CHAPTER 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: stated broadly,

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. The resulting rules have a long, complicated, and astonishing history, full of lessons for scientific methodology in general (see Comments at the end of several Chapters).

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 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. $\qquad (B|C)$

(2) Having accepted B as true, decide that A is true. $\qquad (A|BC)$

Or, equally well,

(1') Decide that A is true. $\qquad (A|C)$

(2') Having accepted A as true, decide that B is true. $\qquad (B|AC)$

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

$$(AB|C) = F[(A|C), (B|C)]$$

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

$$u = (AB|C), \qquad v = (A|C), \qquad w = (B|AC), \qquad x = (B|C), \qquad y = (A|BC)$$

If u is to be expressed as a function of two or more of v, w, x, y, there are eleven 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) shows 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:

$$B|C' > B|C \ ,$$

$$A|BC' = A|BC \ ,$$

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

$$AB|C' \geq AB|C$$

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

$$B|C'' = B|C$$

$$A|BC'' > A|BC$$

we require that

$$AB|C'' \geq AB|C$$

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 the same 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 differentiable [this is not necessary; see the discussion following (2–4)], then we have

$$F_1(x, y) \equiv \frac{\partial F}{\partial x} \geq 0 \tag{2–2a}$$

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

$$F_2(x, y) \equiv \frac{\partial F}{\partial y} \geq 0 \tag{2–2b}$$

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

Next we impose the Desideratum III(a) 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):

$$(ABC|D) \; = \; F[(BC|D),\ (A|BCD)]$$

and then in the plausibility $(BC|D)$ we can again apply (2–1) to give

$$(ABC|D) \; = \; F\{F[(C|D),\ (B|CD)],\ (A|BCD)\} \tag{2–3a}$$

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:

$$(ABC|D) \; = \; F[(C|D),\ (AB|CD)] \; = \; F\{(C|D),\ F[(B|CD), (A|BCD)]\} \tag{2–3b}$$

If this rule is to represent a consistent way of reasoning, the two expressions (2–3a), (2–3b) 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,

$$F[F(x, y), z] \; = \; F[x, F(y, z)] \quad . \tag{2–4}$$

This equation has a long history in mathematics, starting from a work of N. H. Abel in 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 [Eq. (2–17) below] without assuming differentiability; unfortunately, the proof fills eleven pages (256–267) of his book. We give here the shorter proof by R. T. Cox (1961), which assumes differentiability.

It is evident that (2–4) has a trivial solution, $F(x, y) =$const. But that violates our monotonicity requirement (2–2) and is in any event useless for our purposes. Unless (2–4) has a nontrivial solution, this approach will fail; so we seek the most general nontrivial solution. Using the abbreviations

$$u \equiv F(x,y), \qquad\qquad v \equiv F(y,z) \quad , \tag{2–5}$$

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

$$F(x,v) \;=\; F(u,z)\;. \tag{2–6}$$

Differentiating with respect to x and y we obtain, in the notation of (2–2),

$$\begin{aligned} F_1(x,v) &= F_1(u,z)\; F_1(x,y) \\ F_2(x,v)\; F_1(y,z) &= F_1(u,z)\; F_2(x,y) \end{aligned} \tag{2–7}$$

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

$$G(x,v)\; F_1(y,z) \;=\; G(x,y) \tag{2–8}$$

where we use the notation $G(x, y) \equiv F_2(x, y)/F_1(x, y)$. Evidently, the left-hand side of (2–8) must be independent of z. Now (2–8) can be written equally well as

$$G(x,v)\; F_2(y,z) \;=\; G(x,y)\; G(y,z) \tag{2–9}$$

and, denoting the left-hand sides of (2–8), (2–9) 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

$$G(x,y) \;=\; r\,\frac{H(x)}{H(y)} \tag{2–10}$$

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 (2–10), (2–8) and (2–9) become

