Lecture 2

PLAUSIBLE REASONING

Suppose some dark night a policeman walks down the street, and the place is completely deserted apparently; but all of a sudden he hears a burglar alarm, he looks across the street, and sees a jewelry store with a broken window. Also, there's a gentleman wearing a mask, crawling out through the broken window, carrying a bag which turns out to be full of watches and diamond rings. The policeman doesn't hesitate at all in deciding this gentleman is dishonest. But by what reasoning process does he arrive at this conclusion?

2.1 Deductive and Inductive Reasoning

A moment's thought makes it clear that our policeman's conclusion was *not* a logical deduction from the evidence; for there may have been a perfectly innocent explanation for everything. It might be, for example, that this gentleman was the owner of the jewelry store and he was coming home from a masquerade party, and didn't have the key with him. He noticed that a passing truck had thrown a stone through the window, and he was merely protecting his own property. You see, the conclusion which seems so easily made was certainly not an example of logical deduction.

Now while we agree that the policeman's reasoning process was not an example of logical deduction, we still will grant that it had a certain degree of validity. The evidence didn't make the gentleman's dishonesty *certain*, but it did make it extremely *plausible*. This is an example of the kind of reasoning which we all have to use a hundred times a day. We're always faced with situations where we don't have enough information to permit deductive reasoning, but still we have to decide what to do.

The formation of plausible conclusions is a very subtle process and it's been discussed for centuries, and I don't think anyone has ever produced an analysis of it which anyone else finds completely satisfactory. These problems haven't been solved and they're certainly not going to be solved in these talks; but I do hope that we'll be able to say a few new things about them.

All discussions of these questions start out by giving examples of the contrast between deductive reasoning and plausible reasoning. The syllogism is the standard example of deductive reasoning:

If A is true, then B is true

A is true

Therefore, B is true

or, its inverse:

If A is true, then B is true

B is false

Therefore, A is false

This is the kind of reasoning we'd like to use all the time; but, unfortunately, in almost all the situations we're confronted with we don't have the right kind of information to allow this kind of reasoning. We fall back on weaker forms:

If A is true, then B is true

B is true

Therefore, A becomes more plausible

The evidence doesn't prove that A is true, but verification of one of its consequences does give us more confidence in A. Another weak syllogism, still using the same major premise, is:

If A is true, then B is true

A is false

Therefore, B becomes less plausible

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

Now the reasoning of the policeman in this example was not even of the above types. It is best described by a still weaker form:

If A is true, then B becomes more plausible

B is true

Therefore, A becomes more plausible

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

This shows that the brain, in doing plausible reasoning, not only decides whether something becomes more plausible or less plausible, but it evaluates the degree of plausibility in some way. And it does it in some way that makes use of our past experience as well as the specific data of the problem we're reasoning on. To illustrate, for example, that the policeman was making use of the past experience of policemen in general, we have only to change that experience. Suppose that these events happened several times every night to every policeman, and in every case the gentleman turned out to be completely innocent. Well, very soon policemen would be ordered to ignore such trivial things. This shows that in our reasoning we depend very much on past experience--or as we will presently call it, on _prior information_--to help us in evaluating the degree of plausibility. This reasoning process goes on unconsciously, almost instantaneously, and we conceal how complicated it really is by calling it _common sense_.

Professor George Polya has written three books on plausible reasoning (Polya, 1945, 1954), pointing out all sorts of interesting examples, showing that there are fairly definite rules by which we do plausible reasoning (although in his work they remain in qualitative form). Evidently, the deductive reasoning described above has the property that you can go through arbitrarily long chains of reasoning of this type and the conclusions have just as much certainty as the premises. With the other kinds of reasoning, the reliability of the conclusion attenuates if you go through several stages. Polya showed that even a pure mathematician actually uses these weaker kinds of reasoning most of the time. Of course, when he publishes a new theorem, he'll be very careful to invent an argument which uses only the first kind of reasoning; and his professional reputation depends on his ability to do this. But the process which led him to the theorem in the first place almost always involves one of the weaker forms.

Now the problem I'm concerned with is this. Is it possible to reduce this process of plausible reasoning to quantitative terms? The idea of inventing a mathematical theory of reasoning, both deductive and inductive, is a very old one. Leibnitz had speculated on such a "Characteristica Universalis," almost 200 years before Boole's _The Laws of Thought_ (1854) provided a calculus of deductive reasoning. When the theory of probability was developed, culminating in Laplace's Theorie Analytique (1812), it was believed to be the long-awaited "calculus of inductive reasoning," fully developed. Throughout the 19th century this was the prevailing view, expounded by such people as Laplace, de Morgan, Maxwell, Poincaré, and many others. And yet, in the 20th century we find that probability theory has erupted into controversy, almost all of this fruitless, inconclusive kind, in which one person attacks the assumptions of another person.

