Lecture 3

LAPLACE'S MODEL OF COMMON SENSE

We have now formulated our problem, and it ought to be a matter of straightforward mathematics to work out the consequences of our three desiderata:

1. Representation of degrees of plausibility by real numbers.
2. Qualitative correspondence with common sense.
3. Consistency.

This seems in retrospect an obvious and natural thing to do; but historically, the rules we are about to deduce were first stated as arbitrary axioms, on intuitive grounds, without any attempt to demonstrate their uniqueness or consistency. This, of course, left room for practically endless controversy; if the rules are introduced in that way, what right have we to suppose that they are any better than a hundred other arbitrary ones we could invent? It was just this kind of doubt, strengthened by some ridiculous misapplications, that led many to reject Laplace's work and to deny that probability theory has any connection with inductive reasoning. As a result, the development of statistical theory was delayed for many years, and the very "latest" advances in this field amount only to a rediscovery of methods that had been described and used by Laplace and Daniel Bernoulli in the 18th century.

To the best of my knowledge, the first person to see that there is a better way of developing the theory was Professor R. T. Cox (Cox, 1946; 1961). Instead of stating the rules in a way that leaves their consistency and

uniqueness open to doubt, the requirement that they be consistent can be imposed from the start as one of the basic conditions of the theory; and then their uniqueness can be deduced mathematically. Cox's argument, which we follow here, therefore cuts the ground out from under more than a century of unjust criticisms of Laplace's methods.

3.1 Deduction of Rule 1.

We first seek a consistent rule for obtaining the plausibility of AB from the plausibilities of A and B separately. In particular, let us find the plausibility $(AB|C)$; on what others must it depend? Now in order for AB to be a true proposition, it is certainly necessary that B be 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 anything about A, so if we have $(B|C)$ and $(A|BC)$ we will not need $(A|C)$. It would tell us nothing about AB that we didn't already have. Similarly, $(A|B)$ and $(B|A)$ are not needed; whatever plausibility A or B might have in the absence of data C could not be relevant to judgments of a case in which we know from the start that C is true.

We could, of course, interchange A and B in the above paragraph, so the knowledge of $(A|C)$ and $(B|AC)$ would also suffice to determine $(BA|C) \equiv (AB|C)$. The fact that we must obtain the same value for $(AB|C)$ no matter which procedure we choose will be one of our conditions of consistency.

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

$$(AB|C) = F[(B|C),\ (A|BC)] \tag{3-1}$$

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 the conditions that we've imposed on our robot. 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, if I'm told that Mr. Jones lives in Dallas, it might be quite plausible that his eyes are blue, and it might be quite plausible that his hair is black; and it's reasonably plausible that both are true. But, if I'm told that Mr. Smith lives in St. Louis, it is quite plausible that his left eye is blue, and it's quite plausible that his right eye is brown; but it's extremely implausable 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 human beings do, even qualitatively, with that kind of functional relation.

You might try further a relation of the form

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

in which you try to take the above cases into account by allowing all four of these simple plausibilities to determine $(AB|C)$. But even here you can produce counter-examples which show that a function of this form could not reproduce plausible reasoning even qualitatively like ours.

You can blow this up into a whole research project, if you like. Thus, introduce the real numbers

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

If u is to be expressed as a function of two or more of v, w, x, and 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 implies A false, etc. If you do this, Myron

Tribus has shown (Tribus, 1969) 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 $u = F(w,v)$, which are just the two possibilities already suggested.

Another way of looking at this, suggested by Mr. Alfred S. Gilman, may seem more attractive than this laborious elimination of alternatives, one by one. We may regard the process of deciding that AB is true as a sequence of two "mental transitions" in which there are only two possible routes, illustrated by the decision tree diagram, Fig. 3.1. In order to decide that AB is true, we

(1) decide that B is true,

(2) having accepted B as true, decide that A is true.

or, we can

(1') decide that A is true,

(2') having accepted A as true, decide that B is true.

Along either route, the state of knowledge in which we decide to make the next transition is indicated by the plausibility symbols on the arrows.

