How Does the Brain do Plausible Reasoning?

By

E. T. Jaynes

Microwave Laboratory and Department of Physics

Stanford University, Stanford, California

ABSTRACT

We start from the observation that the human brain does plausible reasoning in a fairly definite way. It is shown that there is only a single set of rules for doing this which is consistent and in qualitative correspondence with common sense. These rules are simply the equations of probability theory, and they can be deduced without any reference to frequencies.

We conclude that the method of maximum-entropy inference and the use of Bayes' theorem are statistical techniques fully as valid as any based on the frequency interpretation of probability. Their introduction enables us to broaden the scope of statistical inference so that it includes both communication theory and thermodynamics as special cases.

The program of statistical inference is thus formulated in a new way. We regard the general problem of statistical inference as that of devising new consistent principles by which we can translate "raw" information into numerical values of probabilities, so that the Laplace-Bayes model is enabled to operate on more and more different kinds of information. That there must exist many such principles, as yet undiscovered, is shown by the simple fact that our brains do this every day.

1. INTRODUCTION

Shannon's theorem 2, in which the formula $H(p_1 \ldots p_n) = - \sum p_i \log p_i$ is deduced,[1] is a very remarkable argument. He shows that a qualitative requirement, plus the condition that the information measure be consistent, already determines a definite mathematical function. Actually, this is not quite true, because he chooses the condition of consistency (the composition law) in a particular way so as to make H additive. Any continuous differentiable function $f(H)$ for which $f'(H) > 0$ would also satisfy the qualitative requirements and a different, but equally consistent, composition law. Thus a qualitative requirement plus the condition of consistency determines the function H only to within an arbitrary monotonic function. The content of communication theory would, however, be exactly the same regardless of which monotonic function was chosen. Shannon's H thus involves also a convention which leads to simple rules of combination.

This interesting situation led the writer to ask whether it might be possible to deduce the entire theory of probability from a qualitative requirement and the condition that it be consistent. It turns out that this is indeed possible. In terms of the resulting theory we are enabled to see that communication theory, thermodynamics, and current practice in statistical inference, are all special cases of a single principle of reasoning.

In developing this theory we find ourselves in the fortunate position of having all the hard work already done for us. The methodology has been supplied by Shannon, the necessary mathematics has been worked out by Abel[2] and Cox,[3] and the qualitative principle was given by Laplace.[4] All we have to do is fit them together.

Laplace's qualitative principle is his famous remark[4] that "Probability theory is nothing but common sense reduced to calculation." The main

object of this paper is to show that this is not just a play on words, but a literal statement of fact.

One of the most familiar facts of our experience is this: that there is such a thing as common sense, which enables us to do plausible reasoning in a fairly consistent way.[5,6] People who have the same background of experience and the same amount of information about a proposition come to pretty much the same conclusions as to its plausibility. No jury has ever reached a verdict on the basis of pure deductive reasoning. Therefore the human brain must contain some fairly definite mechanism for plausible reasoning, undoubtedly much more complex than that required for deductive reasoning. But in order for this to be possible, there must exist consistent rules for carrying out plausible reasoning, in terms of operations so definite that they can be programmed on the computing machine which is the human brain. This is the "experimental fact" on which our theory is based. We know that it must be true, because we all use it every day. Our direct knowledge about this process is, however, only qualitative in much the same way as is our direct experience of temperature. For that reason it is necessary to use the methodology of Shannon.

2. LAPLACE'S MODEL OF COMMON SENSE

We now turn to development of our first mathematical model. We attempt to associate mental states with real numbers which are to be manipulated according to definite rules. Now it is clear that our attitude toward any given proposition may have a very large number of different "coordinates." We form simultaneous judgments as to whether it is probable, whether it is desirable, whether it is interesting, whether it is amusing, whether it is important, whether it is beautiful, whether it is morally right, etc.

If we assume that each of these judgments might be represented by a number, a fully adequate description of a state of mind would then be represented by a vector in a space of a very large, and perhaps indefinitely large, number of dimensions. Not all propositions require this. For example, the proposition, "The refractive index of water is 1.3," generates no emotions; consequently the state of mind which it produces has very few coordinates. On the other hand, the proposition, "Your wife just wrecked your new car," generates a state of mind with an extremely large number of coordinates. A moment's introspection will show that, quite generally, the situations of everyday life are those involving the greatest number of coordinates. It is just for this reason that the most familiar examples of mental activity are the most difficult ones to reproduce by a model. We might speculate that this is the reason why natural science and mathematics are the most successful of human activities; they deal with propositions which produce the simplest of all mental states. Such states would be the ones least perturbed by a given amount of imperfection in the human brain.

The simplest possible model is one-dimensional. We allow ourselves only a single number to represent a state of mind, and wish to discover how much of mental activity we can reproduce subject to that limitation. For the time being we call these numbers plausibilities, reserving the term "probability" for a particular quantity to be introduced later.

The way in which states of mind are to be reduced to numbers is at this stage very indefinite. For the time being we say only that greater plausibility must always correspond to a greater number, and we assume a continuity property which can be stated only imprecisely: infinitesimally greater plausibility should correspond only to an infinitesimally greater number.

We denote various propositions by letters A, B, C,··· . By the product AB we mean the proposition "Both A and B are true." The expression (A+B) is to be read, "At least one of the propositions A, B is true." The plausibility of any proposition A will in general depend on whether we accept some other proposition B as true. We indicate this by the symbol

$(A|B)$ = conditional plausibility of A, given B.

Thus, for example,

$(AB|C)$ = plausibility of A and B, given C.

$(A+B|CD)$ = plausibility that at least one of the propositions A, B is true, given that both C and D are true,

$(A|C) > (B|C)$ means that, on data C, A is more plausible than B.

In order to find rules for manipulation of these symbols, we are guided by two requirements:

1) The rules must correspond qualitatively to common sense. (2-1)

2) The rules must be consistent. This is used in two ways:

{ If a result can be arrived at in more than one way, we must obtain the same result for every possible sequence of operations on our symbols. } (2-2)

{ The rules must include deductive logic as a special case. In the limit where propositions become certain or impossible in any way, every equation must reduce to a valid example of deductive reasoning. } (2-3)

By a successful model we mean any set of rules satisfying these conditions. If we find that we have any freedom of choice left after imposing them, we can exercise that freedom to adopt conventions so as to make the rules as simple as possible. If we find that these requirements are so restrictive that there is in effect only one possible model satisfying them, are we entitled to claim that we have discovered the mechanism by which the brain does "one-dimensional" plausible reasoning? Except for the proviso that the human mind is imperfect, it seems that to deny that claim would be to assert that the human mind operates in a deliberately inconsistent way.

We now 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 $(AB|C)$. Now in order for AB to be true on data C, it is first of all necessary that B be true; thus the plausibility $(B|C)$ must be involved. If B is true, it is further necessary that A be true; thus $(A|BC)$ is needed. If, however, B is false, then AB is false independently of any statement about A. Therefore $(A|C)$ is not needed; it tells us nothing about AB that we did not already have in $(A|BC)$. 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 where we know from the start that C is true.

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

Thus, we seek some function $F(x,y)$ such that

$$(AB|C) = F[(A|BC),(B|C)]. \tag{2-4}$$

It is easy to exhibit special cases which show that no relation of the form $(AB|C) = F[(A|C),(B|C)]$, or of the form $(AB|C) = F[(A|C),(A|B),(B|C)]$, could satisfy conditions (2-1), (2-2), (2-3).