This issue has been framed rather sharply by Ludwig von Mises (von Mises, 1957; 1963) who is really violent in denouncing any idea that probability theory has anything to do with inductive reasoning. He insists that it is, instead, "the exact science of mass phenomena and repetitive events." On the other hand, Sir Harold Jeffreys (Jeffreys, 1939; 1955) is equally vigorous in upholding the opposite view, and insists that probability theory is exactly what Laplace thought it was: the "calculus of inductive reasoning."

Well, which is it? I want to point out that it makes a big difference in applications. Science and engineering offer many problems where use of probability theory is entirely legitimate on one interpretation, and entirely unjustified on the other. Even in cases where both viewpoints would allow the use of probability theory, your decision as to which mathematical problems are important and worth working on, can still depend on which viewpoint you adopt. (As an example, whose meaning will become clear later: when approximations are necessary, is it the sampling distribution of a statistic or the posterior distribution of a parameter that should be approximated? The two different schools of thought will give opposite answers to this; and each regards the mathematical labors of the other as effort wasted on a false problem.)

Sooner or later, such an unsettled condition in probability theory couldn't fail to have pretty serious repercussions in theoretical physics and engineering--both of which make more and more use of probability methods. And so now you see why any serious student of physics or engineering must become worried about this situation. I hope to show in these talks that some of the outstanding unsolved problems in both physics and communication theory have their origin in this state of utter confusion which exists in the foundations of probability theory.

2.2 Analogies with Physical Theories

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

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

The ultimate test of a physical theory is not, "Can you demonstrate it by logic?" but only; "Is it free of obvious inconsistencies and does it agree with experiment?" It has taken the human race thousands of years to comprehend this simple fact. It was utterly unknown to the ancient philosophers, and Galileo was the first to demonstrate clearly the advantages of recognizing it.

It is exactly the same in our present problem. The test of any model of plausible reasoning is not "Can you prove that it is correct?" Real life, unfortunately, does not permit such a Utopian program. The only test which can actually be applied in practice is: "Is it free of inconsistencies and in agreement with common sense?" It has taken us a long time to realize this, and I'm sure that there are still many people who will dispute it vigorously.

The analogy with physical theories goes a lot deeper than a mere analogy of method. Often, the things which are most familiar to us turn out to be the hardest to understand. Our universities can train people to perform surgery on the living heart and measure the internal charge distribution of the proton; but nobody seems to know how to prevent the common cold, and all of modern science is practically helpless when faced with the complications of such a commonplace thing as a blade of grass. Accordingly, we must not expect too much of our models; we must be prepared to find that some of the most familiar features of mental activity may be ones for which we have the greatest difficulty in constructing any adequate model.

There are many more analogies. In physics we are accustomed to find that any advance in knowledge ultimately leads to consequences of the greatest practical value, but of a totally unpredictable nature. Roentgen's discovery of x-rays led to important new possibilities of medical diagnosis; Maxwell's discovery of one more term in the equation for curl H led to the possibility of practically instantaneous communication all over the earth.

Our mathematical models for common sense also exhibit, although on a more modest scale, this feature of practical usefulness. Any successful model, even though it may reproduce only a very few features of common sense, will prove to be a powerful extension of common sense in some field of application. Within this field, it enables us to solve problems of plausible reasoning which are so involved that we would never attempt to solve them without its help. Thus the problem of optimum design of an electrical filter or an antenna (which is just a particular kind of filter, operating in space instead of in time) can sometimes be solved by applying a model of common sense. Similarly, we will show that the prediction of the laws of thermodynamics, including all experimentally reproducible features of irreversible processes, can be viewed as an application of a single, formally very simple model of common sense.

Models may have practical uses of a quite different type. Many people are fond of saying, "They will never make a machine to replace the human mind--it does many things which no machine could ever do." One of the best answers to this attitude was given by J. von Neumann in a talk on computers given at the Institute for Advanced Study in Princeton in 1948, which I was privileged to attend. In reply to the canonical question from the audience ("But of course, a mere machine can't really <u>think</u>, can it?"), he said: "Look here. You insist that there is something a machine cannot do. If you will tell me precisely what it is that a machine cannot do, then I can always make a machine which will do just that!"