$$F_1(y,z) = H(v)/H(y) \tag{2–11}$$

$$F_2(y,z) = r\;H(v)/H(z) \tag{2–12}$$

and the relation $dv = dF(y, z) = F_1dy + F_2dz$ takes the form

$$\frac{dv}{H(v)} \;=\; \frac{dy}{H(y)} \;+\; r\,\frac{dz}{H(z)} \tag{2–13}$$

or, on integration,

$$w[F(y,z)] \;=\; w(v) \;=\; w(y)\; w^r(z) \tag{2–14}$$

where

$$w(x) \equiv \exp\left[\int^x \frac{dx}{H(x)}\right]\,, \tag{2–15}$$

the absence of a lower limit on the integral signifying an arbitrary multiplicative factor in w. But taking the function $w(\cdot)$ of (2–6) and applying (2–14), we obtain $w(x)w^r(v) = w(u)w^r(z)$; applying (2–14) again, our functional equation now reduces to

$$w(x)w^r(y)[w(z)]^{r^2} = w(x)w^r(y)w^r(z)$$

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

$$w[F(x,y)] = w(x)\,w(y) \tag{2–16}$$

$$F(x,y) = w^{-1}[w(x)w(y)] \quad . \tag{2–17}$$

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

$$w(AB|C) = w(A|BC)\,w(B|C) = w(B|AC)\,w(A|C) \tag{2–18}$$

which we shall call henceforth the *product rule*. By its construction (2–15), $w(x)$ 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 (2–18) has been derived as a necessary condition for consistency in the sense of Desideratum III(a). Conversely, it is evident that (2–18) 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 (2–3); but if (2–18) 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 (2–18) 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:

$$AB|C = B|C$$

and also we will have

$$A|BC = A|C$$

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, (2–18) reduces to

$$w(B|C) = w(A|C)\,w(B|C) \tag{2–19}$$

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

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

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

$$AB|C = A|C$$

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:

$$A|BC = A|C \quad .$$

In this case, equation (2–18) reduces to

$$w(A|C) \;=\; w(A|C)\,w(B|C) \tag{2–20}$$

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 0 or $+\infty$ (the choice $-\infty$ is ruled out because then by continuity $w(B|C)$ would have to be capable of negative values; (2–20) 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 $0 \le w(x) \le 1$ 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.]

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 $u \equiv w(A|B)$, $v \equiv w(\overline{A}|B)$, there must exist some functional relation

$$v \;=\; S(u) \quad . \tag{2–21}$$

Evidently, qualitative correspondence with common sense requires that $S(u)$ be a continuous monotonic decreasing function in $0 \le u \le 1$, with extreme values $S(0) = 1$, $S(1) = 0$. 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 $A\overline{B}$:

$$w(AB|C) \;=\; w(A|C)\,w(B|AC) \tag{2–22}$$

$$w(A\overline{B}|C) \;=\; w(A|C)\,w(\overline{B}|AC). \tag{2–23}$$

Thus, using (2–21) and (2–23), Eq. (2–22) becomes

$$w(AB|C) = w(A|C)\,S[w(\overline{B}|AC)] = w(A|C)\,S\left[\frac{w(A\overline{B}|C)}{w(A|C)}\right]. \tag{2–24}$$

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

$$w(A|C)\;S\left[\frac{w(A\overline{B}|C)}{w(A|C)}\right] \;=\; w(B|C)\;S\left[\frac{w(B\overline{A}|C)}{w(B|C)}\right] \quad . \tag{2–25}$$

This must hold for all propositions A, B, C; in particular, (2–25) must hold when

$$\overline{B} = AD \tag{2–26}$$

where D is any new proposition. But then we have the truth–values noted before in (1–8):

$$A\overline{B} = \overline{B}, \qquad B\overline{A} = \overline{A}, \tag{2–27}$$

and in (2–25) we may write

$$\begin{aligned} w(A\overline{B}|C) &= w(\overline{B}|C) = S[w(B|C)] \\ w(B\overline{A}|C) &= w(\overline{A}|C) = S[w(A|C)]\,. \end{aligned} \tag{2–28}$$

Therefore, using now the abbreviations

$$x \equiv w(A|C), \qquad y \equiv w(B|C) \tag{2–29}$$

Eq. (2-25) becomes a functional equation

$$x\,S\left[\frac{S(y)}{x}\right] = y\,S\left[\frac{S(x)}{y}\right], \qquad \begin{aligned} &0 \le S(y) \le x, \\ &0 \le x \le 1 \end{aligned} \tag{2–30}$$

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

$$S[S(x)] = x \tag{2–31}$$

which states that $S(x)$ is a self-reciprocal function; $S(x) = S^{-1}(x)$. Thus, from (2–21) it follows also that $u = S(v)$. But this expresses only the evident fact that the relation between A, $\overline{A}$ 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 (1–6); if it had not been obvious before, we should be obliged to recognize it at this point.