Condition (2-1) imposes the following limitations on the function $F(x,y)$. An increase in either of the plausibilities $(A|BC)$ or $(B|C)$ must never produce a decrease in $(AB|C)$. Furthermore, $F(x,y)$ must be a continuous function, otherwise we could produce a situation where an arbitrarily small increase in $(A|BC)$ or $(B|C)$ still results in the same large increase in $(AB|C)$. Finally, an increase in either of the quantities $(A|BC)$ or $(B|C)$ must always produce <u>some</u> increase in $(AB|C)$, unless the other one happened to represent impossibility. Thus condition (2-1) requires that

$$\left\{\begin{array}{l} F(x,y) \text{ must be a continuous function, with } \dfrac{\partial F}{\partial x} \geq 0 \\ \text{and } \dfrac{\partial F}{\partial y} \geq 0. \text{ The equality sign can apply only when} \\ (AB|C) \text{ represents impossibility.} \end{array}\right\} \qquad (2\text{-}5)$$

The condition of consistency (2-2) places further limitations on the possible form of the function $F(x,y)$. For we can calculate $(ABD|C)$ from (2-4) in two different ways. If we first group AB together as a single proposition, two applications of (2-4) give us

$$(ABD|C) = F[(AB|DC),(D|C)] = F\{F[(A|BDC),(B|DC)],(D|C)\}.$$

But if we first regard BD as a single proposition, (2-4) leads to

$$(ABD|C) = F[(A|BDC),(BD|C)] = F\{(A|BDC),\ F[(B|DC),(D|C)]\}.$$

Thus, if (2-4) is to be consistent, $F(x,y)$ must satisfy the functional equation

$$F[F(x,y),z] = F[x,F(y,z)]. \qquad (2\text{-}6)$$

Conversely, it is easily shown by induction that if (2-6) is satisfied, then (2-4) is automatically consistent for all possible ways of finding any number of joint plausibilities, such as $(ABCDEF|G)$. This functional equation turns out to be one which was studied by N. H. Abel.[2] Its solution, given also by Cox,[3] is

$$p[F(x,y)] = p(x)\ p(y), \tag{2-7}$$

where $p(x)$ is an arbitrary function. By (2-5) it must be a continuous monotonic function. Therefore our rule necessarily has the form

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

which we will also write, for brevity, as[7]

$$p(AB|C) = p(A|BC)\ p(B|C). \tag{2-8}$$

The condition (2-3) above places further restrictions on the function $p(x)$. Assume first that A is certain, given C. Then $(AB|C) = (B|C)$, and $(A|BC) = (A|C) = (A|A)$. Equation (2-8) then reduces to

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

and this must hold for all $(B|C)$. Therefore,

<u>Certainty must be represented by p = 1</u>. (2-9)

If for some particular degree of plausibility $(A|BC)$, the function $p(A|BC)$ becomes zero or infinite, then (2-8) says that $(B|C)$ becomes irrelevant to $(AB|C)$. This contradicts common sense unless $(A|BC)$ corresponds to impossibility. Therefore

<u>p cannot become zero or infinite for any degree of plausibility other than impossibility</u>. (2-10)

Now assume that A is impossible, given C. Then $(AB|C) = (A|BC) = (A|C)$, and (2-8) reduces to

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

which must hold for all $(B|C)$. There are three choices for $p(A|C)$ which satisfy this; $p(A|C) = 0$, or $+\infty$, or $-\infty$. But by (2-9) and (2-10) the choice $-\infty$ must be excluded, for any continuous monotonic function which has the values $+1$ and $-\infty$ at two given points necessarily passes through zero at some point between them. Therefore

Impossibility must be represented by $p = 0$, or $p = \infty$. (2-11)

Evidently the plausibility that A is false is determined by the plausibility that A is true in some reciprocal fashion. We denote the denial of any proposition by the corresponding small letter; i.e.

$$a = \text{"}A \text{ is false"}$$

$$b = \text{"}B \text{ is false"}$$

We could equally well say that $A =$ "a is false," etc. Clearly, $(A+a)$ is always true, and Aa is always false.

Since we already have some rules for manipulation of the quantities $p(A|B)$, it will be convenient to work with $p(A|B)$ rather than $(A|B)$. For brevity in the following derivation we use the notation

$$[A|B] \equiv p(A|B).$$

Now there must be some functional relationship of the form

$$[a|B] = S[A|B] \tag{2-12}$$

where by (2-1), $S(x)$ must be a monotonic, decreasing function. Since the propositions a and A are reciprocally related, we must have also

$$[A|B] = S[a|B]. \tag{2-13}$$

Therefore the function $S(x)$ must satisfy the functional equation

$$S[S(x)] = x. \tag{2-14}$$

To find another condition which $S(x)$ must satisfy, apply (2-8) and (2-12) alternately as follows:

$$[AB|C] = [A|BC][B|C] = S[a|BC][B|C] = [B|C]S\left\{\frac{[aB|C]}{[B|C]}\right\} \tag{2-15}$$

The original expression $[AB|C]$ is symmetric in A and B. So also, therefore, is the final expression; thus

$$[AB|C] = [A|C]\ S\left\{\frac{[bA|C]}{[A|C]}\right\}. \tag{2-16}$$

The expressions (2-15) and (2-16) must be equal whatever A, B, C, may be. In particular, they must be equal when $b = AD$. But in this case,

$$[bA|C] = [b|C] = S[B|C],$$
$$[aB|C] = [a|C] = S[A|C].$$

Substituting these into (2-15) and (2-16), we see that $S(x)$ must also satisfy the functional equation

$$x\ S\left[\frac{S(y)}{x}\right] = y\ S\left[\frac{S(x)}{y}\right]. \tag{2-17}$$

R. T. Cox[3] has shown that the only continuous differentiable function satisfying both (2-14) and (2-17) is

$$S(x) = (1 - x^m)^{1/m} \tag{2-18}$$

where m is any non-zero constant. Therefore the reciprocal relation between $[a|B]$ and $[A|B]$ necessarily has the form

$$[A|B]^m + [a|B]^m = 1. \tag{2-19}$$

Suppose we represent impossibility by $p = 0$. Then, from (2-19), m must be chosen positive. However, use of different values for m does not represent any freedom of choice that we did not already have in the arbitrariness of the function $p(x)$. The only condition on $p(x)$ is that it be a continuous monotonic function which increases from 0 to 1 as we go from impossibility to certainty. If the function $p_1(x)$ satisfies this condition, so also does the function

$$p_2(x) = [p_1(x)]^m .$$

Therefore if we write (2-19) in the form

$$p(A|B) + p(a|B) = 1 \tag{2-20}$$

in which $p(x)$ is understood to be an arbitrary monotonic function, Eq. (2-20) is just as general as is (2-19).

Suppose, on the other hand, that we represent impossibility by $p = \infty$. Then we must choose m negative. Once again, to say that we can use different values of m does not say anything that is not already said in the statement that $p(x)$ is an arbitrary monotonic function which increases from 1 to ∞ as we go from certainty to impossibility. The equation

$$\frac{1}{p(A|B)} + \frac{1}{p(a|B)} = 1 \tag{2-21}$$

is also just as general as (2-19).

An entire consistent theory of plausible reasoning can be based on (2-21) as well as on (2-20). They are not, however, different theories, for if $p_1(x)$ satisfies (2-21), the equally good function

$$p_2(x) = 1/p_1(x)$$

satisfies (2-20), and says exactly the same thing. If we agree to use only functions of type (2-20), we are not excluding any possibility of representation, but only removing a certain redundancy in the mathematics.

From (2-20) we can derive the last of our fundamental equations. We seek an expression for the plausibility of $(A+B)$, the statement that at least one of the propositions A, B is true. Noting that if $D = A+B$, then $d = ab$, we can apply (2-20) and (2-8) in alternation to get

$$p(A+B|C) = 1 - p(ab|C) = 1 - p(a|bC)\, p(b|C)$$
$$= 1 - [1 - p(A|bC)]\, p(b|C) = p(B|C) + p(Ab|C)$$
$$= p(B|C) + p(A|C)\, [1 - p(B|AC)]$$

or,

$$p(A+B|C) = p(A|C) + p(B|C) - p(AB|C)\,. \tag{2-22}$$

Equations (2-8) and (2-22) are the fundamental equations of the theory of probability. From them all other relations follow.