The only operations which a machine cannot perform for us are those which we cannot describe in detail. The only limitations on making "machines which think" are our own limitations in not knowing exactly what "thinking" consists of. For further comments on this, see my recent Letter (Jaynes, 1963a). But in our study of common sense we will be led to some very explicit ideas about the detailed mechanism of thinking. Every time we can construct a mathematical model which reproduces a part of common sense by prescribing a definite set of operations, it becomes a kind of blueprint showing us how to build a machine which operates on incomplete data and does plausible reasoning instead of deductive reasoning. In science fiction, such machines have been an accomplished fact for many years. In fact, I want to turn this idea around and instead of asking, "How can we build a mathematical model of common sense?" I want to ask, "How could we build a machine which would do plausible reasoning?"

2.3 Introducing the Robot

Now the question of the process of plausible reasoning that actual human brains use is very charged with emotion and misunderstanding, to the extent that the only solution is to avoid it. Also, it is so complicated that we can make no pretense of explaining all its mysteries; and in any event we are not trying to explain all the abberations and inconsistencies of human brains. That is an interesting and important subject, but it is not the subject we are studying here. We are trying rather to understand some of the good features of human brains.

In order to direct attention to constructive things and away from controversial things which we can't answer at present, we will follow the methodology of Gibbs and invent an imaginary beast. His brain is to be designed _by us_, so that he reasons according to certain definite rules. The rules are suggested by properties of human brains which we think, or hope, exist; but by introducing the beast we accomplish the following. You can't object to the theory on the grounds that we have failed to prove the "correctness" of the rules, whatever that may mean. We are free to adopt any rules we please. That's our way of _defining_ which beast we are going to study. After we've worked out the properties of this beast, we can then compare the results of his reasoning process with the results of ours. If you find no resemblance between the way the beast reasons and the way you reason, then you're free to decide that the beast is nothing but an idle, useless toy. But if you find a very strong resemblance, which makes it almost impossible to avoid concluding "I am this beast," then that will be an accomplishment of the theory, not a premise.

Now, let's take a problem with maybe some science fiction overtones. We've been assigned the job of designing the brain case of a robot. This is supposed to be a very sophisticated robot. He doesn't just receive orders and carry them out. He also has to have the ability to learn, he has to be able to make judgments on his own, he has to decide on the best course of action even when we fail to give him full instructions. This means that his brain has got to contain some kind of computing machine which will carry out plausible reasoning whenever the information we give him is insufficient to permit deductive reasoning. How shall we design his brain case? This is a fairly definite engineering problem.

Well, our robot is going to reason about propositions. We denote various propositions by letters A, B, C, and so on, and for the time being we'll have to require that any proposition we use will have, at least to the robot, an unambiguous meaning. It must also be of such a "logical type" that it makes sense to say that the proposition must be either true or false. Of course, not all propositions are of that type at all. Later on we'll see whether there are any possibilities of relaxing that restriction.

Now to each proposition the robot is going to associate some plausibility, which represents his degree of belief in the truth of the proposition, based on all the evidence we have given him up to this time. In order that these plausibilities can be handled in the circuits of his brain, they must be associated with some physical quantity such as voltage or pulse duration or frequency, and so on, however you want to design him. This means that there will have to be some kind of association between degrees of plausibility and real numbers. This assumption, you see, is practically forced on us by the requirement that the robot's brain must operate by the carrying out of some definite physical process.

Let me emphasize the contrast between such a robot and a human brain. We have decided that we will attempt to associate mental states with 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." You and I form simultaneous judgments not only as to whether it is plausible, but also whether it is desirable, whether it is important, whether it is interesting, whether it is amusing, whether it is morally right, etc. If we assume that each of these judgments might be represented by a number, then a fully adequate description of a state of mind would be represented by a vector in a space of a rather large number of dimensions.

Not all propositions require this. For example, the proposition, "The refractive index of water is less than 1.3" generates no emotions; consequently the state of mind which it produces has very few coordinates. On the other hand, the proposition, "Your mother-in-law just wrecked your new car" generates a state of mind with an extremely large number of coordinates. A moment's introspection will show that, quite generally, the situations of everyday life are those involving many coordinates. It is just for this reason, I suggest, that the most familiar examples of mental activity are often the most difficult to reproduce by a model.