The domain of validity given in (2–30) 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

$$0 \le w(D|AC) \le 1 \quad . \tag{2–32}$$

But $S(y) = w(AD|C) = w(A|C)w(D|AC)$, and so (2–32) is just $(0 \le S(y) \le x)$, as stated in (2–30). This domain is symmetric in x, y; it can be written equally well with them interchanged. Geometrically, it consists of all points in the $x - y$ 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 (2–30) says for points lying infinitesimally above it. For if we set $y = S(x) + \epsilon$, then as $\epsilon \to 0+$ two terms in (2–30) 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

$$\frac{S(x)}{y} = 1 - e^{-q} \tag{2–33}$$

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

$$S(1 - \delta) = S(1 - e^{-q}) = \exp[-J(q)] \ , \tag{2–34}$$

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

Considering now x, q as the independent variables, we have from (2–33)

$$S(y) = S[S(x)] + e^{-q}\, S(x)\, S'[S(x)] + O(e^{-2q}) \ .$$

Using (2–31) and its derivative $S'[S(x)]\, S'(x) = 1$, this reduces to

$$\frac{S(y)}{x} = 1 \ - e^{-(\alpha+q)} + O(e^{-2q}) \tag{2–35}$$

where

$$\alpha(x) \ \equiv \ \log\left[\frac{-x\ S'(x)}{S(x)}\right] \ > \ 0 \ . \tag{2–36}$$

With these substitutions our functional equation (2–30) becomes

$$J(q+\alpha) \ - \ J(q) \ = \ \log\left[\frac{x}{S(x)}\right] \ + \ \log(1 \ - \ e^{-q}) \ + \ O(e^{-2q}) \, , \qquad \begin{array}{c} 0 < q < \infty \\ 0 < x \le 1 \end{array} \tag{2–37}$$

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

$$J(q) \sim a + bq + O(e^{-q}) \ , \tag{2–38}$$

with positive slope

$$b = \alpha^{-1} \ \log\left[\frac{x}{S(x)}\right] \ . \tag{2–39}$$

In (2–38) there is no periodic term with period α, because (2–37) 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 (2–39) must be independent of x. This gives, using (2–36),

$$\frac{x}{S(x)} = \left[\frac{-x\, S'(x)}{S(x)}\right]^b \, , \qquad 0 < b < \infty \tag{2–40}$$

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

$$S^{m-1} dS \ + \ x^{m-1} dx \ = \ 0 \quad . \tag{2–41}$$

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

$$S(x) \ = \ (1 - x^m)^{1/m} \ , \qquad \begin{array}{c} 0 \le x \le 1 \\ 0 < m < \infty \end{array} \tag{2–42}$$

and conversely, we verify at once that (2–42) is a solution of (2–30).

The result (2–42) 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 non–differentiable functions, we would have rejected this whole theory as a qualitative violation of common sense]. In any event, (2–42) is the most general function satisfying the functional equation (2–30) 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 (2–30) used the special choice (2–26) for B, we have shown thus far only that (2–42) is a necessary condition to satisfy the general consistency requirement (2–25). To check its sufficiency, substitute (2–42) into (2–25). We obtain

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

a trivial identity by virtue of (2–18) and (2–23). Therefore, (2–42) is the necessary and sufficient condition on $S(x)$ for consistency in the sense (2–25).

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 (2–18). Our result (2–42) states that this same function must also obey a sum rule:

$$w^m(A|B) + w^m(\overline{A}|B) = 1 \tag{2–43}$$

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

$$w^m(AB|C) = w^m(A|C)\; w^m(B|AC) \;=\; w^m(B|C)\; w^m(A|BC) \tag{2–44}$$

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

$$p(x) \equiv \; w^m(x) \tag{2–45}$$

and our rules take the form

$$p(AB|C) \;=\; p(A|C)\, p(B|AC) \;=\; p(B|C)\, p(A|BC) \tag{2–46}$$

$$p(A|B) \;+\; p(\overline{A}|B) \;=\; 1\;. \tag{2–47}$$

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 (2–46), (2–47) in which $p(x)$ is any continuous monotonic increasing function with the range $0 \le p(x) \le 1$.

