1 Bits and quanta

1.1 Information and bits

On the evening of 18 April 1775, British troops garrisoned in Boston prepared to move west to the towns of Lexington and Concord to seize the weapons and capture the leaders of the rebellious American colonists. The colonists had anticipated such a move and prepared for it. However, there were two possible routes by which the British might leave the city: by land via Boston Neck, or directly across the water of Boston Harbor. The colonists had established a system of spies and couriers to carry the word ahead of the advancing troops, informing the colonial militias exactly when, and by what road, the British were coming.

The vital message was delivered first by signal lamps hung in the steeple of Christ Church in Boston and observed by watchers over the harbor in Charlestown. As Henry Wadsworth Longfellow later wrote,

> One if by land, and two if by sea;
> And I on the opposite shore will be,
> Ready to ride and spread the alarm
> Through every Middlesex village and farm . . .

Two lamps: the British were crossing the harbor. A silversmith named Paul Revere, who had helped to organize the communication network, was dispatched on horseback to carry the news to Lexington. He stopped at houses all along the way and called out the local militia. By dawn on 19 April, the militiamen were facing the British on Lexington Common. The first battle of the American Revolutionary War had begun.

In the United States, Paul Revere[1] is remembered as a hero of the Revolutionary War, not for his later military career but for his "midnight ride" in 1775. Revere is famous to this day as a carrier of information.

Information is one of our central themes, and over the course of this book we will formalize the concept, generalize it, and subject it to extensive analysis. The story of Paul Revere illustrates several key ideas. What, after all, is information? We can give a heuristic definition that, though we will later stretch and generalize it almost beyond recognition, will serve to guide our discussion.

Information is the ability to distinguish reliably between possible alternatives.

[1] William Dawes and Dr. Samuel Prescott, who accompanied Revere on the ride and helped spread the word, are almost forgotten – perhaps because they were never immortalized in verse by Longfellow.

Before Paul Revere's ride, the militiamen of Lexington could not tell whether the British were coming or not, or by what route. After Revere had reached them, they could distinguish which possibility was correct. They had gained *information*.

We can also distinguish between an abstract *message* and the physical *signal* that represents the message. The association between message and signal is called a *code*. For instance, here is the code used by the colonists for their church steeple signal.

Signal	Message
0 lamp	The British are not coming.
1 lamp	The British are coming by land.
2 lamps	The British are coming by sea.

Mathematically, the code is a function from a set of possible messages to a set of possible states of the physical system that will carry the message – in this case, the lamp configuration in the church tower.

There is no requirement that all possible signals are used in the code. (The Boston spies could have hung three lamps, or six, though their distant compatriots would have been rather perplexed.) On the other hand, the association between message and signal should be one-to-one, so that distinct messages are represented by distinct signals. Otherwise, it is not possible to deduce the message reliably from the signal.

Another very important point is that information can be *transformed* from one physical representation to another, so that the same message can be encoded into quite different physical signals. Paul Revere's message was not only represented by lamps in a church, but also by neural activity in his brain and then by patterns of sound waves as he cried, "The British are coming!"

Exercise 1.1 Identify at least seven distinct physical representations that this sentence has had from the time we wrote it to the time you read it.

The fact that the same message can be carried by very different signals is a fundamental truth about information.

This *transformability* of information allows us to simplify matters considerably, for we can always represent a message using signals of a standard type. The universal "currency" for information theory is the *bit*. The term *bit* is a generic term for a physical system with two possible distinguishable states. The states may be designated *no* and *yes*, *off* and *on*, or by the binary digits 0 and 1. These two states may be distinct voltage levels in an electrical device, two directions of magnetization in a small region of a computer disk, the presence or absence of a light pulse, two possible patterns of ink on a piece of paper, etc. All of these bits are *isomorphic*, in that information represented by one type of bit can be converted to another type of bit by a physical process.

A single bit has a limited "capacity" to represent information. We cannot store an entire book in one bit. The reason is that there are too many possible books (that is, possible messages) to be represented in a one-to-one manner by the two states 0 and 1. Somewhere in the code, we would inevitably have something like this:

Signal	Message
⋮	
0	*Alice in Wonderland* by Lewis Carroll
0	*The Guide for the Perplexed* by Moses Maimonides
⋮	

From a bit in the state 0, it would be impossible to choose the correct book in a reliable way.

