Black holes and information: A crisis in quantum physics

John Preskill

Caltech Theory Seminar, 21 October 1994

(This is a lightly edited transcript of the talk.)

Abstract. According to quantum mechanics, information concerning a physical state can never be destroyed (though it may become inaccessible in practice). Thought experiments involving black holes seem to challenge this principle, and so threaten the foundations of modern physics. If the principles of quantum mechanics are regarded as inviolable, deep insights can be attained into the nature of physics at exceedingly small distances (of order 10^{-33} centimeters) that cannot be directly explored in foreseeable experiments.

PARADOX. When the theories that we use to describe Nature lead us to unacceptable or self-contradictory conclusions, we're faced with a great challenge...and a great opportunity. Real history is always more complicated (and more interesting) than the history that we recount in our physics textbooks, and it didn't really happen this way. But it's fun to imagine a physicist in the late 19th century waking up one morning and saying, "Hey, wait a minute...The principle of equipartition of energy say that an oscillator carries energy kT in thermal equilibrium, and the electromagnetic field in a cavity has an infinite number of oscillating modes. That means that the total energy in the cavity is infinite in thermal equilibrium...Something smells wrong here. Maybe classical physics is flawed. Maybe we need to cast classical physics aside, and seek new foundations for physical theory!"

Well, we are now nearing the end of the first century of quantum theory. Quantum *mechanics*, in much the same form that we still use it and teach it today, has been serving us well for nearly 70 years. But there's trouble on the horizon. (That's a joke, but very few people get it.) This time, we have a revolutionary spokesperson. It's Stephen Hawking. For nearly twenty years, he has been saying that a black hole can destroy quantum-mechanical information. The type of information destruction envisioned by Hawking is in violation of the principles of quantum mechanics. Therefore, says Hawking, quantum mechanics is flawed. It must be cast aside, and we must seek new conceptual foundations for all of physics.

This is a revolutionary claim. We should not accept it without subjecting it to the closest scrutiny. That is what I propose to do in this talk.

Black hole. First of all, we'll need to understand what a black hole is. When an object such as a (sufficiently massive) dying star is unable to resist its own gravitational pull, the result is catastrophic gravitational collapse to an object whose gravity is so strong that nothing can escape from it. The boundary of the region of no escape is called the event horizon of the black hole. Anyone foolish enough to cross the event horizon will be forever unable to return to the region outside the black hole, or even to send a signal to a friend who remains outside.

Black holes are very interesting in astrophysics, but I won't talk about astrophysics today. Instead I will consider how black holes can illuminate questions of principle about fundamental physics.

Light cone. We can gain a deeper understanding of the event horizon by invoking the concept of a light cone. Imagine a light source that emits a flash at a particular point and at some instant of time. The pulse travels outward from that point as a spherical shell that expands at light speed. If we plot the position of the expanding shell as a function of the time, we obtain a cone in spacetime, called the future light cone of the point P where the original flash event occurred. The significance of the future lightcone is that, since no signal can travel faster than light speed, all of the events that can be influenced by the event P—all points in spacetime where a signal can be received that is emitted at P—lie inside the future lightcone.

Tipping of the light cones. In the case of the black hole, we find by solving the Einstein field equations that govern gravitation in general relativity that the lightcones tip inward as one approaches the black hole. Inside the event horizon, this tipping of the lightcones becomes so extreme that the whole future light cone lies inside the horizon. Any signal that gets emitted from a point inside the horizon necessarily travels more deeply into the black hole. The unfortunate astronaut who enters the black hole is inevitably drawn to the *singularity*, a region of enormous gravitational forces. As he approaches the singularity, he is stretched in one direction and compressed in the other, until he is ultimately torn apart. Not a pleasant way to go.

Gravitational redshift. Another useful way to characterize the event horizon is: it is a surface of infinite gravitational redshift. Here's what that means. Someone near the black

hole, if he does nothing to prevent it, is gravitationally drawn toward the black hole, and quickly plummets through the horizon. To remain a fixed distance above the horizon, you must turn on the thrust of your rocket motor—in other words, you must accelerate. Since the light cones tip more and more as you get closer and closer to the horizon, you require more thrust to maintain your position as you lower yourself closer to the horizon, until, right at the horizon, infinite thrust (or infinite acceleration) is required.

