.102}$$

[1] In Chapter 1 we noted the tricky distinction between the weak property of formal implication and the strong one of logical deducibility; by 'implications of a proposition C' we really mean 'propositions logically deducible from *C and the totality of other background information*'. Conventional expositions of Aristotelian logic are, in our view, flawed by their failure to make explicit mention of background information, which is usually essential to our reasoning, whether inductive or deductive. But, in the present argument, we can understand A as including all the propositions that constitute that background information; then 'implication' and 'logical deducibility' are the same thing.

Now, whatever T may assert, the fact that T can be deduced from the axioms cannot prove that there is no contradiction in them, since, if there were a contradiction, T could certainly be deduced from them!

This is the essence of the Gödel theorem, as it pertains to our problems. As noted by Fisher (1956), it shows us the intuitive reason why Gödel's result is true. We do not suppose that any logician would accept Fisher's simple argument as a proof of the full Gödel theorem; yet for most of us it is more convincing than Gödel's long and complicated proof.[2]

Now suppose that the axioms contain an inconsistency. Then the opposite of T and therefore the contradiction $\overline{T}\,T$ can also be deduced from them:

$$A \Rightarrow \overline{T}. \tag{2.103}$$

So, if there is an inconsistency, its existence can be proved by exhibiting any proposition T and its opposite $\overline{T}$ that are both deducible from the axioms. However, in practice it may not be easy to find a T for which one sees how to prove both T and $\overline{T}$.

Evidently, we could prove the consistency of a set of axioms if we could find a feasible procedure which is guaranteed to locate an inconsistency if one exists; so Gödel's theorem seems to imply that no such procedure exists. Actually, it says only that no such procedure *derivable from the axioms of the system being tested* exists.

We shall find that probability theory comes close to this; it is a powerful analytical tool which can search out a set of propositions and detect a contradiction in them if one exists. The principle is that probabilities conditional on contradictory premises do not exist (the hypothesis space is reduced to the empty set). Therefore, put our robot to work; i.e. write a computer program to calculate probabilities $p(B|E)$ conditional on a set of propositions $E = (E_1\,E_2\,\ldots\,E_n)$. Even though no contradiction is apparent from inspection, if there is a contradiction hidden in E, the computer program will crash.

We discovered this 'empirically', and, after some thought, realized that it is not a reason for dismay, but rather a valuable diagnostic tool that warns us of unforeseen special cases in which our formulation of a problem can break down.

If the computer program does not crash, but prints out valid numbers, then we know that the conditioning propositions E_i are mutually consistent, and we have accomplished what one might have thought to be impossible in view of Gödel's theorem. But of course our use of probability theory appeals to principles not derivable from the propositions being tested, so there is no difficulty; it is important to understand what Gödel's theorem does and does not prove.

When Gödel's theorem first appeared, with its more general conclusion that a mathematical system may contain certain propositions that are undecidable within that system, it seems to have been a great psychological blow to logicians, who saw it at first as a devastating obstacle to what they were trying to achieve. Yet a moment's thought shows us

[2] The 1957 edition of Harold Jeffreys' *Scientific Inference* (see Jeffreys, 1931) has a short summary of Gödel's original reasoning which is far clearer and easier to read than any other 'explanation' we have seen. The full theorem refers to other matters of concern in 1931 but of no interest to us right now; the above discussion has abstracted the part of it that we need to understand for our present purposes.

that many quite simple questions are undecidable by deductive logic. There are situations in which one can prove that a certain property must exist in a finite set, even though it is impossible to exhibit any member of the set that has that property. For example, two persons are the sole witnesses to an event; they give opposite testimony about it and then both die. Then we know that one of them was lying, but it is impossible to determine which one.

In this example, the undecidability is not an inherent property of the proposition or the event; it signifies only the incompleteness of our own information. But this is equally true of abstract mathematical systems; when a proposition is undecidable in such a system, that means only that its axioms do not provide enough *information* to decide it. But new axioms, external to the original set, might supply the missing information and make the proposition decidable after all.

In the future, as science becomes more and more oriented to thinking in terms of information content, Gödel's result will be seen as more of a platitude than a paradox. Indeed, from our viewpoint 'undecidability' merely signifies that a problem is one that calls for *inference* rather than deduction. Probability theory as extended logic is designed specifically for such problems.

These considerations seem to open up the possibility that, by going into a wider field by invoking principles external to probability theory, one might be able to prove the consistency of our rules. At the moment, this appears to us to be an open question.

Needless to say, no inconsistency has ever been found from correct application of our rules, although some of our calculations will put them to a severe test. Apparent inconsistencies have always proved, on closer examination, to be misapplications of the rules. On the other hand, guided by Cox's theorems, which tell us where to look, we have never had the slightest difficulty in exhibiting the inconsistencies in the *ad hoc* rules which abound in the literature, which differ in content from ours and whose sole basis is the intuitive judgment of their inventors. Examples are found throughout this book, but particularly in Chapters 5, 15, and 17.

2.6.3 Venn diagrams

Doubtless, some readers will ask, 'After the rather long and seemingly unmotivated derivation of the extended sum rule (2.66), which in our new notation now takes the form

$$P(A+B|C) = P(A|C) + P(B|C) - P(AB|C), \tag{2.104}$$

why did we not illustrate it by the Venn diagram? That makes its meaning so much clearer.' (Here we draw two circles labeled A and B, with intersection labeled AB, all within a circle C.)

The Venn diagram is indeed a useful device, illustrating – in one special case – why the negative term appears in (2.104). But it can also mislead, because it suggests to our intuition more than the actual content of (2.104). Looking at the Venn diagram, we are encouraged to ask, 'What do the points in the diagram mean?' If the diagram is intended to illustrate (2.104), then the probability for A is, presumably, represented by the area of circle A; for then the total area covered by circles A, B is the sum of their separate areas, minus the area of overlap, corresponding exactly to (2.104).