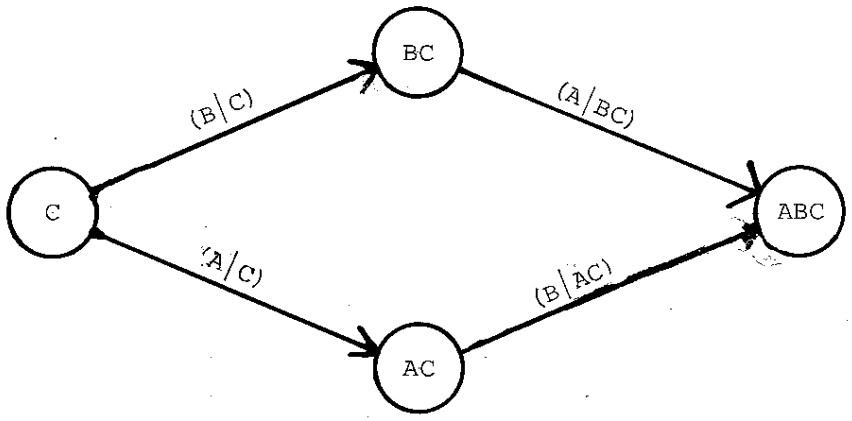

Fig. 3.1. The possible "mental transitions" in deciding that A and B are true, given that C is true.

However you like to view this, I don't think you'll be able to produce any situation where equation (3-1) does not reproduce qualitatively the way you would reason about the situation. (If you can, then all I can say is that your common sense is qualitatively different from mine--and Laplace's--and you are free to design your own robot!)

Now let's start imposing our conditions on the form of this function and see if we can nail down what function it has to be. If anything increases the plausibility $(B|C)$, then that must produce only an increase, never a decrease, in the plausibility $(AB|C)$. Similarly, if anything increases $(A|BC)$, this must also produce an increase, not a decrease, in $(AB|C)$. The only case where it would not produce an increase is where the other independent variable happened to represent impossibility; if we know that A is impossible given C, then, of course, the plausibility of B could increase without affecting $(AB|C)$. Also, the function $F(x,y)$ must be continuous; for otherwise we could produce a situation where an arbitrarily small increase in one of the plausibilities on the right side still results in the same big increase in $(AB|C)$.

In summary, $F(x,y)$ must be a continuous monotonic increasing function of both x and y. I will assume that it's a differentiable function. The derivatives cannot be negative, and they can be zero only in the case where AB is impossible. Now for the condition that it should be consistent.

Suppose that I try to find the plausibility $(ABC|D)$ that three propositions would be true simultaneously. I can do this in two different ways. If the rule is going to be consistent, we've got to get the same result for either order of carrying out the operations. I can first say that BC will be considered a single proposition, and then apply our rule. This plausibility would then be

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

and now in this plausibility of $(BC|D)$ we can again apply the rule to give us

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

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:

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

So by doing it in the other order, we come out with a different expression. If this rule is to represent a consistent way of reasoning, these two expressions must always be the same. The condition that our robot will reason consistently in this case takes the form of a functional equation,

$$F[F(x,y),z] = F[x,F(y,z)]. \qquad (3\text{-}2)$$

Conversely, if this functional equation is satisfied, then our original rule is automatically consistent for all possible ways of finding the joint plausibility of any number of propositions; $(ABCDE|F)$, for example. You can see that there are an enormous number of different ways you can work this out by successive applications of Equation (3-1). And you can show by induction that if the functional Equation (3-2) is satisfied, then you're guaranteed to get the same answer for every possible way of doing it.

This functional equation is one which has quite a long history in mathematics. The earliest reference to it that I know about goes back to 1826, and is a paper by N. H. Abel in the first issue of Crelle's journal. Abel considered equation (3-2) merely as an amusing exercise, and found the general solution by reducing it to a differential equation. The solution has been rediscovered probably dozens of times since 1826. In particular, this is done in a paper by R. T. Cox (Cox, 1946) which I rate as one of the most important ever written on the foundations of probability theory. Cox established the conditions for consistency of this theory in the sense (a) given above, and

my only contribution was to add the qualitative requirements and the other conditions of consistency, which are needed to make the result unique. In a later book (Cox, 1961) Cox's work is given more fully, with some improvements in the derivations. For an appreciation of the importance of Cox's contributions to probability theory, see my review of his book (Jaynes, 1963b).

A particularly neat mathematical treatment of our functional equation (3-2) has been given by J. Aczel in a paper (Aczel, 1948) and in his monumental book on functional equations (Aczel, 1966; Sec. 6.2). He calls it, "The associativity equation." Let me just quote you the theorem that Aczel gives. He says, "Let's let

$$z = x \circ y$$

where

$$x \circ y$$

represents any operation which maps z into the same interval with x and y. In other words, if x is in the interval from a to b, and y is in the interval from a to b, then this operation is one which will always put z into the same interval." He gives a theorem which is exactly backwards from the way we would want it for our application. He considered a formula for the design of the most general slide rule. The general condition that z could be calculated without ambiguity on a slide rule calibrated with numbers x and y is, of course, that there is some monotonic function $f(z) = f(x) + f(y)$. If this is true then you can make a slide rule which gives z in terms of x and y. Aczel shows that a necessary and sufficient condition for that is that the operation x o y must have the following properties:

(1) It must be monotonic: if $x' > x$, then $x' \circ y > x \circ y$, and similarly for y.