Now consider two of these static observers (in rockets) who are maintaining their distance from the horizon. Suppose that an observer one meter above the horizon has a source that emits a light signal that is detected by another observer who is one centimeter above the horizon. Now suppose that yet another observer, who has turned off his rocket engine so that he is falling freely toward the horizon is watching this process. To the freely falling observer, the two other observers are *not* static; they're accelerating. Suppose that, as seen by this freely falling observer, the source and the detector are both at rest at the time that the signal is emitted. Then, because of the acceleration, the detector is rushing toward the source at the (later) time that the signal is detected. As a result, the time and distance between the successive wave crests of the signal is compressed—the signal is shifted toward the blue. Conversely, if the source is one centimeter above the horizon, and the detector is one meter above, the signal is shifted toward the red. That's the gravitational redshift. For a source getting closer and closer to the horizon, the amount of the redshift becomes greater and greater, reaching an infinite redshift at the horizon. In other words, as a static observer watches a clock being quasistatically lowered toward the horizon, the clock seems to run slower and slower, with time finally freezing right at the horizon.

What does the static observer see as the freely falling observer plunges through the horizon? Because of the slowing of time, the freely falling observer seems to fall more and more slowly—he never gets to the horizon, but only asymptotically approaches it, with the signal that he emits becoming shifted more and more to the red. But the freely falling observer knows nothing about this freezing of time at the horizon. From his perspective, he rushes past the horizon unimpeded, and soon thereafter meets his doom at the singularity. This extreme difference between the description of the same process in two different frames of reference is responsible for many of the bizarre and amazing properties of black holes.

A black hole has no hair. One remarkable property of a black hole is that (in the memorable aphorism coined by John Wheeler) "a black hole has no hair." This means the following. Though the process of gravitational collapse to form a black hole may be

extremely complicated, the end result, after the black hole has settled down to a time-independent (stationary) state, is astonishingly simple. All properties of the black hole can be completely characterized by just a few numbers—its mass, angular momentum (a black hole can spin), and electric charge (if it has any).

Ordinary objects have hair. The sun, for example, is not perfectly spherical. It has a complicated shape, and gravitational multipoles that could in principle be detected from afar if we measured the falloff of the sun's gravitational field with distance. But the gravitational field of a black hole has no higher multipoles. This property is a consequence of the fact that the horizon (in a sense that should not be taken too literally) is infinitely far away. It isn't really true—we've seen that a falling observer can reach the horizon in a finite time. But consider measuring distance using what we might call the "optical metric." That is, adopt as the unit of length the wavelength of an infalling light signal as measured by static observers. Because of the gravitational redshift, this wavelength get shorter and shorter closer and closer to the horizon, so distance, in terms of the optical metric, gets stretched out more and more, and the horizon is infinitely far away in these optical units.

The significance of the optical metric is that it is the natural unit of length to use when we integrate field equations on the black hole background. Now suppose that the gravitational field of a stationary black hole had a higher multipole falling off like $1/r_*^l$, where r_* is the distance in optical units. If we now try to extend this multipole inward toward the black hole, we find that it blows up to infinity at the horizon, since the horizon is infinitely far away in optical units. That infinite behavior isn't acceptable. So what happens instead is that the black hole refuses to be stationary. It is a time-dependent configuration that quickly radiates away the multipole—either pushes it away to infinity or swallows it up through the horizon. The final stationary state has no hair. Although a black hole can be macroscopic, it rivals an elementary particle in its simplicity (at least in classical general relativity).

Black hole radiance. Another remarkable property of a black hole is that it is not black. Although in classical general relativity a black hole is an object from which nothing can escape, quantum-mechanical effects allow a black hole to *radiate* (as Stephen Hawking discovered in 1974). Why is this so? In quantum theory, the vacuum is an interesting and busy place, with pairs of virtual quanta being continually created in pairs, and then annihilating back into the vacuum. Typical virtual quanta with wavelength of order λ can separate by a distance of order λ before they must recombine and annihilate. Consider how such a quantum fluctuation would be perceived by a static observer close to the event horizon

of a black hole, if one member of the virtual pair happens to be behind the horizon. This observer never sees the pair annihilation, and has no choice but to interpret the quantum that he does see as a *real* quantum. A static observer a distance λ above the horizon will see a bath of real quanta, with a typical wavelength of order λ (and in fact, it turns out that the bath has a thermal spectrum with a characteristic temperature). Actually, this phenomenon has not so much to do with black holes as with the *acceleration* of the observers—an astronaut in a rocket accelerating through this room would also see such a bath. And the bath is not a mathematical fiction, but a real physical effect. If the astronaut were carrying a block of ice with him, the ice would melt.