We have found that the most general consistent rules for plausible reasoning can be expressed in the form of the product and sum rules (2-8) and (2-22), in which $p(x)$ is an arbitrary continuous monotonic function ranging from 0 to 1. It might appear that different choices of the function $p(x)$ will lead to models with different content, so that we have found in effect an infinite number of different possible consistent rules for plausible reasoning. This, however, is not the case, for regardless of which function $p(x)$ we choose, when we start to use the theory we find that it is always p, not x, that has a definitely ascertainable numerical value. To demonstrate this in the simplest case, consider n propositions A_1, A_2, ..., A_n which are mutually exclusive; i.e., $p(A_iA_j|C) = p(A_i|C)\,\delta_{ij}$. Then repeated application of (2-22) gives the usual sum rule

$$p(A_1+\ldots+A_n|C) = \sum_{k=1}^{n} p(A_k|C). \tag{2-23}$$

If now the A_k are all equally likely on data C (this means only that data C gives us no reason to expect that one of them is more valid than the others), and one of them must be true on data C, the $p(A_k|C)$ are all equal and their sum is unity. Therefore we necessarily have

$$p(A_k|C) = \frac{1}{n} . \qquad (2\text{-}24)$$

This is Laplace's "Principle of Insufficient Reason." No matter what function $p(x)$ we choose, there is no escape from the result (2-24). Therefore, rather than saying that p is an arbitrary monotonic function of $(A|C)$, it is more to the point to say that $(A|C)$ is an arbitrary monotonic function of p, in the interval $0 \leq p \leq 1$. It is the connection of the numbers $(A|C)$ with intuitive states of mind that never gets tied down in any definite way. In changing the function $p(x)$, or better $x(p)$, we are not changing our model, but just displaying the fact that our intuitive sensations provide us only with the relation "greater than," not any definite numbers. Throughout these changes, the numerical values of, and relations between, the quantities p remain unchanged.

All this is in very close analogy with the concept of temperature, which also originates only as a qualitative sensation. Once it has been discovered that, out of all the monotonic functions represented by the readings of different kinds of thermometers, one particular definition of temperature (the Kelvin definition) renders the equations of thermodynamics especially simple, the obvious thing to do is to recalibrate the scales of the various thermometers so that they agree with the Kelvin temperature. The Kelvin temperature is no more "correct" than any other; it is simply more convenient. Similarly, the obvious thing for us to do at this point is to adopt the convention $p(x) \equiv x$, so that the distinction between a plausibility and the quantity p (which we henceforth call the probability) disappears. This means only that we have found a way of calibrating our "plausibility-meters" so that the consistent rules of reasoning take on a simple form. The content of the theory would, however, be exactly the same no matter what function $p(x)$ was chosen. Thus, there is only one consistent model of common sense.

From now on, we write our fundamental rules of calculation in the form

$$(AB|C) = (A|BC)(B|C) = (B|AC)(A|C) \tag{2-25}$$

$$(A+B|C) = (A|C) + (B|C) - (AB|C)\,. \tag{2-26}$$

Laplace's model of common sense consists of these rules, with numerical values determined by the principle of insufficient reason.

Out of all the propositions which we encounter in this theory, there is one which must be discussed separately. The proposition X stands for all of our past experience. There can be no such thing as an "absolute" or "correct" probability; all probabilities are conditional on X at least, and X is not only different for different people, but it is continually changing for any one person. If X happens to be irrelevant to a certain question, then this observation is unnecessary but harmless. We often suppress X for brevity, with the understanding that even when it does not appear explicitly, it is still "built into" all bracket expressions: $(A|B) \equiv (A|BX)$. Any probabilities conditional on X alone are called a-priori probabilities. In an a-priori probability we will always insert X explicitly: $(A|X)$.

It is of the greatest importance to avoid any impression that X is some sort of hidden major premise representing a universally valid proposition about nature; it is simply whatever initial information we have at our disposal for attacking the problem. Alternatively, we can equally well regard X as a set of hypotheses whose consequences we wish to investigate, so that all equations may be read, "If X were true, then - - - ." It makes no difference in the formal theory.

3. DISCUSSION

It is well known that criticism of the theory of Laplace, and pointing out of its obvious absurdity, has been a favorite indoor sport of writers on probability and statistics for decades. In view of the fact that we have just shown it to be the only way of doing plausible reasoning which is consistent and in agreement with common sense, it becomes necessary to consider the objections to Laplace's theory and if possible to answer them.

Broadly speaking, there are three points which have been raised in the literature. The first is that any quantity which is only subjective, i.e. which represents a "degree of reasonable belief," in Jeffreys' terminology,[8] cannot be measured numerically, and thus cannot be the object of a mathematical theory. Secondly, there is a widespread impression that even if this could be accomplished, a quantity which is different for different observers is not "real," and cannot be relevant to applications.[9] Thirdly, there is a long history of pathology associated with this view; it is tempting and easy to misuse it.

The latter is of course not a valid objection to any theory, and we need only answer the first two. The arguments of Sec. 2 almost answer the first, but there remains the question of finding numerical values of probabilities in cases where there is no apparent way of reducing the situation to one of "equally possible" cases. We must hasten to point out that the notion of "equally possible" has, at this stage, nothing whatsoever to do with frequencies. The notion of frequency has not yet appeared in the theory. Now the question of how one finds <u>numerical values</u> of probabilities is evidently an entirely different problem than that of finding a consistent <u>definition</u> of probability, and consistent rules for calculation.

In physics, after the Kelvin temperature is defined, there remains the difficult problem of devising experiments to establish its numerical value. Similarly, after our model has been set up, the problem of reducing "raw" information to a statement of probability numerical values remains.

Most of the objections to Laplace's theory which one finds in the literature[11] consist of applying it to some simple problem, and pointing out that the result flatly contradicts common sense. However, study of these examples will show that in every case where the theory leads to results which contradict common sense, the person applying the theory has additional information of some sort, relevant to the question being asked, but not actually incorporated into the equations. Then his common sense utilizes this information unconsciously and of necessity comes to a different conclusion than that provided by the theory.

Here is one of Polya's examples.[11] A boy is ten years old today. According to Laplace's law of succession, he has the probability 11/12 of living one more year. His grandfather is 70. According to the same law, he has the probability 71/72 of living one more year. Obviously, the result contradicts common sense. Laplace's law of succession, however, applies only to the case where we have absolutely no prior information about the problem.[13] In this example it is even more obvious that we do have a great deal of additional information relevant to this question, which our common sense used but we did not allow Laplace's theory to use.

Laplace's theory gives the result of consistent plausible reasoning on the basis of the information which was put into it. The additional information is often of a vague nature, but nevertheless highly relevant, and it is just the difficulty of translating it into numerical values which causes all the trouble. This shows that the human brain must have extremely

powerful means, the nature of which we have not yet imagined, for converting raw information into probabilities.

We can see from this why Laplace's theory was incomplete and why it will always remain incomplete. It is simply that there is no end to the variety of kinds of partial information with which we might be confronted, and therefore no end to the problem of finding consistent ways of translating that information into probability statements. Here again there is a close analogy with physics. Whenever research involving temperature extends into some new field, science is dependent on the ingenuity of experimenters in devising new procedures which will give the Kelvin temperature in terms of observed quantities. Physicists must continually invent new kinds of thermometers, and statisticians must continually invent new kinds of "plausimeters." Laplace's theory is incomplete in the same sense, and for the same reason, that physics is incomplete; but Laplace's basic model occupies the same fundamental position in statistics as do the laws of thermodynamics in physics.

The principle of insufficient reason is only one of many techniques which one needs in current applications of probability theory, and it needs to be generalized before it is applicable to a very wide range of problems.[14] In the following sections we will show two principles available for doing this. The first has been made possible by information theory, and the second comes from a relation between probabilities and frequencies.

Consider now the second objection, that a probability which is only subjective and different for different people cannot be relevant to applications. It seems to the writer that this is the exact opposite of the truth: <u>it is only a subjective probability which could possibly be relevant to applications</u>. What is the purpose of any application of probability theory?