(2) It must be continuous: $\lim (x \circ y) = (\lim x) \circ (\lim y)$.

(3) It must be associative: $(x \circ y) \circ z = x \circ (y \circ z)$.

You see that these are precisely the conditions that we have imposed on our

function $z = F(x,y)$. It had to be a monotonic, continuous operation in order to agree qualitatively with common sense. The condition that it should represent a consistent kind of reasoning was just the condition that it be associative. We conclude that the general relation between x, y, z, implied by $z = F(x,y)$ must be expressible in the form

$$F(x,y) = f^{-1}[f(x) + f(y)], \text{ or}$$

$$f(z) = f(x) + f(y).$$

Now, of course, we can write this equally well as a product,

$$p(z) = p(x)\ p(y),$$

where $p(x) \equiv \exp[f(x)]$ is still an arbitrary continuous monotonic function. It makes no difference which form we choose, but the second choice will prove more convenient later on.

So our rule for finding the plausibility of both A and B takes the form

$$p(AB|C) = p(A|BC)\ p(B|C). \tag{3-3}$$

The condition that this shall represent reasoning qualitatively like ours can tell us something more about this function p(x). For example, let's imagine first that A is certain, given C. What would happen then? Well, if A is certain given C, then in the "environment" produced by knowledge of C, AB and B are the same proposition, in the sense that one is true if and only if the other is true. So, the plausibility that AB is true must be just the plausibility that B is true:

$$(AB|C) = (B|C).$$

And also we would have:

$$(A|BC) = (A|C),$$

because if A is already certain given C, the fact that we may also have B given would not be relevant; it's still certain. To what is our equation (3-3) reduced in this case? It then says

$$p(B|C) = p(A|C)\ p(B|C),$$

and this would have to hold no matter how plausible or implausible B might be. So our function p(x) has to have the property that certainty must always be represented by p = 1.

Now suppose that A is impossible, given C. In this case, the proposition AB is also impossible given C:

$$(AB|C) = (A|C)$$

and if A is already impossible given C, then if we had been given B also, A would still be impossible:

$$(A|BC) = (A|C).$$

In this case, equation (3-3) reduces to

$$p(A|C) = p(B|C)\ p(A|C) \qquad (3\text{-}4)$$

and again this equation would have to hold no matter what plausibility B might have. Well, there are two possible values of $p(A|C)$ that might satisfy this condition. It could be zero or plus infinity. The choice minus infinity can be ruled out [see what happens in (3-4) if B also becomes impossible], but at present there's nothing to tell us to choose zero rather than plus infinity; either one is equally good.

All right, let's sum up what we know about p(x) so far. It is a continuous monotonic function. It may be either increasing or decreasing. If it's an increasing function, it must range from zero for impossibility up to one for certainty; if it's a decreasing function, it must range from one for certainty up to infinity for impossibility. The way in which it varies between these limits, of course, our rule says nothing at all about.

3.2 Deduction of Rules 2 and 3.

Now there are still other conditions of consistency which these rules must satisfy. Let me introduce another notation. By a small letter I'll mean the denial of the big letter. In other words, proposition a stands

for the proposition "A is false." Conversely, A stands for the proposition "a is false." Most of the literature follows the notation of Boole, who indicated denial by placing a bar over the letter. This is fine except that it's a little hard to do reproducibly on a typewriter, so I've taken the liberty of changing it in a way that makes typed notes easier to produce, and less ambiguous to the reader. Actually, we will have little use for this notation beyond the present derivation; so it hardly matters.