This claim may surprise you if you have a really hot car, and you have sometimes floored the accelerator when the light changed. You probably didn't find that your eyes were suddenly flooded by a bath of thermal photons. To understand why not, we better put in some numbers. For an acceleration equal to one earth gravity, the temperature of the radiation is about $10^{-20}\ ^\circ K$. (The inverse frequency of a typical quantum is the time it would take the rocket to accelerate from rest to a speed of the order of the speed of light.) It's a small effect, but it is still interesting.

So the static observers outside the black hole see that the horizon is surrounded by a gas of radiation quanta. Close to the horizon, typical quanta in this gas are deflected by the gravity of the black hole so that they fall back toward the horizon. However, at a distance from the horizon comparable to the size of the black hole, the quanta have a reasonable chance of escaping, if they are directed close enough to the vertical. The black hole is surrounded by what Kip Thorne called a "thermal atmosphere," and the atmosphere leaks away very slowly. The radiation that leaks out of the atmosphere and escapes to infinity is the black hole radiation that Hawking discovered. The typical quanta that escape have a wavelength of the order of the size of the black hole, so an observer far away detects a flux of thermal radiation with a temperature corresponding to this wavelength.

Black hole entropy. From the phenomenon of black hole radiance, we can infer another remarkable property: A black hole has an enormous intrinsic *entropy*. We can regard a process in which a black hole emits or accretes a little bit of radiation as a reversible thermodynamic process, and then apply the thermodynamic identity that says that the change in the entropy of the black hole is the amount of energy that it emitted or absorbed, divided by its temperature. We know how to express the energy of the black hole in terms of its size, and we now also know how to express its temperature in terms of its size. So we

can integrate this identity, and find the black hole entropy (up to an additive constant of integration). The result is a beautiful formula. The entropy is one quarter the *area* of the event horizon of the black hole, divided by the Planck unit of area. The Planck unit of area is the square of the *Planck length*—the quantity with the dimensions of length that can be constructed from the fundamental constants: Planck's constant $\hbar$, the speed of light c, and Newton's gravitational constant G. This Planck length turns out to be remarkably small, of order 10^{-33} centimeters. Correspondingly, the entropy of a black hole is enormous—for a solar mass black hole it is about 10^{78}, which is about 20 orders of magnitude larger than the entropy of the sun. So a black hole provides a remarkably efficient way to squeeze a great amount of entropy into a small region.

Planck scale. The Planck length that appeared in this formula is a very interesting length from the point of view of fundamental physics. At this distance, and at the corresponding energy scale of order 10^{19} GeV (the Planck energy), the gravitational forces between elementary particles become very strong. This is also the distance scale at which our usual ideas about space and time cease to apply, because spacetime itself, at the Planck scale, is subject to very large quantum fluctuations. To see why, suppose that we want to resolve the structure of space at this scale. We need to build a microscope that uses very short wavelength particles. But these particles have a large gravitational effect on the region that we are trying to probe, and so disturb the structure that we are trying to see. So space itself is subject to large uncontrollable quantum fluctuations at the Planck scale.

Well, 10^{19} GeV is a bit beyond the capacity of today's particle accelerators—we'll need to wait for the *Planckatron*. Though it is only about 15 orders of magnitude above SSC energies, the Planckatron would probably be more than a factor of 10^{15} more expensive. So we won't be exploring the physics of the Planck scale *directly* in accelerator experiments any time soon. Still, we can anticipate that a qualitatively new sort of physics sets in at this scale.

Notice that there is a certain tension between two of the things that we have learned about black holes. First, black holes have no hair—they are remarkably simple objects with very little structure. Second, black holes have enormous entropy. That's odd. Ordinarily, large entropy is associated with *complexity*. In quantum statistical physics, large entropy is associated with a large number of accessible quantum states to which a system can fluctuate in thermal equilibrium. So which is it? Are black holes simple or are they complex?

I will return to this question, but first I want to talk for a few minutes about the concept

of *information* in quantum physics.

Quantum information. The fundamental unit of information is the *bit*: 0 or 1, on or off. But in quantum theory there is an interesting generalization: the *quantum bit*. We can regard 0 and 1 as the elements of an orthonormal basis for a two-dimensional complex vector space. The state of the quantum bit can be an arbitrary linear combination (of unit norm) of 0 and 1. The interpretation is that, if we measure the bit, it will read 0 with probability $|a|^2$ and 1 with probability $|b|^2$ (where $|a|^2 + |b|^2 = 1$). But there's more to it than that.