Simply to help us in forming reasonable judgments in situations where we do not have complete information. Whether some other person may have complete information is quite irrelevant to our problem. We must do the best we can with the information we have, and it is only when this is incomplete that we have any need for probability theory. The only "objective" probabilities are those which describe frequencies observed in experiments already completed. Before they can serve any purpose in applications they must be converted into subjective judgments about other situations where we do not know the answer.

If a communication engineer says, "The statistical properties of the message and noise are known," he means only that he has some knowledge about the past behavior of some particular set of messages and some particular sample of noise. When he infers that some of these properties will hold also in the future and designs a communication system accordingly, he is making a subjective judgment of exactly the type accounted for by Laplace's theory, and the sole purpose of the statistical analysis of past events was to obtain that subjective judgment.

Two engineers who have different amounts of statistical information about messages will assign different n-gram probabilities and design different coding systems. Each represents rational design on the basis of the available information, and it is quite meaningless to ask which is "correct." Of course, the man who has more advance knowledge about what a system is to do will generally be able to utilize that knowledge to produce a more efficient design, because he does not have to provide for so many possibilities. This is in no way paradoxical, but just simple common sense.

Similarly, if a medical researcher says, "This new medicine is effective in 85 per cent of the cases," he means only that this is the frequency

observed in past experiments. If he infers that it will hold approximately in the future, he is making a subjective judgment which might be (and often is) entirely erroneous. Nevertheless, it was the most reasonable judgment he could have made on the basis of the information available. The judgment, and also its level of significance, are accounted for by Laplace's theory. Its conclusions are, for all practical purposes, identical with those provided by the method of confidence intervals,[15] and it is our contention that the validity of the latter method depends on this agreement.

4. THE PRINCIPLE OF INSUFFICIENT REASON

Two conditions are necessary before we can assign probabilities by means of the principle of insufficient reason:

> We must be able to analyze the situation into an enumeration of the different possibilities which we recognize as mutually exclusive and exhaustive. (4-1)

> Having done this, we must then find that the available information gives us no reason to prefer any possibility to any other. (4-2)

In practice these conditions are hardly ever met unless there is some evident element of symmetry in the problem, as is usually the case in games of chance. Note, however, that there are two different ways in which condition (4-2) may be satisfied. It may be the consequence of complete ignorance, or it may be the consequence of positive knowledge.

Suppose a person, known to be very dishonest, is going to toss a die. Observer A is allowed to examine the die, and he has at his disposal all the facilities of the National Bureau of Standards. He performs thousands of experiments with scales, calipers, microscopes, magnetometers, x-rays, neutron beams, etc., and finally is convinced that the die is perfectly symmetrical. Observer B is not told this; he knows only that a die is being tossed by a shady character. He suspects that it is biased, but has no idea in which direction. Condition (4-2) is satisfied for both, and they will both assign probability 1/6 to each face. The same probability assignment may describe either knowledge or ignorance. This seems paradoxical: why doesn't A's extra knowledge make any difference?

Well, it does make a difference, and a very important one, but the difference requires time to "develop." Suppose that the first toss gives a "3." To observer B this constitutes evidence that the die is biased to favor 3, and so on the second throw B will assign different probabilities which take this into account. Observer A, however, will continue to assign probability 1/6 to each face, because to him the evidence of symmetry carries overwhelmingly greater weight than does the evidence of one throw.

It is now fairly clear what will happen. To observer B, every throw of the die represents new evidence about its bias, which causes him to change his probability assignments for the next throw. Under certain circumstances, his assignments are given by a generalization of Laplace's law of succession. To observer A, the evidence of symmetry continues to carry greater weight than does the evidence of the random experiment, and he persists in assigning probability 1/6. Each observer has done consistent plausible reasoning on the basis of the information available to him, and

Laplace's theory accounts for the behavior of each (Sec. 6).

This difference in behavior is not, however, accounted for by any theory based on a frequency definition of probability, because when you define a probability simply as a frequency you deprive yourself of any way of saying that you have evidence unless it is in the form of an observed frequency. Everything which the National Bureau of Standards can tell us must be ignored, because it has no frequency interpretation.

5. THE ENTROPY PRINCIPLE

A biased die, colored black with white spots, has been tossed many times onto a black table, and we have recorded the experiment with a camera, obtaining a multiple exposure of uniform density. From the blackening of the film we cannot determine the relative frequencies of the different faces, but only the average number of spots which were on top. This average is not 3.5, as we might expect from an honest die, but 4.5. On the basis of this information, what are the probabilities for the different faces?

Automobiles of make i have weight W_i and length L_i. We observe a cluster of 1000 cars packed bumper to bumper, occupying a total length of 3 miles. As these cars pass an intersection they go over a machine which weighs each one and totals the result, not retaining the record of the individual weights. Therefore we have only the total length and total weight of the 1000 cars. What can we infer about the number of cars of each make in the cluster?

During an earthquake, 100 windows were broken into 1000 pieces. What is the probability for a window to be broken into exactly m pieces?

These are examples of problems where condition (4-1) is satisfied but not condition (4-2). They can be formulated in a general way as follows.

The quantity x can assume the discrete values $x_1 \ldots x_n$. There are k functions $f_1(x), \ldots, f_k(x)$ for which we know the average values

$$f_r = \sum_{i=1}^{n} p_i f_r(x_i) , \qquad 1 \leq r \leq k . \tag{5-1}$$

The problem is to find the p_i. If $k < (n-1)$, there are not enough conditions to determine the p_i in the sense of a mathematical solution of (5-1) and $\sum p_i = 1$. We cannot use the principle of insufficient reason because we have too much information; there <u>are</u> reasons for preferring some possibilities to others. There are many probability assignments which would all agree with the available information. Which is the most reasonable one to adopt?

Consider the third example above, and restate it as: the average window is broken into 10 pieces. If we were to conclude that <u>each</u> window is broken into 10 pieces, this would be in complete agreement with all the available information. However, our common sense tells us that it would not be a <u>reasonable</u> probability assignment; we would be assuming far more than was given in the statement of the problem. It is more reasonable to assign probability $p_m = 1/5$ for a window to be broken into m pieces, where $m = 8,9,10,11,12$. But this still assumes more than was warranted by the given information. It says, for example, that it is impossible for a window to be broken into 13 pieces. Evidently we regard a broad distribution as more reasonable than a sharply peaked one, and there is no value of m for which we would be justified in assigning $p_m = 0$.

To make a long story short, we want the probability assignment which assumes <u>nothing</u> beyond what was given in the statment of the problem. Shannon's theorem 2 tells us that the consistent measure of the "amount of uncertainty" in a probability distribution is its entropy, and therefore

we must choose the distribution which has maximum entropy subject to the constraints (5-1). Any other distribution would represent an arbitrary assumption of some kind of information which was not given to us. The maximum-entropy distribution is "maximally noncommittal" with respect to missing information.

The solution follows immediately from the method of Lagrangian multipliers, by arguments which are very well known in a different context. The results are expressed compactly if we define the partition function:

$$Z(\lambda_1 \ldots \lambda_k) = \sum_{i=1}^{n} \exp\left[- \lambda_1 f_1(x_i) - \ldots - \lambda_k f_k(x_i) \right] \tag{5-2}$$

Then the maximum-entropy distribution is

$$p_i = \exp\left[- \lambda_0 - \lambda_1 f_1(x_i) - \ldots - \lambda_k f_k(x_i) \right] \tag{5-3}$$

with the λ_r determined by

$$\lambda_0 = \log Z \tag{5-4}$$

$$\left\langle f_r(x) \right\rangle = - \frac{\partial}{\partial \lambda_r} \log Z, \qquad 1 \leq r \leq k \ . \tag{5-5}$$

At first glance it seems idle and trivial that we should have to do all this in order to learn how to say nothing. The important point, however, is that we have here found a consistent way of saying nothing in a new language; the language of probability theory. The triviality fades away entirely when we notice that the problem of inferring the macroscopic properties of matter from the laws of atomic physics is of exactly the type we are considering. All of thermodynamics, including the prediction of every experimentally reproducible feature of irreversible processes, is contained in the above solution.[16,17,18]

This is so easy to demonstrate that we will sketch the argument here. In any macroscopic experiment the exact microscopic state of a system is never under control or observation; there will be perhaps $10^{10^{20}} = (10^{10^{10}})^{10^{10}}$ different quantum states compatible with a given set of experimental conditions. Although the microscopic state is changing rapidly, the time required for any reasonably complete "sampling" of so many states is still rather long; perhaps $10^{10^{10}}$ years. When we repeat the experiment we will surely not repeat the microscopic state. Therefore, any property which is experimentally reproducible must be characteristic of <u>each</u> of the great majority of the class C_e of microscopic states allowed by the experimental conditions. This is not necessarily the same as the <u>subjective</u> class C_s consisting of all reasonably probable states in the maximum-entropy distribution.[19] Clearly, the only properties which we will be able to predict definitely from the maximum-entropy distribution will be those characteristic of the great majority of the states in class C_s.