Because of the fact that these propositions are of the type which must be either true or false, we see that the logical product aA is always false, and the logical sum a+A will always be true. Now the plausibility of a, given some data B, depends in some reciprocal way on the plausibility of A; if we define $x \equiv p(A|B)$, $y \equiv p(a|B)$, then

$$y = S(x). \tag{3-5}$$

Evidently, if this is going to agree qualitatively with common sense, the function S(x) must be some continuous monotonic decreasing function. But the relation between propositions a and A is a symmetrical one; it doesn't matter which I choose to call a capital letter and which the small letter. I can equally well say that

$$x = S(y). \tag{3-6}$$

It would have to be the same function. So S(x) must satisfy a functional equation that when we apply it twice we get back to where we started:

$$S[S(x)] = x \tag{3-7}$$

Now this alone is not enough to tell us much about this function. It says only that the graph of y = S(x) has mirror reflection symmetry about the line y = x. So now I'd like to give you another argument. There's another condition which S(x) will have to satisfy in order to represent a consistent way of reasoning, and for this we already have one rule of calculation worked out:

$$p(AB|C) = p(B|C)\ p(A|BC) \tag{3-8}$$

We'll call this Rule 1 from now on. Now we can make this step:

$$p(AB|C) = p(B|C)\ S[p(a|BC)]$$

but Rule 1 also says that $p(aB|C) = p(B|C)\ p(a|BC)$, and so

$$p(AB|C) = p(B|C)\ S\left\{\frac{p(aB|C)}{p(B|C)}\right\} \tag{3-9}$$

This looks like a very strange thing to do. But notice that the quantity we started with involved A and B in a symmetric way. If I interchange A and B, I don't change $p(AB|C)$. Therefore, although it doesn't look like it at all, this final expression must also be symmetric in A and B. In other words,

$$p(A|C)\ S\left\{\frac{p(bA|C)}{p(A|C)}\right\} = p(B|C)\ S\left\{\frac{p(aB|C)}{p(B|C)}\right\} \tag{3-10}$$

These two expressions must be equal no matter what propositions A, B, and C are. In particular, they must be equal when the denial of B is the same as the proposition "both A and D are true," that is, when $b = AD$, or

$$B = a + d.$$

But in that particular case, equation (3-10) simplifies. If B has this meaning, then what is $p(bA|C)$? Well, b is the statement that A is true and also that D is true. But this means that $bA = ADA = AD = b$; the propositions bA and b are the same, in the sense that they have the same "truth value." One of them is true if and only if the other is true. Therefore, they must have the same plausibility:

$$p(bA|C) = p(b|C) = S[p(B|C)].$$

Likewise, $aB = a(a+d) = a + ad = a$; in other words, aB and a are the same proposition in the sense that they have the same truth value, and so

$$p(aB|C) = p(a|C) = S[p(A|C)]$$

Substituting these into (3-10), we get a rather awful looking functional equation:

$$x\, S\left[\frac{S(y)}{x}\right] = y\, S\left[\frac{S(x)}{y}\right] \tag{3-11}$$

Here is another functional equation which has to be satisfied in order to have a consistent set of rules for reasoning.

At this point, we will simply turn again to the paper by R. T. Cox (Cox, 1946), or to his later book (Cox, 1961), which solves this problem. He shows that the only twice differentiable function which satisfies all of our conditions is

$$S(x) = (1 - x^m)^{1/m}.$$

and you easily verify that this does satisfy (3-7) and (3-11). This means that our reciprocal relation between the proposition and its denial would then have to take the form

$$p^m(a|B) + p^m(A|B) = 1. \tag{3-12}$$

m can be any constant except zero. I might say that I'm not entirely satisfied with the argument that we went through to get this; not because I think it's wrong, but because I think it's too long. The final result we get is so simple that there must be a simpler way of deriving it; but I haven't found it.

Now suppose that we make the choice that p = 0 is going to represent impossibility. In that case, we'll have to choose m as a positive number in order that (3-12) can be satisfied; but notice that choosing different values of m is really idle, because the only condition on this function p is that it is a continuous monotonic function which increases from zero to one as we go from impossibility to certainty. But if $p_1(x)$ satisfies these conditions, then $p_2(x) \equiv [p_1(x)]^m$ also satisfies them. So the statement that we could use different values of m doesn't give us any freedom that we didn't already have in the fact that p(x) was an arbitrary monotonic function. This means that if I choose to write equation (3-12) in the form

$$p(a|B) + p(A|B) = 1 \qquad (3\text{-}13)$$

this is just as general.

