The Concept of Probability in Quantum Mechanics

Richard P. Feynman

1951

From about the beginning of the twentieth century experimental physics amassed an impressive array of strange phenomena which demonstrated the inadequacy of classical physics. The attempts to discover a theoretical structure for the new phenomena led at first to a confusion in which it appeared that light,and electrons, sometimes behaved like waves and sometimes like particles. This apparent inconsistency was completely resolved in 1926 and 1927 in the theory called quantum mechanics. The new theory asserts that there are experiments for which the exact outcome is fundamentally unpredictable, and that in these cases one has to be satisfied with computing probabilities of various outcomes. But far more fundamental was the discovery that in nature the laws of combining probabilities were not those of the classical probability theory of Laplace.

I want to discuss here the laws of probability of quantum mechanics. The subject is over twenty years old and has been expertly discussed in many places. My only excuse for speaking about it again is the hope that, being mathematicians, all of you may not have heard of it in detail. And you may be delighted to learn that Nature with her infinite imagination has found another set of principles for determining probabilities; a set other than that of Laplace, which nevertheless does not lead to logical inconsistencies. We shall see that the quantum mechanical laws of the physical world approach very closely the laws of Laplace as the size of the objects involved in the experiments increases. Therefore, the laws of probabilities which are conventionally applied are quite satisfactory in analyzing the behavior of the roulette wheel but not the behavior of a single electron or a photon of light.

I should say, that in spite of the implication of the title of this talk the concept of probability is not altered in quantum mechanics. When I say the probability of a certain outcome of an experiment is p, I mean the conventional thing, that is, if the experiment is repeated many times one expects that the fraction of those which give the outcome in question is roughly p. I will not be at all concerned with analyzing or defining this concept in more detail, for no departure from the concept used in classical statistics is required.

What is changed, and changed radically, is the method of calculating probabilities. The effect of this change is greatest when dealing with objects of atomic dimensions. For this reason we shall illustrate the laws of quantum mechanics by describing the results to be expected in some experiments deal-

ing with a single electron. The experiment is illustrated in figure 1.

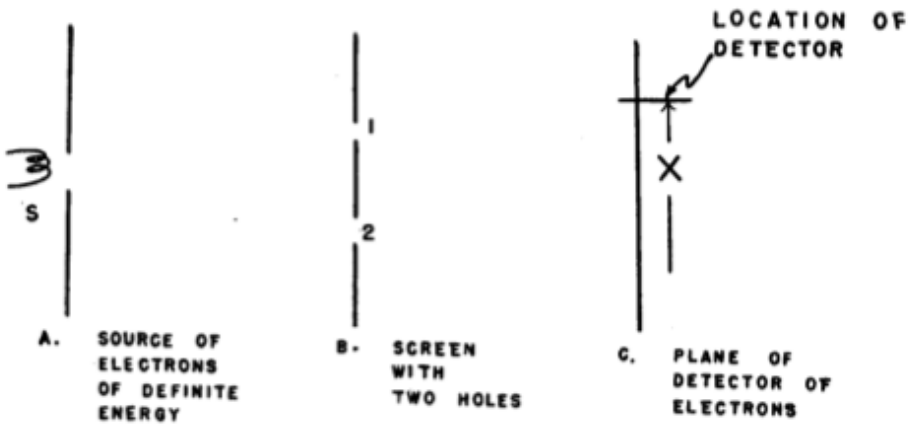

Figure 1: An experiment to determine the probability that electrons arrive at a detector at X

At A we have a source of electrons S. The electrons at S all have the same energy but come out in all directions to impinge on a screen B. The screen B has two slits, 1 and 2 through which the electrons may pass. Finally behind the screen B at a plane C, we have a detector of electrons which may be placed at various distances X from the center of the screen.

If the detector is extremely sensitive (such as a Geiger counter) it will be discovered that the current arriving at X is not continuous, but corresponds to a rain of particles. If the intensity of the source S is very low the detector will record pulses representing the arrival of a particle, separated by gaps in time during which nothing arrives. This is the reason we say electrons are particles. If we had detectors simultaneously all over the screen C, with a very weak source S, only one detector would respond, then after a little time, another would record the arrival of an electron, etc. There would never be a half response of the detector, either an entire electron arrives or nothing happens. And two detectors would never respond simultaneously (except for the coincidence that the source emits two electrons within the resolving time of the detectors - a coincidence whose probability can be decreased by further decrease of the source intensity). In other words the detector records the passage of a single corpuscular entity traveling from S through the holes in screen B to the point X.