Now if it is found that the class P_s of properties predictable by maximum-entropy inference is identical with the class P_e of experimentally reproducible properties, the theory is entirely successful. This would by no means imply that the class C_s is identical with the class C_e. If, however, the class P_s is found to differ in any way from the class P_e, we would be forced to conclude that $C_s \neq C_e$. But this could be true only if there exist new physical states, or new constraints on the possible physical states, which we did not take into account in our initial enumeration.

Therefore, strictly speaking, we should not assert that maximum-entropy inference <u>must</u> lead to correct predictions. But we can assert something

even more important: if the class of predictable properties is found to differ in any way from the class of experimentally reproducible properties, that fact would in itself demonstrate the existence of new laws of physics. Assuming that this occurs and the new laws are eventually worked out, then maximum-entropy inference based on the new laws will again have this property.

From this we see that maximum-entropy inference is precisely the appropriate tool for reasoning from the microscopic to the macroscopic. Its characteristic property is that it does not allow us to form any conclusions which are not indicated by the available evidence. Any other distribution would permit one to draw conclusions not warranted by the evidence.

Historically, maximum-entropy inference was discovered, in its mathematical aspects, by Boltzmann about 1870, and greatly advanced by Gibbs around 1900. The result is what the physicist calls statistical mechanics. However, the interpretation of the mathematical rules has always been a subject of great confusion, because of the illusion that probabilities must be given a frequency interpretation. This made it appear that the rules could be justified only by demonstrating a certain physical property called ergodicity, or in modern terms, metric transitivity. All attempts to demonstrate this have, however, failed. Until the discovery of Shannon's theorem 2, it was not possible to understand just what we were doing in statistical mechanics, or to have any confidence in it for the prediction of irreversible processes. However, we can now see that statistical mechanics is a much more powerful tool than physicists had realized.

6. PROBABILITY AND FREQUENCY

Although the word "frequency" has appeared a few times above, we have not so far made any use of it in developing the basic theory or in demonstrating its application to thermodynamics. This has been done deliberately in order to emphasize the fact that the notions of probability and frequency are entirely distinct. Many of the most important applications of probability theory can be justified and carried to completion without ever introducing the notion of frequency. However, in cases where a random experiment provides most or all of the available information, there should exist some relationship between the observed frequency of the event and the probability which we assign to it. Similarly, if an event can be regarded as a possible result of a random experiment, there may in some cases be a relation between the probability which we assign to it, and the relative frequency with which we expect it to occur. Such relations must, of course, be deduced from the theory and not postulated.

To demonstrate the latter relation, we introduce the propositions,

$$A_p \equiv \text{"The probability of } A \text{ in each case is } p." \tag{6-1}$$

$$N_n \equiv \text{"In } N \text{ trials, } A \text{ was (or will be) true } n \text{ times."} \tag{6-2}$$

The probability $(N_n|A_p)$, obtained immediately from the sum and product rules (2-26), (2-25), is the binomial distribution

$$(N_n|A_p) = \binom{N}{n} p^n (1-p)^{N-n}. \tag{6-3}$$

As a function of n, this attains a maximum value when n is within one unit of Np, so that the most probable frequency is substantially equal to the probability.

Note that the phrase "in each case," in (6-1) is essential. To demonstrate this, we look more closely at the derivation of (6-3) from our basic rules. Define the proposition

$$B_n \equiv \text{"A is true in the n'th trial."} \tag{6-4}$$

Now according to (2-25) we have

$$(B_2 B_1 | A_p) = (B_2 | B_1 A_p)(B_1 | A_p) ,$$

which reduces to $(B_2|A_p)(B_1|A_p) = p^2$ only if $(B_2|B_1A_p) = (B_2|A_p)$; i.e., the probability of A at the second trial which is involved in (6-3) is that based on A_p and knowledge of the result of the first trial. It is equal to p, as assumed in (6-3), only if knowing the result of the first trial would have given us no reason to change the assignment. This in spite of the fact that in (6-3) we are predicting a frequency entirely on the basis of A_p, since only A_p appears to the right of the vertical stroke. Even though we are not given the results of any trial, the expected frequency still depends on whether such knowledge would have been relevant.

This again corresponds to common sense. To take the most extreme case, suppose we are tossing a coin and A stands for "heads." Let it be a very dishonest coin, and define the proposition

$$C_p \equiv \text{"The coin has either two heads or two tails,}$$
$$\text{and the probability of the former is } p." \tag{6-5}$$

Now on the basis of this evidence alone, it is still true that the probability of "heads" in each particular throw is p. But no one expects the relative frequency of heads to be p: We now have $(B_2|B_1C_p) = 1$, so that

$$(B_2 B_1 | C_p) = (B_2 | B_1 C_p)(B_1 | C_p) = p$$

and by repeated applications of (2-25), we find that the only sequences of

N throws which do not have probability zero, correspond to

$$(B_N \ldots B_2 B_1 | C_p) = p$$
$$(b_N \ldots b_2 b_1 | C_p) = 1 - p$$

so that in place of (6-3) we have

$$(N_n | C_p) = p\,\delta(n,N) + (1 - p)\,\delta(n,0), \tag{6-6}$$

which is exactly what our common sense told us without any calculation.

This shows that <u>before we can infer any definite frequency from a probability assignment, the evidence on which that probability assignment is based must be very good evidence indeed</u>. It corresponds to that possessed by the man from the Bureau of Standards in the dice game of Sec. 3. In order for (6-3) to hold, the evidence on which A_p is based must carry overwhelmingly more weight than does the evidence of N throws. For this reason, the probabilities obtained from maximum-entropy inference have no reasonable frequency interpretation, and we can see why statistical mechanics was so confusing as long as we tried to interpret it this way.[18]

Now introduce the proposition,

$$D_f \equiv \text{"In an infinitely long sequence of trials, the relative frequency of } A \text{ approaches f."} \tag{6-7}$$

In the limit as $N \to \infty$, the binomial distribution becomes infinitely sharp, and so we obtain the Dirac delta-function[20]

$$(D_f | A_p) = \delta(f - p). \tag{6-8}$$

Equation (6-8) is loaded with logical booby-traps, which we must hasten to point out. Note first that it by no means says that the relative frequency $f = p$ <u>must</u> occur. It says only that, on the basis of the information which led to the assignment A_p, this is the only relative frequency which it is reasonable to expect; the available evidence gives no support

at all to any other value. The probability (6-8) is still only a subjective quantity.

Equation (6-8) represents a limiting case which can never be justified in practice, because in order for (6-3) to continue to hold as $N \longrightarrow \infty$, the evidence on which A_p is based must carry overwhelmingly more weight than do the results of an infinite number of trials. Not even the Bureau of Standards can provide us with evidence this good.