On the other hand, we could represent impossibility by $p = \infty$. In that case, we would have to choose m negative. Once again, to say that we can use different values of m wouldn't say anything that wasn't already implied by the fact that p was an arbitrary monotonic function which increased from one to infinity as we went from certainty to impossibility. So I could equally well write this reciprocal law in the form

$$\frac{1}{p(a|B)} + \frac{1}{p(A|B)} = 1.$$

Now we could go through our entire theory of the design of this robot's brain with the choice of $p = \infty$ to represent impossibility, and we would not get stopped any place. Everything would go through just fine. We would end up with equations which don't look quite so familiar to you as the ones that the other choice will give us. But notice that they're not different theories, because if $p_1(x)$ is a possible choice which goes to plus infinity to represent impossibility, then

$$p_3(x) = \frac{1}{p_1(x)}$$

is a function which represents impossibility by zero, and has all the properties that we needed. So regardless of which choice I make to represent impossibility, it makes the form of equations look different but their content will be exactly the same. You can go from one to the other simply by replacing all p's by the reciprocals of the p's. So if we agree not to use this choice of $p = \infty$ and always to use the choice $p = 0$ to represent impossibility, we're not throwing away any possibility of representation as far as content is concerned. We're just removing a redundancy in how you could have stated the theory. Let us agree, then, to use the choice:

$$0 \le p \le 1.$$

(for impossibility) (for certainty)

You recognize, of course, that this equation (3-13)

$$p(a|B) + p(A|B) = 1$$

which we henceforth call Rule 2, plus our Rule 1

$$p(AB|C) = p(B|C)\ p(A|BC)$$

are actually the fundamental equations of probability theory. Everything in probability theory follows from those by sufficiently complicated arguments. For example, I'd like to get the formula for

$$p(A + B|C),$$

the plausibility that at least one of the propositions A or B would be true, given C. This follows from the rules we already have; we just apply Rule 1 and Rule 2 over and over again:

$$\begin{aligned} p(A + B|C) &= 1 - p(ab|C) \\ &= 1 - p(a|bC)\ p(b|C) \\ &= 1 - [1 - p(A|bC)]\ p(b|C) \\ &= p(B|C) + p(Ab|C) \\ &= p(B|C) + p(b|AC)\ p(A|C) \\ &= p(B|C) + p(A|C)\ [1 - p(B|AC)]. \end{aligned}$$

Finally, we get

$$p(A + B|C) = p(A|C) + p(B|C) - p(AB|C). \qquad (3\text{-}14)$$

At long last we come out with the above form. And it's this result that I will take as our Rule 3.

We can summarize what we have learned up to this point by writing down our fundamental rules:

Rule 1: $p(AB|C) = p(A|BC)\ p(B|C) = p(B|AC)\ p(A|C)$ (3-15)

Rule 2: $p(A|B) + p(a|B) = 1$ (3-16)

Rule 3: $p(A + B|C) = p(A|C) + p(B|C) - p(AB|C)$ (3-17)

Rule 1, of course, involves A and B in a symmetric way and we could have interchanged A and B in all the argument leading up to it, so we have the liberty of writing it with A and B interchanged, as shown.

3.3 Deduction of Rule 4.

We've found so far the most general consistent rules by which our robot can manipulate plausibilities, granted that he must associate them with real numbers in some way so that his brain can operate by the carrying out of a definite physical process, and we are encouraged by the familiar appearance of these rules. But there are two evident circumstances which show that our job isn't yet finished. In the first place, while Rules 1, 2, and 3 show 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'd be a different set of rules.

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 his calculations. How is the robot to make his initial encoding of the given information, into definite numerical values of plausibilities?

The following analysis answers both of these questions, in a way that I think you will find both interesting and unexpected. Let's 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 Rule 3, 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 Rule 3 again to the first and third of these expressions, we obtain seven terms which can be grouped as follows:

$$\begin{aligned} p(A_1+A_2+A_3|B) &= p(A_1|B) + p(A_2|B) + p(A_3|B) \\ &\quad - p(A_1A_2|B) - p(A_2A_3|B) - p(A_3A_1|B) \\ &\quad + p(A_1A_2A_3|B) \end{aligned} \tag{3-18}$$

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

$$p(A_iA_j|B) = p(A_i|B)\delta_{ij} \tag{3-19}$$

where δ_{ij} is the Kronecker delta

$$\delta_{ij} \equiv \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} .$$

If the A_i are mutually exclusive, then the last four terms of (3-18) vanish, and we have

$$p(A_1+A_2+A_3|B) = p(A_1|B) + p(A_2|B) + p(A_3|B) \tag{3-20}$$

Adding more propositions A_4, A_5, etc., it is easy to show by induction that if we have n mutually exclusive propositions $\{A_1 \ . \ . \ . \ A_n\}$, (3-20) generalizes to

$$p(A_1+ \ . \ . \ . \ +A_m|B) = \sum_{i=1}^{m} p(A_i|B), \quad m \leq n \tag{3-21}$$

a rule which we will be using constantly from now on. In conventional expositions, Eq. (3-21) is usually introduced directly as one of the basic axioms of the theory, without any attempt to demonstrate its uniqueness or consistency. The present approach shows that this rule is deducible from simpler relations, which in essence represent the conditions for this theory to be consistent in the sense (a) given in Sec. 2.3.