(Incidentally, if one prefers one can just as well use light instead of electrons in this experiment. The same points would be illustrated. The source S could be a source of monochromatic light and the sensitive detector a photoelectric cell or better a photomultiplier which would record pulses, each being the arrival of a single photon.)

What we shall measure for various positions X of the detector is the mean number of pulses per second. In other words we shall determine experimentally the (relative) probability P that the electron passes from S to X, as a function of X.

The graph of this probability as a function X is the complicated curve illustrated qualitatively in ?gure 2(a). It has several maxima and minima, and there are locations near the center of the screen at which electrons hardly ever arrive. It is the problem of physics to discover the laws governing the structure of this curve.

We might at first suppose (since the electrons behave as particles) that

I. Each electron which passes from S to X must go either through hole 1 or hole 2. As a consequence of I we expect that:

II. The chance of arrival at X should be the sum of two parts, P_1, the chance of arrival coming through hole 1, plus P_2, the chance of arrival coming through hole 2.

We may find out if this is true by direct experiment. Each of the component probabilities is easy to determine. We simply close hole 2 and measure the chance at arrival at X with only hole 1 open. This gives the chance P_1 of arrival at X for those coming through 1. The result given in figure 2(b). Similarly, by closing 1 we find the chance P_2 of arrival through hole 2, [figure 2(c)].

The sum of these [figure 2(d)] clearly does not agree with the curve (a). Hence experiment tells us definitely that, $P \neq P_1 + P_2$ or that II is false.

The chance of arrival at X with both holes open is not the sum of the chance with just hole 1 open plus that with just hole 2 open.

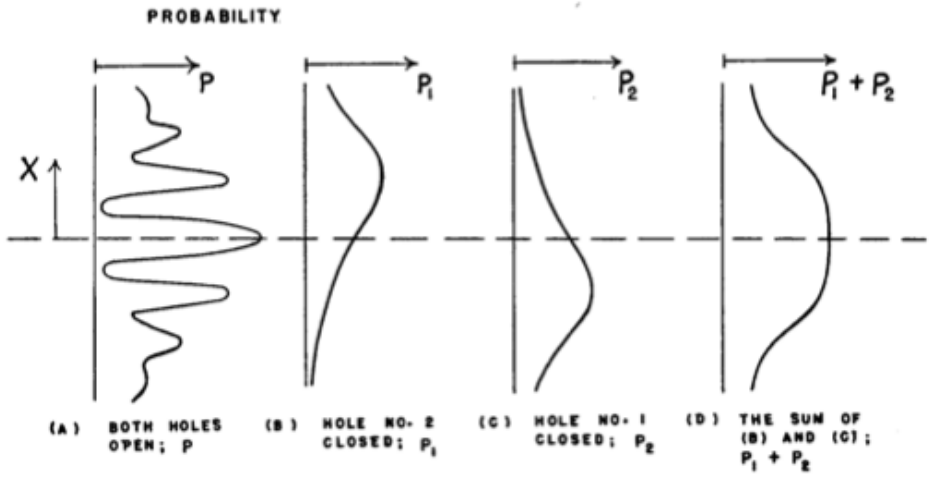

Figure 2: Results of the experiment. Probability of arrival of electrons at X plotted against the position X of the detector.