To represent an entire book, we must use strings or sequences of many bits. If we have a string of n bits, then the number of distinct states available to us is

$$\#\text{ of states} = \underbrace{2 \times \cdots \times 2}_{n\text{ times}} = 2^n. \tag{1.1}$$

Exercise 1.2 A *byte* is a string of eight bits. How many possible states are there for one byte?

Suppose that there are M possible messages. If the number of possible signals is at least as large as M, then we can find a code in which the message can be reliably inferred from the signal. We can thus determine whether the message can be represented by n bits.

- If $M > 2^n$, then n bits are not enough.
- If $M \leq 2^n$, then n bits are enough.

The number n of bits necessary to represent a given message is a way of measuring "how much information" is in the message. This is a very practical sort of measure, since it tells us what resources (bits) are necessary to perform a particular task (represent the message faithfully).

We define the *entropy* H of our message to be

$$H = \log M, \tag{1.2}$$

where M is the number of possible messages. (From now on, unless we otherwise indicate, "log" will denote the logarithms with base 2: $\log \equiv \log_2$.) The entropy H is a measure of the information content of the message. From our discussion above, we see that if $n < H$, n bits will not be enough to represent the message faithfully. On the other hand, if $n \geq H$, then n bits will be enough. Thus, H measures the number of bits that the message requires.[2]

We can think of H as a measure of *uncertainty* – that is, of how much we do not know before we get the message. It is also a measure of how our uncertainty is reduced when we identify which message is the right one. In other words, before we receive and decode the signal, there are M possibilities and $H = \log M$. Afterward, we have uniquely identified the right message, and the entropy is now $H' = \log 1 = 0$.

We have called H the "entropy," which is the name used for H by Claude Shannon in his pioneering work on the mathematical theory of information. The name harkens back to

[2] Anticipating later developments, we should note here that our definition of H implicitly assumes that the M possible messages are all *equally likely*. To cope with more general situations, we will need a more general expression for the entropy. However, Eq. 1.2 will do for our present purposes.

thermodynamics, and for very good reason. Ludwig Boltzmann showed that if a macroscopic system has W possible microscopic states, then the thermodynamic entropy S_θ is

$$S_\theta = k_{\text{B}} \ln W, \tag{1.3}$$

where $k_{\text{B}} = 1.38 \times 10^{-23}$ J/K, called *Boltzmann's constant*. This famous relation (which is inscribed on Boltzmann's tomb in Vienna) can be viewed as a fundamental link between information and thermodynamics. Up to an overall constant factor, the thermodynamic entropy S_θ is just a measure of our uncertainty of the microstate of the system.

Exercise 1.3 A liter of air under ordinary conditions has a thermodynamic entropy of about 5 J/K. How many bits would be necessary to represent the microstate of a liter of air?

We will have more to say about the connection between information and thermodynamics later on.

Suppose A and B are two messages, having M_A and M_B possible values respectively. The two messages taken together form a joint message that we denote AB. If A and B are independent of each other, every combination of A and B values is a possible joint message, and so $M_{AB} = M_A M_B$. In this case the entropy is additive:

$$H(AB) = H(A) + H(B). \tag{1.4}$$

On the other hand, if the messages are not independent, it may be that some combinations of A and B are not allowed, so that the joint entropy $H(AB)$ may be less than $H(A) + H(B)$.

Exercise 1.4 Suppose A and B each have 16 possible values. What is the joint entropy $H(AB)$ (a) if the messages are independent, and (b) if B is known to be an exact copy of A?

How much information was contained in the message sent from the Christ Church steeple in 1775? There were three possible messages, and so $H = \log 3 \approx 1.58$. This means that one bit would not suffice to represent the message, but two bits would be more than enough. That much is clear; but can we give a more exact meaning to H? Does it make sense to say that a message contains 1.58 bits of information?

Suppose our message is a decimal digit, which can take on values 0 through 9. There are ten possible values for this message, so the entropy is $H = \log 10 \approx 3.32$. We shall need at least four bits to represent the digit. But imagine that our task is to encode, not just a single digit, but a whole sequence of independent digits. We could simply set aside four bits per digit, but we can do better by considering groups (or *blocks*) of three digits. Each group has $10^3 = 1000$ possible values, and so has an entropy of $3 \log 10 = 9.97$. Therefore we can encode three digits in ten bits, using (on average) $10/3 \approx 3.33$ bits per digit. This is more efficient, and is very close to using $\log 10$ bits per digit.