Now suppose that the propositions $\{A_1 \ . \ . \ . \ A_n\}$ are not only mutually exclusive but also exhaustive; i.e., on data B one and only one of them

must be true. In that case the sum (3-21) over all of them must be unity:

$$\sum_{i=1}^{n} p(A_i|B) = 1 \tag{3-22}$$

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 $p(A_i|B)$ can be a difficult problem.

There is, however, one case in which the answer is particularly simple, requiring only direct application of principles already given. But we are now entering a very delicate and crucial area which has caused trouble and controversy for over a century; so I ask that you suppress all intuitive feelings that you may have, and contemplate the following logical analysis minutely. The point we are about to make cannot be developed too carefully; and unless it is clearly understood, you 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 $\{A_1 \ . \ . \ . \ A_n\}$ and we seek to evaluate $p(A_i|B)$. 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 $\{A_1' \ . \ . \ . \ A_n'\}$, given by

$$\begin{aligned} A_1' &= A_2 \\ A_2' &= A_1 \\ A_k' &= A_k, \quad k \geq 3 \end{aligned} \tag{3-23}$$

and we seek to evaluate the quantities $p(A_i'|B)$, $i = 1, 2, \ . \ . \ ., n$.

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, he must have the same state of knowledge about A_2' in problem II, for they are the same proposition, his given information B is the same in both problems, and he is contemplating the same totality of propositions $\{A_1 \ . \ . \ . \ A_n\}$ in both problems. Therefore we must have

$$p(A_1|B)_I = p(A_2'|B)_{II} \tag{3-24}$$

and similarly

$$p(A_2|B)_I = p(A_1'|B)_{II} \tag{3-25}$$

We will call these the transformation equations. What we have just done may appear utterly trivial to you, but bear with me; this line of reasoning, as Professor Eugene Wigner has aptly remarked (Wigner, 1959), consists of a number of steps each of which appears trivial in itself, but which in their totality are far from trivial. At this point, note that the transformation equations (3-24), (3-25) 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., it gives the robot no reason to prefer either over the other. In this case, problems I and II are entirely equivalent; i.e., he is in exactly the same state of knowledge about the set of propositions $\{A_1' \ . \ . \ . \ A_n'\}$ in problem II, including their labeling, as he is about the set $\{A_1 \ . \ . \ . \ A_n\}$ in problem I.

Now we invoke our requirement of consistency in the sense (b) as given above (Sec. 2.3). This stated that, in two equivalent problems, where the robot has the same state of knowledge, he must assign the same plausibilities. In equations, this statement is

$$p(A_i|B)_I = p(A_i'|B)_{II}, \quad i = 1,2,\ldots,n \tag{3-26}$$

which we will call the equivalence equations. But now, combining equations

(3-24), (3-25), (3-26), we obtain

$$p(A_1|B)_I = p(A_2|B)_I \tag{3-27}$$

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 (3-27). You recall that I asked you to suppress whatever intuitive feelings you may have, and allow yourself to be guided solely by the logical analysis. We will discuss the reasons for this presently; but first let us extend the result. More generally, let $\{A_1" \ . \ . \ . \ A_n"\}$ be any permutation of $\{A_1 \ . \ . \ . \ A_n\}$ and let Problem III be that of determining the $p(A_i"|B)$. If the permutation is such that $A_i = A_k"$, there will be n transformation equations of the form

$$p(A_i|B)_I = p(A_k"|B)_{III} \tag{3-28}$$

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 $\{A_1" \ . \ . \ . \ A_n"\}$ in problem III as he was about the set $\{A_1 \ . \ . \ . \ A_n\}$ in problem I; and again our desideratum of consistency demands that he assign equivalent distributions in equivalent problems, leading to the n equivalence equations

$$p(A_k|B)_I = p(A_k"|B)_{III}, \quad k = 1, 2, \ldots, n \tag{3-29}$$

From (3-28) and (3-29) we obtain n equations of the form

$$p(A_i|B)_I = p(A_k|B)_I \tag{3-30}$$

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 (3-30). 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). Since the $\{A_1 \ldots A_n\}$ are exhaustive, Eq. (3-22) will hold, and the only possibility is therefore

$$p(A_i|B)_I = \frac{1}{n}, \quad i = 1, 2, \ldots, n \tag{3-31}$$

and we have finally arrived at a set of definite numerical values. We will call this result Rule 4.