Actually, the complicated curve $P(X)$, is familiar inasmuch as it is exactly the intensity of distribution in the interference pattern to be expected if waves starting from S pass through the two holes and impinge on the screen C. We can state the correct law mathematically by saying that $P(X)$ is the absolute square of a certain complex quantity (if electron spin is taken into account it is a hypercomplex quantity) $\phi(X)$ which we call the probability amplitude of arrival at X and furthermore $\phi(X)$ is the sum of two contributions ϕ_1, the amplitude of arrival through hole 1 plus ϕ_2 the amplitude of arrival through hole 2. In other words,

III. There are complex numbers in, and 4:2 such that

$$P = |\phi|^2 \quad , \quad \phi = \phi_1 + \phi_2$$

and

$$P_1 = |\phi_1|^2 \quad , \quad P_2 = |\phi_2|^2$$

We discuss in a little more detail later the actual calculation of ϕ_1 and ϕ_2. Here we say only that ϕ_1, for example, may be calculated as a solution of a wave equation representing waves spreading from the source to 1 and from 1 to X. This reflects the wave properties of electrons (or in the case of light,

photons).

To summarize: we *compute* the intensity (that is, the absolute square of the amplitude) of waves which would arrive in the apparatus at X, and then *interpret* this intensity as the probability that a particle will arrive at X.

What is remarkable, is that this dual use of wave and particle ideas does not lead to contradictions. This is only so if great care is taken as to what kind of statements one is permitted to make about the experimental situation.

To discuss this point in more detail we first consider the situation which arises from the observation that our new law III of composition of probabilities implies in general, that it is not true that $P = P_1 + P_2$. We must conclude that when both holes are open it is not true that the particle goes through one hole or the other. For if it had to go through one or the other we could classify all the arrivals at X into two disjoint classes, namely, those arriving via hole 1 and those arriving through hole 2, and the frequency P of arrival at X would be surely the sum of the frequency P_1 of those coming through 1 and of those coming through hole 2, P_2.

To extricate oneself from the logical difficulties introduced by this startling conclusion one might try various artifices.

We might say perhaps, for example, that the electron travels in a complex trajectory going through hole 1, then back through hole 2 and finally out through 1 in some complicated manner. Or perhaps, the electron spreads out somehow and passes partly through both holes so as to eventually produce the interference result III. Or perhaps the chance P_1 that the electron passes through hole 1 has not been determined correctly inasmuch as closing hole 2 might have influenced the motion near hole 1. Many such classical mechanisms have been tried to explain the result III, but none of them has in the end proved successful. In particular, in the case when light photons are used, (in which case the same law III applies) the two interfering paths 1 and 2 can be made to be many centimeters apart (in space) so that the two alternative trajectories must almost certainly be independent. That the actual situation is more profound than might at first be supposed is shown by the following experiment.

We have concluded on logical grounds that since $P \neq P_1 + P_2$, it is not true

that the electron passes through hole 1 or hole 2. But it is easy to design an experiment to test our conclusion directly. One has merely to have a source of light behind the holes and watch to see through which hole the electron passes. For electrons scatter light, so that if light is scattered behind hole 1 we may conclude that an electron passed through hole 1 and if it is scattered in the neighborhood of hole 2 the electron has passed through hole 2.

The result of this experiment is to show unequivocally that the electron *does* pass through either hole 1 or hole 2! That is, for every electron which arrives at the screen C (assuming the light was strong enough that we do not miss seeing it) light is scattered either behind hole 1 or behind hole 2, and never (if the source S is very weak) at both places. (A more delicate experiment could even show that the charge passing through the holes passes either through one or the other, and is in all cases the complete charge of one electron and not a fraction of it.)

It now appears that we have come to a paradox. For suppose that we combine both experiments. We watch to see through which hole the electron passes and at the same time measure the chance that the electron arrives at X. Then for each electron which arrives at X we can say experimentally whether it came through hole 1 or hole 2. First we may verify that P_1 is given by curve (b). Because if we select of the electrons which arrive at X only those which appear to come through hole 1 (by scattering light there) we find they are indeed distributed as in curve (b). (This result is obtained whether hole 2 is open or closed, so we have verified that there is no subtle influence of closing 2 on the motion near hole 1.) If we select the ones scattering light at 2 we get P_2 of figure (c). But now each electron appears at either 1 or 2 so if we take both together we must get the distribution $P = P_1 + P_2$ illustrated in figure (d). And experimentally we do! Somehow now the distribution does not show the interference effects III of curve (a)!

What has been changed? When we watch the electrons to see through which hole they pass we obtain the result $P = P_1 + P_2$ When we do not watch we get a different result $P = |\phi_1 + \phi_2|^2 \neq P_1 + P_2$.