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 (2–46) and sum rule (2–47), formulas for the plausibility of the conjunction AB and the negation $\overline{A}$. But we noted, in the discussion following Eq. (1–12), 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

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

and finally,

$$p(A+B|C) = p(A|C) + p(B|C) - p(AB|C) \,. \tag{2–48}$$

This generalized sum rule is one of the most useful in applications. Evidently, the primitive sum rule (2–47) is a special case of (2–48), with the choice $B = \overline{A}$.

Exercise 2.1 Is it possible to find a general formula for $p(C|A+B)$, analogous to (2–48), 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 $\{A_1, \cdots, A_n\}$ 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 $p(C|A_i X)$:

$$p(C|(A_1 + \cdots + A_n)X) = p(C|A_1 X + A_2 X + \cdots + A_n X) = \frac{\sum_i p(A_i|X)\,p(C|A_i X)}{\sum_i p(A_i|X)} \,. \tag{2–49}$$

To extend the result (2–48), we noted following (1–11) 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 (1–11). 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 (2–48) will yield the plausibility of any logical sum of the Q_i. In fact, these conjunctions are mutually exclusive, so we shall find [Eq. (2–64) below] that this reduces to a simple sum $\Sigma_i p(Q_i|C)$ of at most $(2^n - 1)$ terms.

So, just as conjunction and negation are an adequate set 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 $\{A_1, \cdots, A_n\}$. 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.

Qualitative Properties

Now let us check to see how the theory based on (2–46) and (2–47) 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 (2–47) 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 (1–9) to state the major premise:

$$\begin{array}{cc} A \Rightarrow B & A \Rightarrow B \\ \underline{A \text{ true}} & \underline{B \text{ false}} \\ B \text{ true} & A \text{ false} \end{array} \tag{2–50}$$

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

$$C \equiv \text{“}A \Rightarrow B\text{”} \tag{2–51}$$

then these syllogisms correspond to our product rule (2–46) in the forms

$$p(B|AC) = \frac{p(AB|C)}{p(A|C)}\ , \qquad p(A|\overline{B}C) = \frac{p(A\overline{B}|C)}{p(\overline{B}|C)} \tag{2–52}$$

respectively. But from (2–50) we have $p(AB|C) = p(A|C)$ and $p(A\overline{B}|C) = 0$, and so (2–52) reduces to

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

as stated in the syllogisms (2–50). 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), (1–4). To show that those original qualitative statements always follow from the present rules, note that the first weak syllogism

$$\begin{array}{c} A \Rightarrow B \\ B \text{ is true} \\ \hline \text{Therefore, } A \text{ becomes more plausible} \end{array} \tag{2–53}$$

corresponds to the product rule (2–46) in the form

$$p(A|BC) = p(A|C)\,\frac{p(B|AC)}{p(B|C)}\ . \tag{2–54}$$

But from (2–50), $p(B|AC) = 1$, and since $p(B|C) \le 1$, (2–54) gives

$$p(A|BC) \ge p(A|C) \tag{2–55}$$

as stated in the syllogism. Likewise, the syllogism (1–4)

$$\begin{array}{c} A \Rightarrow B \\ A \text{ is false} \\ \hline \text{Therefore, } B \text{ becomes less plausible} \end{array} \tag{2–56}$$

corresponds to the product rule in the form

$$p(B|\overline{A}C) = p(B|C)\,\frac{p(\overline{A}|BC)}{p(\overline{A}|C)}\ . \tag{2–57}$$

But from (2–55) it follows that $p(\overline{A}|BC) \le p(\overline{A}|C)$; and so (2–57) gives

$$p(B|\overline{A}C) \le p(B|C) \tag{2–58}$$

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 (2–54). Letting now C stand for 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

$$p(B|AC) \ > \ p(B|C) \tag{2–59}$$

and (2–54) gives at once

$$p(A|BC) \ > \ p(A|C) \tag{2–60}$$

as stated in the syllogism.