Exercise 1.5 Devise a binary code for triples of digits (as described above) and use your code to represent the first dozen digits of π.

This motivates the following argument. Consider a message having an entropy H. If we have a long sequence of independent messages of this type, we can group them into blocks and encode the blocks into bits. If the blocks have n messages, each block has an

entropy nH. Let N be the minimum number of bits needed to represent a message block. This will be the smallest integer that is at least as big as nH, and so

$$nH \leq N < nH + 1. \tag{1.5}$$

Calculating the number of bits required on a "per message" basis, we are using $K = N/n$ bits per message, and

$$H \leq K < H + \frac{1}{n}. \tag{1.6}$$

If we consider very large message blocks, $n \gg 1$ and so $1/n$ is very small. The two ends of the inequality chain squeeze together, and for large blocks we will use almost exactly H bits per message to represent the information. Therefore, if we encode our messages "wholesale," the entropy H precisely measures the number of bits per message that we need.

Exercise 1.6 Consider a type of message that has three possible values (like the message of the colonial spies in Boston). Calculate the minimum number of bits required to encode blocks of 2, 3, 5, 10, or 100 such messages. In each case, also calculate the number of bits used per message.

Things become more complicated in the presence of noise. *Noise* is a general term for any process that prevents a signal from being transferred and read unambiguously. For example, imagine that there had been fog on Boston Harbor on that April night in 1775. In a heavy fog, the church steeple might not have been visible at all from Charlestown, and no information would have been conveyed. In a lighter mist, the observers might have been able to see that there were lamps in the steeple, but not been able to count them. They would then have known that the British troops were on the move, but not which way they were going. A part of the information would have been transmitted successfully, but not all.

It is possible to formalize this notion of partial information. Before any communication takes place, there are M possible messages and the entropy is $H = \log M$. Afterward, we have reduced the number of possible messages from M to M', but because of noise $M' > 1$. The amount of information conveyed in this process is defined to be

$$H - H' = \log \frac{M}{M'}. \tag{1.7}$$

Exercise 1.7 A friend is thinking of a number between 1 and 20 (inclusive). She tells you that the number is prime. How much information has she given you?

The concept of information is fundamental in scientific fields ranging from molecular biology to economics, not to mention computer science, statistics, and various branches of engineering. It is also, as we will see, an important unifying idea in physics.

1.2 Wave–particle duality

Since the 17th Century, there have been two basic theories about the physical nature of light. Isaac Newton believed that light is composed of huge numbers of particle-like "corpuscles." Christiaan Huygens favored the idea that light is a wave phenomenon, a moving periodic

disturbance analogous to sound. Both theories explain the obvious facts about light, though in different ways. For example, we observe that two beams of light can pass through one another without affecting each other. In the Newtonian corpuscle theory, this simply means that the light particles do not interact with each other. In the Huygensian wave theory, it implies that light waves obey the *principle of superposition*: the total light wave is simply the sum of the waves of the two individual beams.

To take another example, we observe that the shadows of solid objects have sharp edges. This is easily explained by the Newtonian theory, since the light particles move in straight lines through empty space. On the other hand, this observation seems at first to be a fatal blow to the wave theory, because waves moving past an obstacle should spread out in the space beyond. However, if the wavelength of light were very short, then this spreading might be too small to notice. For over a hundred years, the known experimental facts about light were not sufficient to settle whether light was a particle phenomenon or a wave phenomenon, and both theories had many adherents.

Then, in 1801, Thomas Young performed a crucial experiment in which Huygens's wave theory was decisively vindicated. This was the famous two-slit experiment.

Suppose that a beam of monochromatic light shines on a barrier with a single narrow opening, or "slit." The light that passes through the slit falls on a screen some distance away. We observe that the light makes a small smudge on the screen. (For thin slits, this smudge of light actually gets wider when the slit is made narrower, and on either side of the main smudge there are several much dimmer smudges. These facts are already difficult to explain without the wave theory, but we will skip this point for now.)

Light passing through another slit elsewhere in the barrier will make a similar smudge centered on a different point. But suppose two nearby slits are both open at once. If we imagine that light is simply a stream of non-interacting Newtonian corpuscles, we would expect to see a somewhat broader and brighter smudge of light, the result of the two corpuscle-showers from the individual slits.