Physicists prefer to think of the two states of the bit not as 0 and 1, but as the two states in which the spin of an elementary particle like the electron point either up or down along, say, the z-axis. Why is that? Is it because physicists are unable to think of things in abstract terms? That's not the only reason. Thinking of a spin gives us another interpretation of this linear combination. For any complex a and b, there is some oblique axis such that, if we measure the spin along that axis instead of the z-axis, the answer will be 0 (say) with 100% probability.

Quantum bits, like classical bits, are good for storing information. Suppose I want to record volume A of the *Encyclopedia Britannica*. It's easy. First, I convert the encyclopedia to ASCII, a string of 0's and 1's. Then I line up a bunch of spins, and I put each one either pointing up (0) or down (1) along the z-axis. Then I can come back the next day, and I can read the encyclopedia by measuring all of the spins along the z-axis. But there's a funny thing about quantum information—to read it you have to know what you're doing. Suppose that I have a dumb friend, and he tries to read the encyclopedia by measuring all of the spins along the *x-axis*. Well, he reads each bit as either 0 or 1, each occurring with probability 1/2. So he finds nothing but a sequence of random bits—there's no information at all! Not only that, but if I come back the day after to look up an article in the encyclopedia, I can't read it any more (even though I know what I'm doing), because my friend futzed it up!

Hidden information. A more subtle type of quantum information emerges when we think about two quantum bits—the bits can be *entangled*. We can prepare a state of two spins in which the spins are perfectly *anticorrelated* with each other. If spin A is up then spin B is down, and vice-versa. Furthermore, this is true no matter how we choose the axis. But in this state, if I measure spin A, it carries no information at all. For *any* choice of the axis, I just get a random bit, either up or down, each with probability 1/2. Same for spin B. But I have two quantum bits, which should be able to encode two bits of information. Where *is* this information?

The information is in the *correlations* between the spins. There are actually 4 states like this one, in which the spins are perfectly correlated or perfectly anticorrelated. All 4 behave the same way if we measure the spins one at a time—along any axis, each spin has probability 1/2 of being up or down. We could have prepared the system in any one of the 4—that's two bits of information. But there is no way to recover the information if we measure the spins only one at a time.

If N spins carry N bits of information, we say that the spins are in a "pure" state, and if there is information missing, we say that the state is "mixed." Here, the full system of two spins is in a pure state, but the state of either spin by itself is mixed. The amount of missing information is quantified by the "entropy." Each spin here has entropy 1—one bit of missing information.

Quantum weirdness. Having mentioned these correlated states, I would be remiss if I didn't take a minute to mention, in passing, the essential weirdness of quantum information. Suppose we prepare our correlated state of two spins, and then take one of the spins far away, say to the Andromeda galaxy. We're careful to preserve the entanglement of the spins along the way. Now we have a strategy for measuring the spin that is left behind here in Pasadena along two different axes. First, we measure the spin in Andromeda along the z-axis, say. We don't measure the spin in Pasadena along the z-axis, we don't need to. It's perfectly anticorrelated with the spin in Andromeda, so we know for sure what the result *would have been* if we had measured the Pasadena spin along the z-axis—we just wait for a telegram to arrive from Andromeda with the results of the measurement. Instead, we measure the Pasadena spin along a different axis. One thing's for sure, nothing that our friends did in Andromeda could affect the outcome of the measurement we made in Pasadena. So we can find out the result that would be obtained for measurements of the spin along the two different axes.

Suppose I and my Andromeda friend choose three different axes, each axis differing from the other two by 120° rotations. Think of the spin along each of the three axes as a quantum-mechanical coin—it comes up either heads or tails. The three coins are lying on the table, but they're covered up so we can't see which side is up. We can uncover any two of them (I measure my spin along one axis, and my Andromeda friend measures his spin along one of the other axes.) But when we uncover two coins, the third one always disappears before we can look at it (I never get a chance to measure the spin along the third axis).

According to quantum mechanics, if we uncover any two of the coins, the probability that

they come up the same (both heads or both tails) is 1/4. Well, one thing for sure we know about real coins—if we uncover all three, at least two have to be the same. For quantum-mechanical coins, we can never uncover all three. If we could, though, the probability that at least two of the three coins are the same has to be less than the sum of the probabilities for each pair to be the same, summed over the three possible ways of choosing the pair. But $\frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4} < 1$. The probability that two of the three coins are the same is less than 1. Weird!