Perhaps you intuition had already led you to just this conclusion, without any need for the rather tortuous reasoning we have been through. If so, fine; then your intuition is consistent with our axioms. But merely writing down (3-31) intuitively does not give one a full appreciation of the importance and uniqueness of this result.

To see this importance, note that Eq. (3-31) 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 $p(A_i|B)$, and <u>not</u> the numerical values of the plausibilities $(A_i|B)$ that we started with. This, also, will be found to be true in general. But recognizing this gives us a beautiful answer to the first question posed at the beginning of this Section; after having found Rules 1, 2, and 3 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.

But now we see that no matter what function $p(x)$ we choose, we would still be led to the same result (3-31), 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 Rules 1, 2, and 3 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 you could design the robot's memory circuits. For each proposition A_i about which he is to reason, he will need a storage register in which he enters some number representing the degree of plausibility of A_i, on the basis of all the data he has been given. Of course, instead of storing the number p he could equally well store any monotonic function of p. But no matter what function he used internally, the externally observable behavior of the robot would be exactly 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 is an arbitrary monotonic function of p, defined in the interval $0 \leq p \leq 1$; it is p that is rigidly fixed by the data of a problem. The question of uniqueness is therefore disposed of automatically by the result (3-31); 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 $x \equiv (A|B)$ that we started with 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 name; from now on, we will call these quantities probabilities. I have studiously avoided using the word "probability" in our derivations up to this point, because while the word does have a colloquial meaning to the "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 such unique 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.

This situation is analogous to that in thermodynamics, where out of all possible temperature scales, which are monotonic functions of each other, we finally decide to use the Kelvin scale; not because it is more "correct" than others but because it is more convenient; i.e., the laws of thermodynamics and statistical mechanics take the simplest form in terms of this particular temperature scale.

3.4 Philosophical Digression.

For historical reasons, we still need quite a long discussion of Rule 4, Eq. (3-31). There seem to be only two kinds of people working in probability theory: those who consider Rule 4 to be so utterly trivial and obvious as to be in no need of any proof; and those who regard it as such a foolish and unjustified piece of metaphysical nonsense as to discredit anyone who uses it.

As far as I have been able to determine, there is no middle ground between these opinions; in the past, every writer on probability theory has been an extremist on one side or the other. I myself was an extremist of the first genre for some twenty years, and it was only recently that more mature reflection finally made me realize that Rule 4 is in need of logical demonstration. More important, it now appears to me that the method

of reasoning we have used to find Rule 4 is fundamental to all of probability theory, almost every present application requiring it to give a full logical justification of the result.

The reasoning we have just used is the most rudimentary example of the general group-theoretical approach which has been used with great success in theoretical physics for some forty years (Wigner, 1959). I had been teaching the use of group-theoretical methods for finding solutions of differential equations and boundary-value problems for sixteen years, without realizing that this same technique is the key to several deep unresolved issues in probability theory.

Rule 4 is itself fundamental to all of probability theory; although some will deny it, I don't think I am exaggerating when I assert that there is no known application of probability theory in which Rule 4 is not needed at one place or another. Those who profess to dislike it merely find some way of disguising the fact that they are using it; I will cite some specific examples in a later lecture. To understand this, we have to study the history of probability theory.

Rule 4 appears to have been first stated explicitly by James Bernoulli at the end of the seventeenth century (although it was, of course, implicit in the still earlier work of Cardano and Pascal). In the old literature it is often called the "Principle of Insufficient Reason," and it was used and defended by Laplace on the grounds that, on the given information, there was "no reason to think otherwise." This terminology and reasoning have been most unfortunate--I am tempted to say tragic--for the development of probability theory, because it has created a psychological block which has prevented many from seeing the real point of Rule 4.

But note that, in view of our derivation, we are asserting the validity of Rule 4, not for the weak and negative reason given by Laplace, but for

the strong and positive reason that it is uniquely determined by elementary requirements of consistency. In the state of knowledge defined by B in (3-31), if the robot were to assign any probability distribution other than the uniform one, then by a mere permutation of labels we could exhibit a second problem in which the robot has exactly the same state of knowledge, but in which he is assigning a different probability distribution. It just would not make sense, then, to say that the distribution described the robot's state of knowledge, or to claim that he is behaving in a consistent way.