But now we have more than the mere qualitative statement (2–60). 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 (2–54) 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), Cox (1961), Tribus (1969), de Finetti (1974), 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.

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 (2–46), (2–47) 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 (1–23), 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 (2–48), as follows. The first application gives

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

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

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

Applying (2–48) again, we obtain seven terms which can be grouped as follows:

$$\begin{aligned} p(A_1 + A_2 + A_3|B) &= p(A_1|B) + p(A_2|B) + p(A_3|B) \\ &- p(A_1A_2|B) - p(A_2A_3|B) - p(A_3A_1|B) \\ &+ p(A_1A_2A_3|B) \end{aligned} \qquad (2\text{–}61)$$

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

$$p(A_iA_j|B) = p(A_i|B)\delta_{ij} \ . \qquad (2\text{–}62)$$

Then the last four terms of (2–61) vanish, and we have

$$p(A_1 + A_2 + A_3|B) = p(A_1|B) + P(A_2|B) + P(A_3|B) \ . \qquad (2\text{–}63)$$

Adding more propositions A_4, A_5, *etc.*, it is easy to show by induction that if we have n mutually exclusive propositions $\{A_1 \cdots A_n\}$, (2–63) generalizes to

$$p(A_1 + \cdots + A_m|B) = \sum_{i=1}^{m} p(A_i|B) \ , \qquad 1 \le m \le n \qquad (2\text{–}64)$$

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

In conventional expositions, Eq. (2–64) 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 (2–64) as the primitive, fundamental relation is one which we are particularly anxious to avoid (see Comments at the end of this Chapter).

Now suppose that the propositions $\{A_1 \dots A_n\}$ 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 (2–64) for $m = n$ must be unity:

$$\sum_{i=1}^{n} p(A_i|B) = 1 \ . \qquad (2\text{–}65)$$

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 $\{A_1 \ldots A_n\}$ 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 $\{A'_1 \ldots A'_n\}$, given by

$$\begin{aligned} A'_1 &\equiv A_2 \\ A'_2 &\equiv A_1 \\ A'_k &\equiv A_k\ , \qquad 3 \le k \le n \end{aligned} \tag{2–66}$$

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

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 $\{A_1 \ldots A_n\}$ in both problems. Therefore we must have

$$p(A_1|B)_I = p(A'_2|B)_{II} \tag{2–67}$$

and similarly

$$p(A_2|B)_I = p(A'_1|B)_{II} \quad . \tag{2–68}$$

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.

But 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 $\{A'_1 \ldots A'_n\}$ in Problem II, *including their labeling*, as it is about the set $\{A_1 \ldots A_n\}$ in Problem I.

Now we invoke our Desideratum of Consistency in the sense IIIc in (1–23). This stated that equivalent states of knowledge must be represented by equivalent plausibility assignments. In equations, this statement is

$$p(A_i|B)_I = p(A'_i|B)_{II}\,, \qquad i = 1, 2, \ldots, n \tag{2–69}$$

which we shall call the *symmetry equations.* But now, combining equations (2–67), (2–68), (2–69) we obtain

$$p(A_1|B)_I = p(A_2|B)_I\ . \tag{2–70}$$

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 (2–70). 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 a later Chapter, 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

$$p(A_i|B)_I \;=\; p(A''_k|B)_{III} \qquad (2\text{–}71)$$

which show how Problems I and III are related to each other; and 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

$$p(A_k|B)_I = p(A''_k|B)_{III}\,, \qquad k = 1, 2, \ldots, n \qquad (2\text{–}72)$$

From (2–71) and (2–72) we obtain n equations of the form

$$p(A_i|B)_I \;=\; p(A_k|B)_I \qquad (2\text{–}73)$$

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 in which, for given i, the index k will range over all of the $(n-1)$ others in (2–73). 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, Eq. (2–65) will hold, and the only possibility is therefore

$$p(A_i|B)_I = \frac{1}{n}\,, \qquad (l \le i \le n) \qquad (2\text{–}74)$$

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 (2–74) 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 (2–74), 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 (2–74) 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 get started. 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 (2–74), 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* $(0 \le p \le 1)$.