Just by watching the electrons we have changed the chance that they arrive at X. How is this possible? The answer is that to watch them we used light and the light in collision with the electron may be expected to alter its

motion or more exactly to alter its chance of arrival at X.

On the other hand, can we not use weaker light and thus expect a weaker effect? A negligible disturbance certainly cannot be presumed to produce the finite change in distribution from (a) to (d). But weak light does not mean a weaker disturbance. Light comes in photons of energy $h\nu$ where ν is the frequency, or of momentum h/λ where λ is the wave length. Weakening the light just means using fewer photons so that we may miss seeing an electron. But when we do see one it means a complete photon was scattered and a finite momentum of order h/λ is given to the electron. [Those that we miss seeing are distributed according to the interference law (a), while those we do see and which therefore have scattered a photon arrive at X with the probability $P = P_1 + P_2$ in (d). The net distribution in this case is therefore the weighted mean of (a) and (d). In strong light when nearly all electrons scatter light it is nearly (d), and in very weak light, when very few scatter it becomes more like (a)].

It might still be suggested that since the momentum carried by the light is h/λ, weaker effects could be produced by using light of longer wave length λ. But there is a limit to this. If light of too long a wave length is used, we will not be able to tell whether it was scattered from behind hole 1 or hole 2. For the source of light of wave length λ cannot be located in space with precision greater than that of order λ.

We thus see that any physical agency designed to determine through which hole the electron passes, must produce, lest we have a paradox, enough disturbance to alter the distribution from (a) to (d).

It was first noticed by Heisenberg, and stated in his uncertainty principle, that the consistency of the then new mechanics required a limitation to the subtlety to which experiments could be performed. [1]In our case it says that

[1]The uncertainty principle was first stated for the special case of position and momentum measurements. It said that measurement of a momentum to accuracy Δp implies disturbances sufficient to create an uncertainty in position Δq at least of the order of $h/\Delta p$. That we would be led to a paradox if this were not true can be seen from our experiment in the following way. Instead of determining through which hole the electron passes by using light we may notice that the deflection suffered by the electron in passing from the source to X through hole 1 differs from that suffered in passing through hole 2. Hence the momentum (in the vertical direction in figure 1) given to the electron by the screen

an attempt to design apparatus to determine through what hole the electron passed and delicate enough so as not to deflect the electron sufficiently to destroy the interference pattern, must fail. It is clear that the consistency of quantum mechanics requires that it must be a general statement involving all the agencies of the physical world which might be used to determine through which hole an electron passes. The world cannot be half quantum mechanical half classical. No exception to the uncertainty principle has been discovered.

We are still left with the question, "Do the electrons have to go through hole 1 or hole 2 or don't they?" To avoid the logical inconsistencies into which it is so easy to stumble, the physicist takes the following view. When no attempt is made to determine through which hole the electron passes one cannot say it must pass through one hole or the other. Only in a situation where an apparatus is operating to determine which hole the electron goes through is it permissible to say that it passes through one or the other. When you watch you find that it goes either through one or the other hole, but if you are not looking you cannot say that it either goes one way or the other! Such is the logical tightrope on which Nature demands that we walk if we wish to describe her.

To summarize then: The probability of an event (in an ideal experiment where there are no uncertain external disturbances) is the absolute square

---

is different in the two cases. Call the difference δp. Hence the hole through which the electron passes can be determined by measuring in each case the recoil momentum given to the screen. This can be done by setting screen B free of its supports and measuring its vertical velocity before and after the passage of each electron to determine the change in momentum. The probability distribution must now be (d) instead of the interference pattern(a).This comes about because by freeing the screen from its supports we can no longer be sure of its exact vertical location. In fact, for the passage of each electron the vertical position may differ, by amounts we shall call Δq. Hence the distribution of electrons is that of (a), but smeared out in X by an amount Δq. A simple calculation shows that the separation between maxima and minima in the pattern (a) is just $h/2\delta p$. We must measure the screen momentum with an error Δp which is less than the difference δp if we are to determine the hole through which the electron passes. The uncertainty principle assures us that the vertical uncertainty Δq in the screen position must exceed $h/\Delta p$ and hence exceed $h/2\delta p$ so that the maxima and minima of the diffraction pattern (a) are completely smeared out and the resulting distribution is that of (d).