The moral is that it's wrong, it's mathematically inconsistent, to assign simultaneous probabilities to the outcome of the measurement of the spin along two different axes. And that despite the fact that the spin can be perfectly anticorrelated with a spin in Andromeda. Nothing in quantum theory has caused more consternation among thoughtful people than this observation, which is known as Bell's theorem. But it is a fact of Nature, verified by experiment, so we all better just try to get used to it. Okay?

Unitary evolution. Now suppose that I've encoded the encyclopedia in my spins. But what I don't realize is that the spins are coupled together—they interact. So when I come back to read the encyclopedia the next day, the spins are no longer in the state that I had so carefully prepared—the state has *evolved*. In quantum mechanics, time evolution is just a rotation of the orthonormal basis in the the vector space of possible states, a unitary transformation. So the evolution isn't a big deal. For N spins there are 2^N mutually orthogonal states, and I could have prepared any one of them. That's N bits of information. After the system has evolved, it is still in one of 2^N mutually orthogonal states, the basis has just been rotated. So I perform measurements to determined which of these 2^N *rotated* states the system is in. That way, I recover the N bits of information.

The trouble is that what typically happens when the state evolves is that the spins become *entangled*, correlations are established involving many spins. I encoded the information originally in a nice clean way, with one bit carried by each spin. But after the evolution, there's no way to recover the information by measuring one spin at a time. It's worse then that. If there are a million spins, I might know everything that there is to know about the quantum state of, say, one hundred thousand of the spins (including all of the correlations among those hundred thousand). Yet, I might still be unable to read even a single bit of the encyclopedia! All of the information is distributed now in a highly nonlocal way among nearly *all* of the spins. Only by carrying out exceedingly intricate measurements of correlations among huge numbers of spins will I ever be able to decipher the encyclopedia!

This tendency of quantum information to become harder to read as a system evolves is what we call "thermalization." It is the origin of the second law of thermodynamics. But please note that no information has been destroyed; in principle we could still read the encyclopedia. It has just become very difficult.

Quantum copy machine? There is another very important thing to know about quantum information: you can't copy it. Suppose I want to devise a machine that will make a perfect copy of a single quantum bit. The machine has two input slots. Into slot A I insert a spin with its spin aligned along an arbitrary axis. In slot B I place a spin that reads 0; it points up along the z-axis. Now the machine is to leave the spin in slot A alone, but rotate the spin in slot B so that it lines up with the other spin. Now I have two copies of the original quantum state that I placed in slot A.

But there's no such machine. It's no problem to build a machine that will copy a quantum bit that I *know* is either pointed up or down along the z-axis; that's no harder than copying a classical bit. But now suppose that I feed this machine a spin that points along the x-axis, a sum of the up state and the down state along the z-axis. In quantum mechanics, time evolution is a linear transformation, so I know what my machine will do. The output is the sum of the output when I fed the machine the up state and the output when I fed it the down state (along the z-axis). But that's not what we wanted at all! the output is an *entangled* state of the spins in the two slots. Instead of two copies of the same information, we have a correlated state. There's no information in either slot, it's all in the correlations between the slots.

Now I can come along and erase slot A, rotate it to 0. Then I finally have what I wanted—a copy, in slot B, of the original state. But I succeeded in making the copy only at the cost of erasing the original quantum bit. In quantum mechanics, you cannot make a copy unless you destroy your original! A Xerox machine that did that would not be very useful. Probably not even patentable.

Maybe it's a good thing that you can't build a quantum copy machine. If you could, it would be no problem to turn it into an acausal telephone. It's easy! My friend in Andromeda and I have our perfectly anticorrelated spins. We agree in advance that my friend will measure his spin along either the x-axis or the z-axis. My job is to find out which he chose. When I find out, he has sent me one bit of information. Well, I just put my spin through the quantum Xerox machine, and make a few copies of it. Then I measure some of the copies along the z-axis and some along the x-axis. If he measured his spin up along the z-axis,

then every one my copies will be down along the z-axis when I measure them. And vice versa. Same for the x-axis. As soon as I find two copies that give *different* values when I measure them both along the z-axis, I know that he must have made his measurement along the x-axis. Or if two of my copies give different values along the x-axis, I know he measured along the z-axis. He has successfully sent me a bit.

But instantaneous communication with Andromeda is not allowed in relativity. In fact, by a simple feat of engineering, we could turn this acausal telephone into a time machine (which might make Kip Thorne happy). So it makes sense that we encountered an obstacle when we tried to build the quantum Xerox machine.