But there is still a paradox here. Suppose that the evidence A_p <u>was</u> perfectly reliable. It would still represent only partial information about the random experiment. According to (6-8), the probability that the limiting frequency lies in the interval $(p - \varepsilon) < f < (p + \varepsilon)$ is

$$\int_{p-\varepsilon}^{p+\varepsilon} (D_f | A_p)\, df = 1 \qquad (6\text{-}9)$$

i.e., f was certain, on data A_p, to lie in this interval. How could we have been certain of <u>anything</u> on the basis of only partial information? How could we have been certain that a limiting frequency even exists?

Well, Eq. (6-8) is actually a logical contradiction, but a useful one. We have asked the theory a foolish question, and it has given us a foolish answer. Equation (6-8) refers only to an infinite number of trials. If N is finite, there is no n in $0 \leq n \leq N$ for which $(N_n | A_p) = 0$. We are <u>not</u> certain of the result of any <u>possible</u> experiment. It is only when the experiment is <u>impossible</u> that we can be certain of the result! Any attempt to define a probability as the limit of a frequency is evidently subject to the same logical difficulty, but in a much more acute form, because there is no way at all of avoiding it.

In spite of this, (6-8) is useful if we understand how to use it. If N is large and the evidence A_p fairly good, it may be a perfectly valid approximation to (6-3) for some purposes, and it will then lead to simpler formulas than would (6-3).

Equation (6-8) can also be used in a different way. If we <u>had</u> evidence about limiting frequencies, that evidence would be equivalent to a perfectly reliable assignment A_p. Thus, if E is any proposition, and A_p is perfectly reliable so that (6-8) holds, we would have

$$(E|D_f) = (E|A_p), \qquad f = p.$$

In particular,

$$(N_n|D_f) = \binom{N}{n} f^n (1 - f)^{N-n} \tag{6-10}$$

which is the form used in the frequency theory.

The inverse problem, of inferring a probability from an observed frequency, is much more difficult. The quantity which we have here to evaluate is $(B_{N+1}|N_n X)$, where we denote, as in Sec. 2, the prior evidence by X. It does not seem possible to carry out this calculation once and for all in the most general case, because the prior evidence might provide intricate relations between the probabilities at different trials, in an infinite number of different ways. The order in which "A true" and a = "A false" occurred would in general be relevant to the probability of B_{N+1}, but the above notation implies that we are not going to consider that evidence.

The only case which the frequency school of thought can treat is the one where we ignore completely all the prior evidence; the frequency school regards a-priori probabilities as nonsense. This simplifies our problem,

because it is only that case that we need to exhibit here in order to establish the relation between the frequency theory and Laplace's theory. In other words, the prior evidence X is now to tell us nothing whatsoever. We have, from (2-25) and (2-26),

$$(B_{N+1}|N_n) = \int_0^1 (B_{N+1}D_f|N_n)\,df = \int_0^1 (B_{N+1}|D_fN_n)(D_f|N_n)\,df. \tag{6-11}$$

Also, by (2-25),

$$(D_f|N_n) = (D_f|X)\,\frac{(N_n|D_f)}{(N_n|X)}\,. \tag{6-12}$$

The a-priori probabilities $(D_f|X)$ and $(N_n|X)$ must now say nothing about the values of f or n. The consistent way of saying this is, from the principle of maximum entropy,

$$(D_f|X) = 1; \qquad (N_n|X) = \frac{1}{N+1}\,, \quad 0 \le n \le N.$$

Furthermore, the evidence D_f carries overwhelmingly more weight than does N_n, so that

$$(B_{N+1}|D_fN_n) = (B_{N+1}|D_f) = f.$$

Substituting these results and (6-10) into (6-11), we have

$$(B_{N+1}|N_n) = (N+1)\binom{N}{n}\int_0^1 f^{n+1}(1-f)^{N-n}\,df$$

$$= \frac{n+1}{N+2}\,, \tag{6-13}$$

which is Laplace's law of succession. If N is sufficiently large, the probability which we assign to A at the next trial is substantially equal to its observed frequency in the previous trials.

From these results we conclude that the general relation between the two theories is the following. Whenever all of the available evidence

consists of observed frequencies, the conclusions obtained from the frequency theory approach those given by Laplace's theory asymptotically as the number of observations increases. If we have additional evidence not expressible in terms of frequencies, the conclusions of the theories may differ widely, and it is Laplace's theory which will agree with common sense.

As a simple example of this, suppose that two observers listen to a geiger counter, known by both to have an efficiency of 10 per cent. O_1 has no knowledge about the source of the particles being counted. O_2 knows that the source is a radioactive sample of long lifetime, in a fixed position. He does not know anything about its strength except, of course, that it is not infinite. During the first minute, 10 counts are registered. O_1 infers, by maximum-likelihood, that about 100 particles actually passed through the counter, and O_2 agrees. During the second minute, 16 counts are registered. O_1 infers that about 160 particles were present, and he does not change his estimate for the first minute. O_2, using Bayes' theorem, concludes that the most probable value is only 137, and he revises his estimate for the first minute to 123. Each has done consistent plausible reasoning, but prior evidence which has no frequency interpretation can completely change the conclusions which we draw from random data, and their degree of reliability.

7. "SUBJECTIVE" COMMUNICATION THEORY

Laplace's theory is of such wide scope that in principle it includes every example of plausible reasoning, and thus a fortiori, communication theory. In particular, much of communication theory can be regarded as an

application of maximum-entropy inference. This viewpoint may or may not lead to new mathematical results unlikely to be found without it. However, the conditions for validity of some known results can be extended. Also, it clarifies a constantly recurring question: what parts of communication theory describe measurable properties of messages, and what parts describe only the state of knowledge of some observer?

The current tendency is to state and prove theorems using the frequency terminology. Mathematical properties needed for the proof must then be regarded as objective properties of the messages or noise, and this makes it appear that the theorem is valid only if these properties can be demonstrated as "true." For example, Shannon's proof of theorem 3 starts out with the statement; "We assume the source to be ergodic so that the strong law of large numbers can be applied." But how are we to decide whether a source is "really" ergodic? What measurements could we perform on it? Ergodicity has a precise frequency interpretation only for behavior over infinite periods of time. From an operational viewpoint it is therefore meaningless. How, then, can we ever trust the result of the theorem?

If we look at the problem in Laplace's way this difficulty disappears. When we say, "The source is ergodic," we are not describing the source, but rather our state of knowledge about the source. We mean only that nothing in the available evidence leads us to expect that it has a sub-class of states in which it can get stuck. As far as we know, there is always a possible route by which it can get from any state to any other.

Whether or not this is actually true is irrelevant for the use we make of the theorem. Our job, again, is only to do the best we can with the information we have, and it would be quite unjustified to assume an invariant sub-class of states unless we have evidence to support this. It could, for

example, lead to design of a communication system which turns out to be incapable of handling the actual messages. Ergodicity of this subjective kind is a consequence only of our being conservative and avoiding unwarranted assumptions; the resulting probabilities are the ones which maximize the entropy subject to whatever we do know. Exactly the same argument applies to ergodicity in statistical mechanics.

Many of the fundamental theorems of communication theory can be re-interpreted in this way, and we then see that they are valid and useful in far more general conditions than one would suppose from the frequency definition of probability.

Consider an observer O_n who knows in advance the n-gram frequencies which a source is going to generate, but has no other knowledge about it. What communication system represents rational design on the basis of this much knowledge, what is the best way of encoding into binary digits for the noiseless case, and what channel capacity does O_n require? In principle, the answer is always the same; we need to find the probabilities $p(M)$ which O_n assigns to each of the conceivable messages, and use the method of Fano and Shannon.[21] We wish to emphasize that it makes no sense whatever to say that there exists a "correct" distribution $p(M)$ for this problem; $p(M)$ is an entirely subjective quantity. This becomes especially clear if we suppose that only a single message is ever going to be sent over the communication system, but we wish to transmit it as quickly as possible. Thus there is no conceivable procedure by which $p(M)$ could be measured. This would in no way affect the problem of engineering design which we are considering.