But what happens in fact (as Young observed) is that the region of overlap of the two smudges shows a pattern of light and dark bands called *interference fringes*, see Fig. 1.1.

This is really strange. Consider a point on the screen in the middle of one of the dark fringes. When either one of the slits is open, some light does fall on this point. But when both slits are open, the spot is dark. In other words, we can *decrease* the intensity of light at some points by *increasing* the amount of light that passes through the barrier.

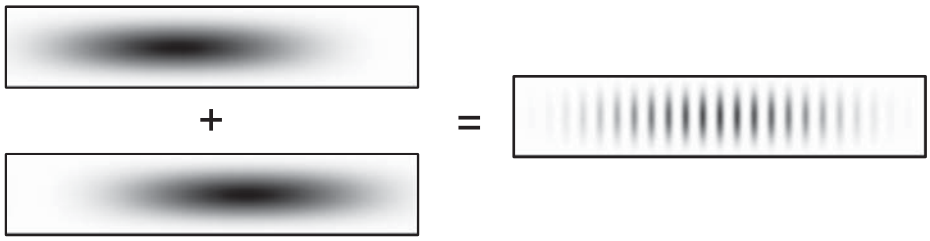

Fig. 1.1 The light patterns from two single slits combine to form a pattern of interference fringes. (For clarity on the printed page, the negative of the pattern is shown; more ink means higher intensity.)

The situation is no less peculiar for the bright fringes. Take a point in the middle of one of these. When either slit is opened, the intensity of light at the point has some value I. But with both slits open, instead of an intensity $2I$ (as we might have expected), we see an intensity of $4I$! The *average* of the intensity over the light and dark fringes is indeed $2I$, but the pattern of light on the screen is less uniform than a particle theory of light would suggest.

Young realized that this curious behavior could easily be explained by the wave theory of light. Waves emerge from each of the two slits, and the combined wave at the screen is just the sum of the two disturbances. Denote by $\phi(\vec{r}, t)$ the quantity that describes the wave in space and time. In sound waves, for example, the "wave function" ϕ describes variations in air pressure. The two slits individually produce waves ϕ_1 and ϕ_2, and by the principle of superposition the two slits together produce a combined wave $\phi = \phi_1 + \phi_2$.

Two further points complete the picture. First we note that ϕ can take on either positive or negative values. By analogy to surface waves on water, the places where ϕ is greatest are called the wave "crests," while the places where ϕ is least (most negative) are called the wave "troughs." Second, the observed intensity of the wave at any place is related to the square of the magnitude of the wave function there: $I \propto |\phi|^2$.

At some points on the screen, the two partial waves ϕ_1 and ϕ_2 are "out of phase," so that a crest of ϕ_1 is coincident with a trough of ϕ_2 and vice versa. At these points, the waves cancel each other out, and $|\phi|^2$ is small. This phenomenon is called *destructive interference* and is responsible for the dark fringes.

At certain other points on the screen, the two partial waves ϕ_1 and ϕ_2 are "in phase," by which we mean that their crests and troughs arrive synchronously. When ϕ_1 is positive, so is ϕ_2, and so on. The partial waves reinforce each other, and $|\phi|^2$ is large. This phenomenon, *constructive interference*, is responsible for the bright fringes.

At intermediate points, ϕ_1 and ϕ_2 neither exactly reinforce one another nor exactly cancel, so the resulting intensity has an intermediate value.

Exercise 1.8 In the two slit experiment, in a particular region of the screen the light from a single slit has an intensity I, but when two slits are open, the intensity ranges over the interference fringes from 0 to $4I$. Explain this in terms of ϕ_1 and ϕ_2.

Young was able to use two-slit interference to determine the wavelength λ of light, which does turn out to be quite small. (For green light, λ is only 500 nm.) Later in the 19th Century, James Clerk Maxwell put the wave theory of light on a firm foundation by showing that light is a travelling disturbance of electric and magnetic fields – an *electromagnetic wave*.

But the wave theory of light was not the last word. In the first years of the 20th Century, Max Planck and Albert Einstein realized that the interactions of light with matter can only be explained by assuming that the energy of light is carried by vast numbers of discrete light *quanta* later called *photons*. These photons are like particles in that each has a specific discrete energy E and momentum p, related to the wave properties of frequency f and wavelength λ:

$$\begin{aligned} E &= hf, \\ p &= \frac{h}{\lambda}, \end{aligned} \tag{1.8}$$

where $h = 6.626 \times 10^{-34}$ J s, called *Planck's constant*. When matter absorbs or emits light, it does so by absorbing or creating a whole number of photons.