I think it is really interesting that, when we try to fool around with quantum mechanics by making it nonlinear, we tend to run into problems of this kind. Perhaps this enables us to understand why quantum mechanics *had* to be linear.

The Paradox. Now we know enough about black holes and information to appreciate Hawking's paradox. Suppose I take my N spins on which volume A of the *Encyclopedia Britannica* has been encoded. I arrange these spins so they form a pressureless dust, poised on the brink of gravitational collapse. Then I let them go. They collapse, forming a black hole. This black hole begins to radiate. The radiation that it emits, according to Hawking, is completely featureless thermal radiation; it carries absolutely no information about the encyclopedia. After all, a black hole has no hair, so a black hole formed from the collapse of volume A is just like a black hole formed from volume B. Furthermore, the spins are now behind the horizon, out of causal contact with the escaping radiation, and so the encyclopedia cannot exert any influence on the radiation.

At first this is neither distressing nor disturbing. We are looking only at part of the full quantum system. The radiation carries no information because all of the information is encoded either in the spins that are behind the horizon (where we can't see them), or in correlations between the radiation and the spins. But suppose that we wait until the black hole has radiated away most of its mass. This may take a long time—for a solar mass black hole it's about 10^{66} years—but we have nothing better to do, so we wait. Now the black hole is microscopic and very hot. It is evaporating fast, getting smaller and smaller, hotter and hotter. There seems to be nothing to prevent it from radiating away all of its mass and disappearing completely. Suddenly (poof!) it's gone.

Now we're in trouble. We can no longer point to the black hole and say that the

information that we're missing is behind the horizon, or in correlations with whatever is behind the horizon. There's no horizon anymore. The information that was encoded in the spins seems to have been irrevocably lost, and there is no conceivable measurement that we could perform on the radiation that would allow us to recover that information. A pure quantum state (with lots of information) has evolved to a mixed quantum state (with lots of *missing* information).

This is the information loss paradox. It is a paradox in the sense that we have followed where the principles of general relativity and quantum mechanics seemed to be leading us, and we have concluded that a pure state can evolve to a mixed state, which is in violation of the principles of quantum mechanics.

Pragmatic view: A pragmatic physicist might react to this story with scorn. "I don't believe this at all," he might say. "I don't see why a black hole should be any different than any other body that emits thermal radiation. And I don't recall anyone ever claiming that emission of thermal radiation violates the principles of quantum mechanics!"

Well, how does it work for other bodies? Suppose that, instead of a black hole, it's a black rock, a lump of coal floating in the perfect vacuum of space. The coal is initially at zero temperature, occupying its unique quantum-mechanical ground state. Now I come flying by in my rocket one day, carrying my laser pistol, and, feeling a little trigger happy, I open fire on this poor rock. By sending a sequence of red and blue photons, I can encode volume A of the *Encyclopedia Britannica*. The rock absorbs the photons, and it warms up. A warm rock radiates. The thermal radiation that it emits carries no information at all (at least at first). There is nothing surprising nor disturbing about that. All of the information about volume A of the encyclopedia is stored either in the internal state of the excited rock, or in the correlations between the state of the rock and the state of the thermal radiation. But we wait for a while. As the rock continues to radiate, it cools down. After a while, most of the energy of the original laser pulses has been radiated away. As a result, the number of accessible internal quantum states that the rock can occupy is much smaller than before. It is no longer possible to encode all of volume A in either the internal state of the rock, or in the correlations of the internal state with the radiation. There just are not enough possible orthogonal states for that.

What must happen, then, is that if we could learn the precise quantum state of the radiation, we would find that the state eventually starts to encode information. If we plot the entropy of the radiation as a function of the total energy of the radiation that has been

emitted, we find that the entropy increases for a time—there is lots of missing information. But after a while, the entropy turns around and starts to decline. More and more information is showing up in the radiation. Finally, the rock cools back to absolute zero. It has returned to its unique quantum-mechanical ground state, completely uncorrelated with the radiation. At this stage, all of the information has been encoded in the radiation—its state has become pure, and its entropy is zero.

The information has not been lost, it has just been *thermalized*. It is encoded in incredibly intricate correlations among many many quanta. The information has become almost impossible to decipher in practice, but it could still be read in principle. No violation of the principles of quantum mechanics has occurred.

Is a black hole like a rock? So the question has become, "Is a black hole like a rock?" No. A black hole is different than a rock. The difference is that a black hole has an event horizon. A rock does not.