In choosing a distribution p(M), it would be possible to assume a particular message structure beyond n symbols. But from the standpoint of O_n this could not be justified, for as far as he knows, an encoding system based on any such structure is as likely to hurt as to help. From O_n's standpoint, rational conservative design consists in carefully avoiding any such assumption. This means, in short, that O_n should choose the distribution p(M) by maximum-entropy inference based on the known n-gram frequencies.[22] For O_1 and O_2 the solution is well known in a different context; the physicist calls them the linear Ising chain with no interactions, and with nearest-neighbor interactions respectively.[23]

Laplace's point of view is helpful also in the problem of detecting a radar signal in noise. Anyone who studies this problem comes to the conclusion that there is no way of evading the notion of a-priori probabilities of different signals. They are an essential part of the problem, because any prior knowledge we have about the signal is extremely relevant to the proper engineering design. The question of how one finds their "true" numerical values then becomes quite embarrassing. They can be given a frequency interpretation only by devices so arbitrary and forced that they could have no relevance to the problem.

We can now see the answer to this. In the first place, no one needs to apologize for, or do any cautious egg-walking around, the use of Bayes' theorem and a-priori probabilities. This is in fact the only consistent way of handling the problem. We have at present no known procedure for translating our prior knowledge about signals into numerical values of probabilities. At least not on paper. But we still have our brains, and until new principles are discovered, we will have to use them. We must take into account everything we know about the signal, and then guess the a-priori probabilities.

8. CONCLUSION

We have tried to show above how a re-interpretation of the probability concept can clarify and extend the power of statistical methods for current applications in science and engineering. Laplace's view of probability theory as the symbolic logic of plausible reasoning enables us to follow the process which our brains must be using, in every case where numerical values of probabilities can be found. It enables us to do this in far greater detail than is possible on the frequency theory, and to take into account additional evidence which cannot even be stated in terms of frequencies.

The analysis of Sec. 2 above is, of course, far from rigorous in the modern sense of the term. However, I believe that all the necessary epsilons and deltas can be supplied by anyone sophisticated enough to feel the need for them. There is always a danger that too much generality will obscure the important points of an argument. Finally, it is interesting to note the increasing importance of the theory of functional equations in this field, shown also by Bellman and Kalaba.[24]

REFERENCES

[1]C. E. Shannon, "A Mathematical Theory of Communication, " Bell Syst. Tech. Jour. Vol. 27, pp. 379-423, 623-656; July, October, 1948. Also in C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication," University of Illinois Press, Urbana, 1949.

[2]N. H. Abel, Crelle's Jour., Bd. 1 (1826).

[3]R. T. Cox, "Probability, Frequency, and Reasonable Expectation," Am. Jour. Physics, Vol. 14, p. 1 (1946). This is a very important, but unfortunately little-known, paper which comes quite close to solving the problem of Sec. 2.

[4]"La théorie des probabilités n'est que le bon sens reduit au calcul." This occurs in the Introduction to P. S. Laplace, "Exposition de la théorie des chances et des probabilités," Paris, 1843. The same statement, with slightly different wording, is found in the Truscott-Emory translation of P. S. Laplace, "A Philosophical Essay on Probabilities," Dover Publications, N. Y. (1951); p. 196.

[5]G. Polya, "Mathematics and Plausible Reasoning," Volumes I and II, Princeton University Press, 1954.

[6]G. Polya, "How to Solve It," Princeton University Press, 1945; Second paperbound edition by Doubleday Anchor Books, Inc., Garden City, N. Y., 1957.

[7]This notation is perhaps confusing. It can be made clearer if we suppose that the symbol for a plausibility is not (A B), but just A B, the parentheses being unnecessary. However, when one writes down more involved equations, the absence of parentheses can cause even greater confusion.[3] The notation adopted here, while not entirely consistent, appears to the

writer as the lesser of two evils.

[8]H. Jeffreys, "Theory of Probability," Oxford University Press, 1939.

[9]This is not a direct quotation from any particular author, but a statement of what is implied by many authors. For example, see Ref. 10, pp 150-151, or Ref. 12, pp 4-6.

[10]H. Cramer, "Mathematical Methods of Statistics," Princeton University Press, 1946.

[11]Reference 5, Vol. II, p. 136. For other examples, see Ref. 8, pp 107-110, and Ref. 12, p. 84.

[12]W. Feller, "An Introduction to Probability Theory and its Applications," John Wiley and Sons, Inc., N. Y., 1950. Any reader familiar with this book will see at once that the present paper is largely a reaction against, and search for an alternative to, the philosophical views expressed therein. I believe this is necessary if probability theory is to meet all the needs of science and engineering. But no one can challenge Feller's beautiful mathematical results, the validity of which does not depend on how we choose to interpret them. They are as useful in Laplace's theory as in the frequency theory.

[13]This is far from being a precise statement. The derivation of Eq. (6-13) shows in more detail what is required for the law of succession to apply.

[14]However, it served Laplace very well indeed. The following procedure led him to some of the most important discoveries in celestial mechanics. Noting a discrepancy between observation and existing theory, he would break down the situation into alternatives which seemed intuitively "equally possible." He would then compare the probability that a discrepancy of this size is

due to a systematic effect, with the probability that it is due to errors of observation. Whenever the ratio was sufficiently high, he would decide that this is a problem worth working on, and attack it. He was, in fact, using Wald's decision theory, in exactly the way developed recently by Middleton, van Meter, and others for the detection of signals in noise!

[15]Ref. 10, pp. 507-524.

[16]E. T. Jaynes, "Information Theory and Statistical Mechanics," Phys. Rev., Vol. 106, pp. 620-630; May 15, 1957. At the time of writing this, I was under the impression that the frequency theory and Laplace's theory are parallel, co-equal theories using the same mathematical rules. However, the arguments of the present paper show that the frequency theory is only a special case of Laplace's theory.

[17]E. T. Jaynes, "Information Theory and Statistical Mechanics II," Submitted to the Phys. Rev.

[18]E. T. Jaynes, "Poincaré Recurrence Times and Statistical Mechanics," Submitted to the Phys. Rev.

[19]This can be stated in a more precise epsilon-delta language, but the reader will anticipate that the conclusions are largely independent of what we mean by "reasonably probable," for the same reason as in Shannon's theorem 4.

[20]$(D_f|A_p)$ is a probability density, $(D_f|A_p)\,df$ being a probability. Since, however, the differentials cancel out of equations and the distinction is already determined by whether the variable is continuous or discrete, there is no need to invent a new notation. On the other hand, it is essential in this theory that we do distinguish in notation between a probability and a frequency.

[21] Reference 1, § 9.

[22] This was recognized by Shannon (Ref. 1, § 10).

[23] G. F. Newell and E. W. Montroll, "On the Theory of the Ising Model of Ferromagnetism," Rev. Mod. Phys. Vol. 25, pp. 353-389; April, 1953.

[24] R. Bellman and R. Kalaba, "On the Role of Dynamic Programming in Statistical Communication Theory," Rand Corp. Report P-949, Dec. 19, 1956.

I R E Professional Group Correspondence

UNIVERSITY OF ILLINOIS
COLLEGE OF ENGINEERING
URBANA, ILLINOIS

DEPARTMENT OF ELECTRICAL ENGINEERING

December 23, 1958

Please address
Reply to:

Prof. G.A. Deschamps
University of Illinois
Electrical Engineering
Research Laboratory
Urbana, Illinois

Dr. E.T. Jaynes
Stanford University
W.W. Hansen Laboratories of Physics
Stanford, California

Dear Dr. Jaynes:

I have accepted the job of editor for the PGIT and one of my first unpleasant tasks is to report on your paper "How Does the Brain do Plausible Reasoning". It has finally been reviewed and the reviewers have recommended its rejection.

I am returning your manuscript and a copy of the reviewers' comments. I have also to apologize, in the name of the editorial committee, for the extremely slow handling of this review and the fact that one reviewer even lost the manuscript that was sent to him!