We might speculate further. Perhaps we have here the reason why science and mathematics are the most successful of human activities; they deal with propositions which produce the simplest of all mental states. Such states would be the ones least perturbed by a given amount of imperfection in the human mind.

I interject these remarks to point out that there is a large unexplored area of possible generalizations and extensions of the theory to be developed here; perhaps this may inspire others to try their hand at developing "multi-dimensional" theories of mental activity, which would more and more resemble the behavior of actual human brains. Such a theory, if successful, might have an importance beyond our present ability to imagine.

For the present, however, we will have to be content with a much more modest undertaking. Is it possible to develop a consistent "one-dimensional" model of reasoning? Evidently, our problem will be simplest if we can manage to represent a degree of plausibility uniquely by a single real number, and ignore the other "coordinates" just mentioned; and at the risk of belaboring it, let me stress again: we are in no way asserting that degrees of plausibility in actual human minds have a unique numerical measure. Our job is not to postulate any such thing; it is to investigate whether it is possible, in our robot, to set up such a correspondence without contradictions. If the attempt to do this should fail, then we will have to consider more complicated kinds of association; but I propose to try out the simplest possibility first.

We'll adopt a convenient but nonessential convention; that this will be done in such a way that a greater plausibility always corresponds to a greater number. It will be convenient to assume also a continuity property, which is hard to state precisely at this stage; but to say it intuitively: an infinitesimally greater plausibility ought to correspond only to an infinitesimally greater number.

To state the above ideas more formally, we introduce some notation of the usual symbolic logic, or Boolean algebra. By the symbolic product

$$AB$$

we mean the proposition "both A and B are true." Obviously, AB and BA are the same proposition. The expression

$$A+B$$

stands for the proposition: "at least one of the propositions A, B is true," and is the same as B+A. The plausibility that the robot associates with proposition A could, in general, depend on whether we told him that some other proposition B is true. And so we indicate this by the symbol

$$(A|B).$$

I'll call this the "conditional plausibility of A, given B;" or just, "A given B." It stands for some real number. Thus, for example,

$$(A|BC)$$

(I'll read this as "A given BC") represents the plausibility that A is true, given that B and C are true. Or,

$$(A+B|CD)$$

represents the plausibility that at least one of the propositions A and B is true, given that both C and D are true, and so on. Now we've decided that we're going to associate greater plausibility with greater numbers, so

$$(A|B)>(C|B)$$

says that given B, A is more plausible than C.

You know that when a computing machine is asked to divide by zero, it develops a psychosis--the poor machine tries its best, but just can't solve the problem. On some old kinds of desk calculators the only thing you can do is to put the machine out of its misery by pulling the plug. In the interest of avoiding impossible problems, we are not going to ask our robot to undergo the agony of reasoning on the basis of mutually contradictory propositions. Thus, we make no attempt to define $(A|BC)$ when B and C are mutually contradictory. Whenever such a symbol appears, we will understand that B and C are compatible propositions.

Now we wouldn't want this robot to behave in a way that's very greatly different from human behavior, because that would make him very hard to live with and nobody would want to keep such a robot in his home. So, we'll want him to reason in a way that is at least qualitatively like the way you and I reason, as described by the above weak syllogisms. As a further example, if he gets new information which increases the plausibility $(A|BC)$ but does not affect the plausibility $(B|C)$, this of course can only produce an increase, never a decrease, in the plausibility $(AB|C)$ that <u>both</u> A and B are true. And it can only produce a decrease, not an increase, in the plausibility that A is false. This qualitative requirement simply gives us the sense of direction in which reasoning goes; it says nothing about <u>how</u> <u>much</u> the plausibilities change.

Also, it would be nice if we could give this robot a very desirable property which we don't have; namely, that he always reasons consistently. By "consistently" I mean three things:

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

(b) If two problems are entirely equivalent; i.e., if the robot's state of knowledge is the same in both, then he must assign the same plausibilities in both.

(c) The robot is completely non-ideological; if he has several pieces of evidence relevant to a question, he does not arbitrarily throw out part of his evidence, basing his conclusions only on what remains; he always takes into account <u>all</u> of the evidence available to him.

All right. Now I claim something which may seem startling. The conditions that we have imposed are:

1. Representation of degrees of plausibility by real numbers.

2. Qualitative correspondence with common sense.

3. Consistency.

These requirements, I claim, uniquely determine the rules according to which this robot must reason; there is only one set of mathematical operations which has all these properties. In the next Lecture we commence the mathematical development by deducing these rules.