This spacetime diagram depicts the collapse of the spins to form a black hole, and the subsequent emission of the Hawking radiation. If a black hole behaves like a rock, then, when most of the mass of the black hole has been radiated away, the Hawking radiation is laden with information. If we only knew how to read it, the encyclopedia would be coming into focus! The green surface in the diagram represents a time when most of the encyclopedia can be recovered from the radiation.

But time in the vicinity of a black hole is peculiar. At the very same time that we are reading the encyclopedia in the Hawking radiation, the original spins are still intact behind the event horizon, and the encyclopedia can be read there, too. It may not look in the diagram like this is the same time, but we need to remember the extreme tipping of the light cones inside the horizon. No point on the green surface is in the future light cone of any other point on the surface, so we can regard it as representing a single tick of a set of clocks suitably distributed throughout space. This is just another way of saying that the spins are out of causal contact with the radiation.

But look what this means. If a black hole really behaves like a rock, then the information must be in two different places at the very same time. A quantum copy machine has operated! But we have seen that such a copy machine is not allowed in quantum mechanics.

If the Hawking radiation really approaches a pure state, then logic dictates that the information that was originally encoded in the spins must have been erased right at the

event horizon. You can't make a copy of a quantum state (in the Hawking radiation) unless you destroy your original (encoded in the spins). A mysterious force must bleach the pages of the book white at the horizon, so that once inside the horizon, volume A has become indistinguishable from volume B. Otherwise, the illicit copying of quantum information has occurred.

But that's absurd! We can adopt the viewpoint of a freely falling observer. That observer can be reading the book as he crosses the horizon, and he knows very well that nothing bleaches the pages white at the horizon. Nothing special happens there. An astronaut, upon crossing the event horizon of a black hole, might be overcome with an acute sense of forboding, but he would not be immediately reduced to his quantum-mechanical ground state!

So, we're stuck. Maybe Hawking is right, and information really is irrevocably lost.

What's so terrible...? Well, what's so bad about that? So quantum mechanics seems to fail in this esoteric thought experiment. Why not just accept that quantum mechanics as we know it cannot apply to this kind of extreme situation?

Something deep in my bones makes me resist this conclusion. Am I just being a reactionary? Maybe. Stephen Hawking says so. On the other hand, we are being asked, on the basis of this esoteric thought experiment, to disavow the foundation on which physics now stands. And in return, we are offered little guidance about how to establish a new foundation on which we can begin to rebuild. We are being asked to repudiate the notion of quantum–mechanical determinism—at least in some cases, we can not say how an initial quantum state will evolve to a final state; we can only assign probabilities to various alternatives.

There is also a technical problem: information loss is highly infectious. It is very hard to modify quantum theory so as to accommodate a little bit of information loss without it leaking into all processes including ordinary ones (having nothing to do with black holes) that we can study in the laboratory. And there is no reason for the violations to be small. So far, no one has been able to formulate a satisfactory generalization of quantum mechanics that can accommodate *some* information loss without admitting *so much* information loss as to be in flagrant disagreement with experiment. One generic problem is that in theories that admit information loss, energy is typically not conserved. We can understand that in a heuristic way. Information loss is rather like coupling the universe to a source of random noise that can overcome the signal encoded in a quantum state. And a random noise source

will heat a system up by pumping energy in.

I think it is healthy to take the attitude: let's suppose that quantum mechanics is really okay. Then what can we learn about physics by thinking about evaporating black holes?

The conflict. But how can I say that? Didn't I convince you just a minute ago that information loss could not be avoided?

Recall the irreconcilable conflict that we uncovered: If a black hole really behaves like a rock, then an observer who stays outside the black hole insists that information must be erased as the encyclopedia crosses the horizon. Yet an observer who falls freely through the horizon *with* the encyclopedia knows very well that it's not so!

But now we must ask, can they ever compare notes? Can they get together to pool their observations and infer that something illicit has happened? In fact, it is very difficult for them to compare their observations. It might not even be possible.

As we saw in our discussion of the rock, the thermal radiation does not start to carry information right away. We need to wait until about half of the radiation has been emitted before we can start to reconstruct the bits of the encyclopedia. In the case of the evaporating black hole, we need to wait until the black hole has radiated away about half of its mass. Suppose that an astronaut who is initially outside the black hole waits outside long enough to verify that quantum information is encoded in the Hawking radiation. Then he dives into the black hole to check that the encyclopedia has not been erased. In this experiment, he could verify to his own satisfaction that copying of quantum information has occurred (though he wouldn't be able to tell us about it, if we stayed outside).