Many interesting examples of this kind have been analyzed, particularly by N. Bohr.

of a complex quantity called the probability amplitude. When the event can occur in several alternative ways the probability amplitude is the sum of the probability amplitude for each alternative considered separately.

If an experiment capable of determining which alternative is actually taken is performed the interference is lost and the probability becomes the sum of the probability for each alternative.

The main point of this paper has been to discuss this relation of probability amplitude to the calculation of probabilities. Of course, the complete physical theory must also supply the exact formulae for calculating the probability amplitudes for a given situation. The amplitude is usually calculated by solving a kind of wave equation. For particles of low velocity it is called the Schrödinger equation.

A more accurate equation valid for electrons of velocity arbitrarily close to the velocity of light is the Dirac Equation. In this case the probability amplitude is a kind of hypercomplex number. It is a problem of the future to discover the exact manner of computing the amplitudes for processes involving the apparently more complicated particles, namely, neutrons, protons, mesons, etc.

The situation for slowly moving particles, which are usually handled by solving the Schrödinger equation, may also be stated in another way. If a particle is released at a certain point X_1 at a time t_1 we may wish to calculate its amplitude of arrival at some other point X_2 at a later time t_2. We can consider that the particle can take any path $X(t)$ going between the given end points $[X(t_1) = X_1, X(t_2) = X_2]$. Then, the total amplitude for arrival can be considered as the sum over all the possible trajectories of an amplitude Φ for each trajectory. It only remains to give the probability amplitude for a given trajectory to state completely the laws of quantum mechanics in the nonrelativistic (low velocity) limit. The amplitudes for the trajectories are complex numbers (all of the same absolute square magnitude) which simply differ from one another in phase. The phase for a given trajectory $X(t)$ is simply the action $S = \int L dt$ (the time integral of the Lagrangian) calculated classically for this trajectory and measured in units of Planck's constant of action h [that is, $\Phi = \text{constant} \exp(2\pi i S/h)$]. It can be demonstrated[2] that

[2]The formulation is discussed in detail in R. P. FEYNMAN "Spacetime approach to

this formulation leads to the Schrödinger equation. Its relation to classical mechanics is interesting. Quantum mechanically we say all trajectories contribute to an effect, each with amplitude $\exp(2\pi iS/h)$; while classically we say only one trajectory is important, namely that which makes the quantity S an extremum. The classical theory arises from the quantum theory in the limit that the action S is large compared to Planck's constant h. For (by the method of stationary phase) the contributions of most trajectories will cancel out by interference because a neighboring trajectory may contribute with a very different phase. Only those trajectories near the one that makes S a maximum or minimum will be important for they all contribute with nearly the same phase.

When the energy is definite and the particles travel in empty space, as in our experiment, the result can be stated in a still simpler manner. For example ϕ_1 is (except for slowly varying factors involving the width of the slits and cosines of the angles of deflection) proportional to $\exp i(d_1/\lambda)$ where d_1 is the total distance from S to hole 1 plus that from hole 1 to X. The quantity λ, the wave length of the waves, is related to the momentum p of the electron by deBroglie's formula $p = h/\lambda$. Likewise ϕ_2 has the phase d_2/λ. It is for those points X for which d_1 and d_2 differ by an odd number of half-wave lengths that we have destructive interference and a minimum in the probability distribution.

For objects of ordinary size the momentum p is so large that the wave length λ is so short that the maxima and minima of the interference pattern occur so close together as to escape ordinary observation. The relative phases are so large and uncertain that the interference terms are not noticed and ordinary probability laws such as $P = P_1 + P_2$ apply with sufficient accuracy.

The amplitude ϕ_1 can be worked out as the product of two factors $\phi_1 = \phi_{S1} \cdot \phi_{1X}$ where ϕ_{S1} is the amplitude to go from S to the hole at 1 and ϕ_{1X} is the amplitude to go from the hole at 1 to X. (If the hole is not small we shall have to consider each differential of area of the hole, calculate the amplitude of going from S to this area times the amplitude of going from this area to X and sum these amplitudes over the total area of the hole. Each differential area constitutes an alternative.) The composition of probability

nonrelativistic quantum mechanics," Review: of Modern Physics, Vol. 20 (1948), pp. 367-387.

amplitudes bears some formal analogy to Laplace's rules for probability. For events occurring in succession we multiply amplitudes, while when various alternatives are available the amplitude is the sum of those corresponding to each alternative.