Personally I enjoyed reading your paper where I found some echos of comments you made to me when I visited Stanford. My impression is that your title has misled the reviewers. You are not really discussing functioning of the brain but rather showing a possible axiomatic derivation of probability theory. It is almost as if Euchid had called his Elements (at least the axiomatic part) "How does the Brain do Geometric Reasoning"!.

I have still a very pleasant memory of my visit with you, three years ago, and hope that we may meet again. I am presently teaching Electromagnetic Theory at the University of Illinois. How is your project of a book on the subject coming along?

I wish you a pleasant holiday season and happy new year.

With my best personal regards,

Cordially yours,

Georges Deschamps

Georges A. Deschamps
Editor of IRE
Transactions on
Information Theory

GAD:wjh

Reviewers' Comments on

"How Does the Brain do Plausible Reasoning"

by Dr. E. T. Jaynes

The author attempts to show that there is only a single set of rules that the brain uses for plausible reasoning. He then relates these rules to statistics more generally and includes communication theory and thermodynamics. He fails to convince me of the relation of these things to the brain, and I will explain this below. The rest of the paper might be termed philosophizing about the justification for probability theory and related subjects. I am inclined to think that this would not be particularly appropriate for the Trans. PGIT.

Many of Prof. Jaynes remarks about the brain do not seem adequately supported by evidence that he cites. Furthermore there are many places where his chain of reasoning is weak. This is illustrated by the following quotations.

On page 2, line 7, he says, "No jury has ever reached a verdict on the basis of pure deductive reasoning". This is a rather sweeping statement which I suspect that Prof. Jaynes would weaken if it were called to his attention.

Right after the statement about juries, he says, "Therefore, the human brain must contain some fairly definite mechanism for plausible reasoning." I gather that he means that there must be similarity between the mechanism in one brain and the mechanism in the next. To show that this does not follow from the evidence he cites beforehand, let me propose a set of affairs that is different.

> Perhaps different people have widely differing mechanisms for plausible reasoning, however, they are all subject to the social pressures of our culture. These pressures enable the widely differing brains to, by and large, learn to produce similar behavior. Those brains that never manage to learn to produce acceptably conforming behavior are eventually restrained by prisons or mental hospitals and would, therefore, not show up on juries. In other words, we might have, at large, brains with many different sets of rules, however, these different sets will all produce roughly a similar reasoning in just those situations required for staying out of constraining institutions.

Prof. Jaynes has not eliminated this possibility and since he says that, "This is the 'experimental fact' on which our theory is based", some patching is needed.

The kind of evidence needed is the kind that experimental psychologists use when they support assertions about the brain. He should cite experimental evidence that (1) can be tested by the reader, and (2) makes any alternative to his assertion quite unlikely.

In the next section he says, "Now it is clear that our attitude toward any given proposition may have a very large number of different 'coordinates'. We form simultaneous judgments as to whether it is probably, whether it is desirable,..." He gives no evidence to indicate that these judgments are made every time the proposition appears as a conscious thought. It may be that these judgments are merely potential sequels and have no existence except when the train of thought continues in some particular direction.

To illustrate this objection more clearly, let us consider it in machine terms. Suppose we have a serial machine, and suppose a 1 came into the units position of the address register. By an argument similar to that of Prof. Jaynes, I believe that I can say that this 1 would have as "coordinates" every number that was stored in memory at an address ending in 1.

On pages 3, line 9, he says, "A moment's introspection will show that..." Introspection is apt to be misleading. It may safely be used as a mehcanism to suggest a solution but it isn't good as a way to prove something to somebody else. Different people's introspection leads to different answers.

In what follows in the middle of page 3 he makes some statements that I interpret to mean that reasoning about natural science and mathematics is the kind of reasoning that is least perturbed by a given amount of imperfection of the human brain. It seems to me that the behavior of mildly intoxicated scientists at a party provides a counter example. The ability to react successfully to many situations in daily life is retained long after the individual has lost the ability to do rapid correct arithmetic.

My judgment is that this paper does not now present convincing evidence about the nature of the brain. If additional evidence could be found and the reasoning in the paper made more secure, it would be an important paper. It is not clear to me that our knowledge of the brain has reached the stage of refinement at which we can say with any degree of certainty how it does any reasoning, plausible or not. At most he should assert that he is conjecturing about the brain, and, if so, he should provide more substantial support for his conjectures.

The presentation is very uneven. The author seems to address alternately people with great sophistication and highly specialized knowledge and, on other occasions, reduces himself to trivialities. In many instances he seems cryptic when he says, for instance, on page 12: "It is the connection of the numbers "Plausibility A, given C" with intuitive states of mind that never gets tied down in any definite way". On other instances he seems to be just sloppy, e.g., on the same page, what are "intuitive sensations"? Or on page 11 "data" is three times used as a singular. On page 5, I do not believe that he really thinks that

if one has - even a unique - mathematical model of some process, one has discovered its mechanism. For the reader who would like to know a little bit more about the background of the different arguments existing in probability theory I would suggest a more complete bibliography.

Unfortunately, two of the most sweeping statements in the paper are backed by articles by the author which are yet to be published. This makes it very hard on the reader to judge their validity.

The paper does not even touch the problem of the organization of the brain which is capable of performing a stunt like "Plausible Reasoning".

Dec. 31, 1958

Dr. Georges A. Deschamps
Dep't of Electrical Engineering
University of Illinois
Urbana, Illinois

Dear Dr. Deschamps:

Thank you for your letter of Dec. 23, 1958. I am delighted to learn that you are not the PGIT Editor.

I had been awaiting action on my manuscript, "How Does the Brain do Plausible Reasoning," with more amusement than impatience, the long delay indicating the consternation it must have caused. Your diagnosis was, of course, 100 percent correct. The referees completely missed the point in thinking that I was writing about the physiology of the human brain, and attacking the paper on those grounds.

I expected that anyone with a sense of humor would see that we were developing the principles of operation of an idealized "robot" brain which does inductive reasoning, the important result being that elementary qualitative requirements for this brain lead uniquely to classical probability theory as envisaged by Laplace. This has an obvious bearing on an argument that has been going on for a century, concerning the proper use of probability theory.

Since no one questioned the soundness of the mathematical arguments, I think I will write a new paper giving only them, and submit it to you in a few weeks. Also, be warned that I am about to send you another short paper concerning uniquely decipherable codes.

In the sixteen months since submitting my "brain" paper, this theory has been developed much further, and I have obtained U. S. Air Force Contract support to continue it. At present, I have three graduate students here studying its application to statistical problems in physics, and two publications based on it are currently "in the mill."

Recently, I spent a week in Dallas, Texas, giving a series of lectures on this at eμ the Magnolia Petroleum Company Research Laboratory. A set of notes, made from tape recordings of the lectures, will be available soon and I will send you a copy. I think you will agree, after reading them, that classical probability theory, with the principle of insufficient reason generalized to the principle of maximum entropy, unifies and simplifies this whole field.

One of the referees expressed doubt as to the applicability of this theory to physics. I might point out that the approach to statistical mechanics which this gives is now being taught in three U. S. Universities other than Stanford. At UCLA, Professor M. Tribus has just written a new textbook on thermodynamics and statistical mechanics based on it. Beginning next school year, engineering students at UCLA are going to be taught this fo theory in their Junior undergraduate year, with thermodynamics, industrial quality control, and communication theory then e derived from it as special cases.

I will remove the features which the referees did not like (i.e. remove the words "human brain"), and shorten the paper by about a factor of two, before re-submitting it.

It will be at least another year before my book on Microwave Circuit Theory will be out. I have delayed it several times because I wanted to include new material, and give a more rigorous treatment of expansions in orthogonal functions. Particularly, it turns out that you often want to differentiate these expansions term by term, but the resulting seriws diverges. Nevertheless, it can be made faia fairly rigorous that you can still use them if you "subtract out" the divergent part. This has quite a bit to do t with the theory of distributions, and it took me a long time to figure out how to present this in a simple way.

Very truly yours,

E. T. Jaynes
Associate Professor

ETJ:mm