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

The question of uniqueness is therefore disposed of automatically by the result (2–74); 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 while 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 they 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 [$dU = TdS - PdV$, $dG = -SdT + VdP$, *etc.*] in terms of this particular scale. Because of this, numerical values of Kelvin temperatures 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 (2–74). Consider the traditional "Bernoulli Urn" of probability theory; ours is known to contain ten balls of identical size and weight, labelled $\{1, 2, \ldots, 10\}$. 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 (2–74) 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 i'th ball is drawn" , $1 \le i \le 10$. Since the background information is indifferent to these ten possibilities, (2–74) applies and the robot assigns

$$p(A_i|B) \;=\; \frac{1}{10} \;, \qquad 1 \le i \le 10$$

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

$$p(\text{Black}|B) \;=\; p(A_4 + A_6 + A_7|B) \quad .$$

But these are mutually exclusive propositions (*i.e.*, they assert mutually exclusive events) so (2–64) applies and the robot's conclusion is

$$p(\text{Black}|B) \;=\; \frac{3}{10} \tag{2–75}$$

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, $(0 \le M \le N)$, false on the rest, we have

$$p(A|B) \;=\; \frac{M}{N} \quad . \tag{2–76}$$

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 of 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. Limits on Probability Values. 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 $a \le b$, show that the probability of the conjunction cannot exceed that of the least probable proposition: $0 \le P(AB|C) \le a$, and the probability of the disjunction cannot be less than that of the most probable proposition: $b \le P(A + B|C) \le 1$. 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.

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:

$$P(A|B)$$

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

$$f(r|n, p)$$

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 short–cuts 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.

COMMENTS

It has taken us two Chapters of close reasoning to get back to the point (2–76) 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.

'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: "The robot – or 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.

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.

First, 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 $\{A_1, A_2, \ldots, A_n\}$ 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. [For, 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 $A = A_1\,A_2 \cdots A_n$ be the system of axioms underlying a mathematical theory and T any proposition, or theorem, deducible from them:†

$$A \Rightarrow T\ .$$

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 R. A. 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.‡

† In Chapter 1 we noted the tricky distinction between the weak property of formal implication and the strong one of logical deducibility; by 'implication of a proposition C' we really mean '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.

‡ The 1957 Edition of Harold Jeffreys' *Scientific Inference* 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.

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:

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

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.

Yet 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. 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. It will be used for this purpose later, particularly in Chapter 21.

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 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 come to seem 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 still 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 the sequel, but particularly in Chapters 5, 15, 17.

Venn Diagrams. Doubtless, some readers will ask, "After the rather long and seemingly unmotivated derivation of the extended sum rule (2–48), which in our new notation now takes the form:

$$P(A+B|C) = P(A|C) + P(B|C) - P(AB|C) \tag{2–48}$$

why did we not illustrate it by the Venn diagram? That makes its meaning so much clearer." [Here we draw two circles labelled A and B, with intersection labelled 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 (2–48). But it can also mislead, because it suggests to our intuition more than the actual content of (2–48). 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 (2–48), then the probability of 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 (2–48).

Now the circle A can be broken down into non–overlapping 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 sub–propositions. 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. Of course, consistency then requires us to suppose that B and C can also be resolved into these same propositions ω_i.

Already, we have 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, 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 (2–48). But most of the propositions about which we reason, for example,

$$A \equiv \text{“It will rain today,”}$$
$$B \equiv \text{“The roof will leak,”}$$

are simply declarative statements of fact, which may or may not be resolvable into 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:

$$B = BC_1 + BC_2 + BC_3 + \cdots + BC_N$$

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 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?

But in some cases there is a resolution so relevant to the context of the problem that it becomes a useful calculational device; Eq. (2–75) 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 (2–48), we prefer to say that (2–48) 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.

The "Kolmogorov Axioms" In 1933, A. N. Kolmogorov presented an approach to probability theory phrased in the language of set theory and measure theory. 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 group theory 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 relation 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.

However, our system of probability differs conceptually from that of Kolmogorov in that we do not interpret propositions in terms of sets. 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.