Einstein used this idea to explain the photoelectric effect. In this phenomenon, light falling on a metal in a vacuum can cause electrons to be ejected from the surface. If the light intensity is increased, the number of ejected electrons increases, but the kinetic energy of each photoelectron remains the same. In a simple wave theory, this is hard to understand. Why should a more intense light, with stronger electric and magnetic fields, not produce more energetic photoelectrons? Einstein reasoned that each ejected electron gets its energy from the absorption of one photon. A brighter light has more photons, but each photon still has the same energy as before.

Exercise 1.9 The "work function" W of a metal is the amount of energy that must be added to an electron to free it from the surface. Write down an expression for the kinetic energy K of a photoelectron in terms of W and the incident light frequency f. Also find an expression for the minimum frequency f_0 required for the photoelectric effect to take place. (This will depend on W, and so may be different for different metals.)

This "quantum theory" of light poses some perplexities. In view of Young's two-slit interference experiment, there can be no question of abandoning the wave theory entirely. Photons cannot be Newtonian corpuscles. Nevertheless, the fact that light propagates through space as a continuous wave (as seen in the two-slit experiment) does not prevent light from interacting with matter as a collection of discrete particles (as in the photoelectric effect). Furthermore, this bizarre situation is not limited to light. In 1924 Louis De Broglie discovered that the particles of matter – electrons and so forth – also have wave properties, with particle and wave quantities related by Eq. 1.8. It is possible to do a two-slit experiment with electrons and observe interference effects. The general principle that everything in nature has both wave and particle properties is sometimes called *wave – particle duality*.

The effort to put quantum ideas into a solid, consistent mathematical theory led to the development of *quantum mechanics* by Werner Heisenberg, Erwin Schrödinger, and Paul Dirac. Quantum mechanics has proved to be a superbly successful theory of phenomena ranging from elementary particles to solid state physics. It is also a very peculiar theory that challenges our intuitions on many levels. Quantum mechanics involves far-reaching alterations in our ideas about mechanics, probability theory, and even (as we shall see) the concept of information.

To illustrate this in a small way, let us re-examine Young's two-slit experiment with quantum eyes. First, we must understand that the intensity of light is a statistical phenomenon. When we say that light is more intense at one point than it is at another, we simply mean that more photons can be found there. But what can this mean when the number is very small? What can it mean if there is only one photon present?

In the single-photon case, the intensity of the wave at any point is proportional to the *probability* of finding the photon at that point. In general, quantum mechanics predicts only the probability of an event, not whether or not that event will definitely occur. So it is with photons. The behavior of any particular photon cannot be predicted exactly, but the

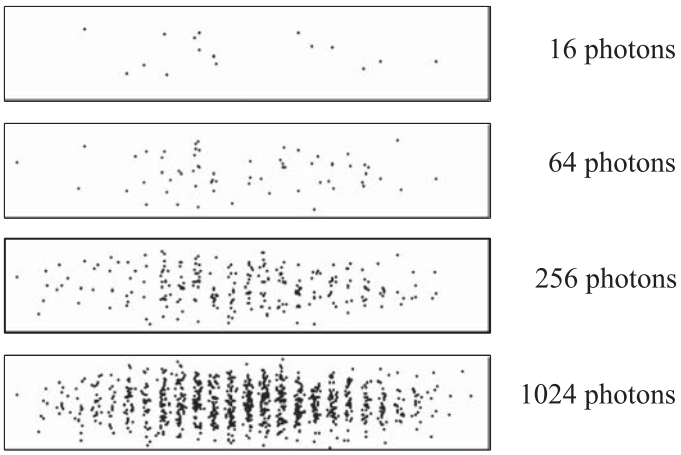

Fig. 1.2 **Photons fall randomly on a screen according to a probability distribution given by two-slit interference. Each image shows four times as many photons as the one before. After many photons, a smooth intensity pattern emerges statistically.**

statistical behavior of a great many photons gives rise to a smooth intensity pattern. See Fig. 1.2 for an illustration of this.