Finally, it is interesting to see how formula $P = |\phi_1 + \phi_2|^2$ becomes altered under the influence of light shining on the holes, so that it assumes the classical form $P = P_1 + P_2$. The light by interacting with the electron going through hole 1 alters the phase with which the electron arrives at X by an amount say θ_1, so that the probability amplitude of arrival through hole 1 is now $\phi_1 e^{i\theta_1}$. The value of θ_1, cannot be exactly determined in a given scattering since the precise phase of the light is lost when the scattered light is absorbed in whatever instrument (eye, photocell, etc.) is used to determine whether the light comes from 1 or 2. Exactly how this comes about has been analyzed in many precise situations by von Neumann. Thus the probability of a particular electron arriving at X is $|\phi_1 e^{i\theta_1} + \phi_2 e^{i\theta_2}|^2$. But each scattering corresponds to a different, unknown and random value of the phase shifts θ_1 and θ_2. We must then average $|\phi_1 e^{i\theta_1} + \phi_2 e^{i\theta_2}|^2$ over all phases θ_1, θ_2 obtaining, as is well known,$|\phi_1|^2 + |\phi_2|^2$ which is just $P_1 + P_2$, in agreement with the experiment.

It is very interesting that in the quantum mechanics the amplitudes ϕ are solutions of a completely deterministic equation. Knowledge of ϕ at $t = 0$ implies its knowledge at all subsequent times. The interpretation of $|\phi|^2$ as the probability of an event is an indeterministic interpretation. It implies that the result of an experiment is not exactly predictable. It is very remarkable that this interpretation does not lead to any inconsistencies. That it is true has been amply demonstrated by analyses of many particular situations by Heisenberg, Bohr, Born, von Neumann and many other physicists. In spite of all these analyses the fact that no inconsistency can arise is not thoroughly obvious. For this reason quantum mechanics appears as a difficult and somewhat mysterious subject to a beginner. The mystery gradually decreases as more examples are tried out, but one never quite loses the feeling that there is something peculiar about the subject.

I believe there are a few interpretational problems on which work may still be done. They are very difficult to state until they are completely worked out. One is to show that the probability interpretation of ϕ is the only consistent interpretation of this quantity. We and our measuring instruments

are part of Nature and so are in principle described by an amplitude function satisfying a deterministic equation. Why can we only predict the probability that a given experiment will lead to a definite result? From whence does the uncertainty arise? Almost certainly it arises from the need to amplify the effects of single atomic events to such a level that they may be readily observed by large systems. The details of this have only been analyzed on the assumption that $|\phi|^2$ is a probability and the consistency of this assumption has been shown. It would be an interesting problem to show that *no other* consistent interpretation can be made.

Other problems which may be further analyzed are those dealing with the theory of knowledge. For example, there seems to be a lack of symmetry in time in our knowledge. Our knowledge of the past is qualitatively different than that of the future. In what way is only the probability of a future event accessible to us while the certainty of a past event can often apparently be asserted? These matters again have been analyzed to a great extent. I believe however a little more can be said to clarify the situation. Obviously we are again involved in the consequences of the large size of ourselves and of our measuring equipment. The usual separation of observer and observed which is now needed in analyzing measurements in quantum mechanics should not really be necessary, or at least should be even more thoroughly analyzed. What seems to be needed is the statistical mechanics of amplifying apparatus.

The analyses of such problems are of course in the nature of philosophical questions. They are not necessary for the further development of physics. We know we have a consistent interpretation of ϕ and almost without doubt, the only consistent one. The problem of today seems to be the discovery of the laws governing the behavior of ϕ for phenomena involving nuclei and mesons. The interpretation of ϕ is interesting. But the much more intriguing question is: What new modifications of our thinking will be required to permit us to analyze phenomena occurring within nuclear dimensions?