But there is still a mystery here. For, no matter what method of reasoning we use, how is it possible that otherwise rational and mathematically competent people could be in violent disagreement on such an apparently simple matter as Equation (3-31)? I think that we have been caught in a semantic trap of our own making; to explain this, let me try to state the position of both extremists.

The extremist of the first camp says, "If the information B gives the robot no reason to prefer any of the propositions A_i over any other, then these propositions must appear equally likely to him; there is obviously no other thing he can possibly do but to assign them equal probabilities by Eq. (3-31). To do anything else would be to jump to conclusions not warranted by the data."

The extremist of the second camp says, "If the information B merely gives the robot no reason to prefer any proposition over another, this provides absolutely no justification for supposing them to be equally likely; they might not be equally likely at all. Unless the information B contains positive evidence that they <u>are</u> equally likely, the problem is simply not well-posed; and to write Eq. (3-31) is to jump to conclusions not warranted by the data."

Perhaps I have not, in spite of some effort, managed to verbalize these two positions in the most felicitous way; but I think you will grant that a more expert verbalizer could make either of these positions seem highly convincing, so at least from a psychological standpoint we can understand how there can be two diametrically opposing camps on this issue.

But, to be more constructive, what is the source of the difference? If you study these two statements, I think you'll agree that it is semantic; the phrase "equally likely" has two entirely different meanings in the two camps. In camp 2, the statement, "A_1 and A_2 are equally likely" is taken to describe a property <u>of the propositions</u> which is either true or false in an objective sense independently of the state of knowledge you or I--or the robot--might have about them. With that interpretation, of course, we have no justification for assuming this property to exist unless there is positive evidence for it.

In camp 1, the statement, "A_1 and A_2 are equally likely" is not regarded as describing any property of A_1 and A_2. In fact, each proposition is, in an objective sense, either true or false; and the only reason for using probability theory is that we are not in a position to say which. In writing Eq. (3-31), we are asserting nothing whatever about the propositions; we are describing only the <u>state of knowledge of the robot</u>.

Now you can, if you like, make value judgments as to which of these interpretations is the more desirable. But this has already been done quite enough to show that arguments on that level are futile. Debate on this issue has been going on more or less furiously in the literature of probability theory since the time of Laplace, one camp and then the other gaining a momentary ascendancy in numbers. But I think you will agree that we have here an issue that can never be settled by philosophical arguments about the meaning of words; much less by taking votes. We are in a situation very

much like the scientist who must decide between two rival theories of physics; and it has taken the human race thousands of years to realize that the only real, objective criterion for deciding such matters is the pragmatic one: casting aside all philosophical or ideological considerations, which viewpoint leads to a theory with the widest range of useful applications?

Therefore, I don't intend to waste any more time on the issue at this point; it is a major objective of these lectures to examine the problem on just the above pragmatic grounds. We are going to study a wide range of problems, covering almost all present applications of probability theory; and whenever possible we will exhibit the actual calculations, and final results, that the two viewpoints lead to.

It is perhaps already clear that viewpoint 1 is more widely applicable; there are many problems which our robot can undertake at once starting from Rule 4, but which on viewpoint 2 are ill-posed, offering no basis for applying probability theory. Now of course, a human statistician belonging to camp 2 may simply refuse to work on a problem (possibly at the cost of his job) if the information available is not as complete as he would like; but our robot is not free to do this, because the whole point of designing him is that he is to do the best he can whatever the information at hand. The issue will then be: in such problems, does the robot arrive at useful and defensible conclusions?

Of course, if the given information is too vague to justify any definite conclusions, we will want the robot to recognize this and tell us that more data are needed. His way of doing this will be to give us a final probability distribution that is very broad, indicating no strong preference for one conclusion over another. If the data do justify definite conclusions, he will find very sharply peaked final distributions, and report, "The data you gave me point to conclusion C as overwhelmingly the most likely to be

correct." And, of course, the robot should have some way of interpolating between these extremes, where most of the really interesting problems of the theory lie.

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. But it is just the function of our consistency requirements to make these probability assignments completely "objective" in the sense that they are independent of the personality of the user; i.e., they are a means of describing (or if you like, of encoding) the given information, independently of whatever personal feelings you or I might have. It is "objectivity" in this sense that is needed for a scientifically respectable theory of plausible reasoning.

The job before us now is, therefore, not to engage in philosophical disputation, but to put our robot to the test by examining just what he will do if he reasons by applying Rules 1 - 4 and their generalizations that we will develop as needed.