In the single-photon case, therefore, the wave ϕ is actually a *probability amplitude*, a curious mathematical creature that is not itself a probability, but from which a probability may be calculated. Roughly speaking, the probability[3] P of finding a photon at a given point is just $P = |\phi|^2$. Probability is the square of the magnitude of a probability amplitude.

The probability amplitude wave ϕ obeys the principle of superposition. In the two-slit experiment, consider a particular point X on the screen. With only slit #1 open, the probability amplitude that the photon lands at X is ϕ_1, so that the probability of finding the photon there is $P_1 = |\phi_1|^2$. Opening only slit #2 yields an amplitude ϕ_2, which gives rise to a probability P_2 of finding the photon at X. But with both slits open, we have a combined probability amplitude $\phi = \phi_1 + \phi_2$, yielding a probability

$$P = |\phi|^2 = |\phi_1 + \phi_2|^2 , \tag{1.9}$$

for the photon to wind up at X. The two probability amplitudes may reinforce one another or cancel each other out, enhancing or suppressing the probability that the photon lands at X.

If the photon can pass through only one slit, the probability of reaching X is P_1. If it can pass only through the other, it is P_2. In ordinary probability theory, if there are two possible mutually exclusive ways that an event can happen, then the combined probability is $P = P_1 + P_2$. For example, if we flip two coins, the probability that they land with the same face upward is

$$P(\text{same face}) = P(\text{both heads}) + P(\text{both tails}). \tag{1.10}$$

[3] In the two-slit experiment, where the photon can be found in a continuous range of positions, P is actually a probability *density* rather than a probability. This technical detail, and a great many others, will be worked out carefully in later chapters!

But quantum probabilities are not ordinary probabilities! In the two slit experiment, the combined likelihood may be either less than or greater than the sum $P_1 + P_2$, depending on the relative phase of the two amplitudes ϕ_1 and ϕ_2. In other words, quantum probabilities can exhibit destructive and constructive interference effects.

Suppose at a point X on the screen the probabilities P_1 and P_2 both equal p. This means that the probability amplitudes at this point satisfy

$$|\phi_1| = |\phi_2| = \sqrt{p}. \tag{1.11}$$

If the two amplitudes constructively interfere at X, then the two amplitudes are "in phase" there: $\phi_1 = \phi_2$, and so

$$P = |\phi|^2 = |2\phi_1|^2 = 4p. \tag{1.12}$$

If the two amplitudes destructively interfere at X, then $\phi_1 = -\phi_2$ (the amplitudes are "out of phase"). Then $\phi = 0$ and so $P = 0$. We can see that the probability P for finding a photon in this region of the screen will vary over the interference fringes between 0 and $4p$.

Exercise 1.10 Consider a point X on the screen at which $P_1 = p$ and $P_2 = 2p$. That is, with only slit #1 open, the photon has a probability p of reaching X, but with only slit #2 open this probability is twice as great. Now open both slits. What are the largest and smallest possible values for P at X due to interference effects?

When analyzing the behavior of a photon in the two-slit experiment, we find that $P = |\phi_1 + \phi_2|^2$. Yet the conventional probability law $P = P_1 + P_2$ does apply to the two-coin example. So we are faced with an apparent inconsistency. Sometimes we must add probabilities, and sometimes we must add probability amplitudes. How do we know which of these rules will apply in a given situation?

The difference cannot be mere size. Quantum interference effects have been observed in surprisingly large systems, including molecules more than a million times more massive than electrons (see Problem 1.4). Conversely, we can often apply ordinary probability rules to microscopic systems. The essential difference between the two situations must lie elsewhere.

Notice that, in the two-coin example, we can check to see which of the two contributing alternatives actually occurred. That is, we can examine the coins and tell whether they are both heads or both tails. But in the two-slit experiment, this is not possible. If the single photon arrives in one of the bright interference fringes, it could have passed through either of the slits. Even a very close examination of the apparatus afterward would not tell us which possible alternative occurred.

Suppose we were to modify the two-slit experiment so that we could tell which slit the photon passed through. We can for instance imagine a very sensitive photon detector placed beside one of the slits, which is able to register the passage of a photon without destroying it. This detector need not be a large device: a single atom would be enough in principle, if the state of that atom were sufficiently affected by the passing light quantum. With such a detector in place, we could perform the two-slit interference experiment and then afterwards determine which path the photon took, simply by checking whether or not a photon had been detected.