Now, the circle A can be broken down into nonoverlapping subregions in many different ways; what do these subregions mean? Since their areas are additive, if the Venn diagram is to remain applicable they must represent a refinement of A into the disjunction of some mutually exclusive subpropositions. We can – if we have no mathematical scruples about approaching infinite limits – imagine this subdivision carried down to the individual points in the diagram. Therefore these points must represent some ultimate 'elementary' propositions ω_i into which A can be resolved.[3] Of course, consistency then requires us to suppose that B and C can also be resolved into these same propositions ω_i.

We have already jumped to the conclusion that the propositions to which we assign probabilities correspond to sets of points in some space, that the logical disjunction $A + B$ stands for the union of the sets, the conjunction AB for their intersection, and that the probabilities are an additive measure over those sets. But the general theory we are developing has no such structure; all these things are properties only of the Venn diagram.

In developing our theory of inference we have taken special pains to avoid restrictive assumptions which would limit its scope; it is to apply, in principle, to any propositions with unambiguous meaning. In the special case where those propositions happen to be statements about sets, the Venn diagram is an appropriate illustration of (2.104). But most of the propositions about which we reason, for example,

$$A \equiv \text{it will rain today}, \tag{2.105}$$

$$B \equiv \text{the roof will leak}, \tag{2.106}$$

are simply declarative statements of fact, which may or may not be resolvable into a disjunction of more elementary propositions within the context of our problem.

Of course, one can always force such a resolution by introducing irrelevancies; for example, even though the above-defined B has nothing to do with penguins, we could still resolve it into the disjunction

$$B = BC_1 + BC_2 + BC_3 + \cdots + BC_N, \tag{2.107}$$

where $C_k \equiv$ the number of penguins in Antarctica is k. By choosing N sufficiently large, we will surely be making a valid statement of Boolean algebra; but this is idle, and it cannot help us to reason about a leaky roof.

Even if a meaningful resolution exists in our problem, it may not be of any use to us. For example, the proposition 'rain today' could be resolved into an enumeration of every conceivable trajectory of each individual raindrop; but we do not see how this could help a meteorologist trying to forecast rain. In real problems, there is a natural end to this resolving, beyond which it serves no purpose and degenerates into an empty formal exercise. We shall

[3] A physicist refuses to call them 'atomic' propositions, for obvious reasons.

give an explicit demonstration of this later (Chapter 8), in the scenario of 'Sam's broken thermometer': does the exact way in which it broke matter for the conclusions that Sam should draw from his corrupted data?

In some cases there is a resolution so relevant to the context of the problem that it becomes a useful calculational device; Eq. (2.98) was a trivial example. We shall be glad to take advantage of this whenever we can, but we cannot expect it in general.

Even when both A and B can be resolved in a way meaningful and useful in our problem, it would seldom be the case that they are resolvable into the *same* set of elementary propositions ω_i. And we always reserve the right to enlarge our context by introducing more propositions $D, E, F, \ldots$ into the discussion; and we could hardly ever expect that all of them would continue to be expressible as disjunctions of the *same* original set of elementary propositions ω_i. To assume this would be to place a quite unnecessary restriction on the generality of our theory.

Therefore, the conjunction AB should be regarded simply as the statement that both A and B are true; it is a mistake to try to read any more detailed meaning, such as an intersection of sets, into it in every problem. Then $p(AB|C)$ should also be regarded as an elementary quantity in its own right, not necessarily resolvable into a sum of still more elementary ones (although if it is so resolvable this may be a good way of calculating it). We have adhered to the original notation $A + B$, AB of Boole, instead of the more common $A \vee B$, $A \wedge B$, or $A \cup B$, $A \cap B$, which everyone associates with a set theory context, in order to head off this confusion as much as possible.

So, rather than saying that the Venn diagram justifies or explains (2.104), we prefer to say that (2.104) explains and justifies the Venn diagram, in one special case. But the Venn diagram has played a major role in the history of probability theory, as we note next.

2.6.4 The 'Kolmogorov axioms'

In 1933, A. N. Kolmogorov presented an approach to probability theory phrased in the language of set theory and measure theory (Kolmogorov, 1933). This language was just then becoming so fashionable that today many mathematical results are named, not for the discoverer, but for the one who first restated them in that language. For example, in the theory of continuous groups the term 'Hurwitz invariant integral' disappeared, to be replaced by 'Haar measure'. Because of this custom, some modern works – particularly by mathematicians – can give one the impression that probability theory started with Kolmogorov.

Kolmogorov formalized and axiomatized the picture suggested by the Venn diagram, which we have just described. At first glance, this system appears so totally different from ours that some discussion is needed to see the close relationship between them. In Appendix A we describe the Kolmogorov system and show that, for all practical purposes, the four axioms concerning his probability measure, first stated arbitrarily (for which Kolmogorov has been criticized), have all been derived in this chapter as necessary to meet our consistency requirements. As a result, we shall find ourselves defending Kolmogorov against his critics on many technical points. The reader who first learned probability theory on the Kolmogorov basis is urged to read Appendix A at this point.

Our system of probability, however, differs conceptually from that of Kolmogorov in that we do not interpret propositions in terms of sets, but we do interpret probability distributions as carriers of incomplete information. Partly as a result, our system has analytical resources not present at all in the Kolmogorov system. This enables us to formulate and solve many problems – particularly the so-called ‘ill posed’ problems and ‘generalized inverse’ problems – that would be considered outside the scope of probability theory according to the Kolmogorov system. These problems are just the ones of greatest interest in current applications.