However, verifying that the encyclopedia is still intact is not at all easy. We can program the encyclopedia to send out a beacon: "I'm still here! I'm still here! I'm still here! ..." The problem arises from the tipping of the light comes inside the horizon. Any signal emitted from the encyclopedia well after it passes through the horizon will have no chance of reaching our astronaut who enters the black hole much later. That signal will fall to the singularity much too quickly, and who knows what happens to it then? For the signal to reach the astronaut, it must be emitted very soon after the encyclopedia reaches the horizon. In fact the time available turns out to be much less than the Planck time! But that means that the signal must be encoded in quanta with frequency greater than the Planck frequency. Therefore, until we understand quantum gravity better than we currently do, we can't be certain that such a signal can really be sent and received. Perhaps, even if a black hole

behaves like a rock, no conceivable observer can verify that quantum information has really been copied. It might be all right for quantum information to be copied, if no one can ever find out! Indeed, if no one can ever find out, perhaps it didn't really happen (in any operationally meaningful sense).

I find this to be a liberating thought. The moral is that it is very difficult to patch together and reconcile the descriptions of the two observers, one who stays outside the black hole, and one who falls in. And the crux of the information loss problem was that we couldn't see how both descriptions could be valid simultaneously. Well, if comparing their observations is really a question for quantum gravity, let's not even try (at least not for now). Let's forget about the poor fellow who fell in. It's too bad what happened to him, but let's not worry about it. Let's just stick with the outside description, and now ask, again, if a black hole can behave like a rock. Is there a consistent description of physics outside the black hole in which information is encoded in the Hawking radiation? Never mind the inside!

Sub-Planckian quantum fluctuations. Once we resolve to stick with the outside description only, we soon recognize that Hawking's conclusion that information is lost is really predicated on implicit assumptions about quantum gravity. Because of the gravitational redshift, to the outside observer, the Hawking radiation seems to originate as very short wavelength quanta very close to the event horizon. Since time freezes at the event horizon, the quanta seem to stick close to the horizon for a long time, then gradually pull away, being redshifted all the while. In fact, you don't need to wait very long before the radiation that's coming out seems to have started out with a frequency greater than the Planck frequency and at a distance from the horizon less than the Planck length. Since gravitational interactions are strong at the Planck scale, the quanta appear to emerge from a layer of strongly interacting "Planckian goo," about a Planck length thick, clinging to the horizon. To understand the detailed microscopic properties of the Hawking radiation, we'll need to understand the quantum physics of Planckian goo. This is a (hard!) problem in quantum gravity.

One's impulse is to protest: What nonsense! This Planckian goo is a complete hoax. It results from adopting the frame of reference of unphysical observers with Planckian acceleration who hover a Planck length above the horizon. Probably, there can be no such observers, even in principle. Freely falling observers plunge right through the horizon without ever seeing any sign of any Planckian goo. It doesn't exist.

Well, I don't care! Forget the freely falling observers—they aren't going to help us understand the information content of the Hawking radiation. If we are to address the question of information loss, we have no choice but to adopt the viewpoint of these highly accelerated observers near the horizon. We've agreed to stick with the observations of those who always stay outside the horizon, and to them Planckian goo *is* real. I don't care what the other guys see!

Now, from this point of view, Hawking's conclusion that information is lost follows from a particular assumption about the Planckian goo. Hawking's (implicit) assumption is that Planckian goo can encode information in quanta of arbitrarily short wavelength. That is a natural assumption in normal quantum theory—there are vacuum fluctuations of arbitrarily small wavelength. Vacuum fluctuations with wavelength much less than the Planck scale look like real sub-Planckian quanta to the static observers, quanta that carry no information, because they are correlated with quanta on the other side of the horizon. In this picture, these quanta eventually pull away from the horizon, and are detected far away as Hawking radiation that carries no information. Information loss occurs because quanta of arbitrarily small wavelength contribute to the entropy (missing information) of the Planckian goo.

But perhaps it doesn't have to be that way. Perhaps in the *right* theory of quantum gravity, the Planckian goo can *not* encode information in modes of arbitrarily small wavelength. In this theory, it may be that the amount of information in the Planckian goo cannot exceed

$$\text{Entropy} = \frac{1}{4}\frac{\text{Area}}{L^2_{\text{Planck}}}$$

This is an *a priori* plausible result—it says that the Planckian goo can store about one bit of information per Planck area. If this is correct, we have a satisfying picture, from the point of view of the observers who always remain outside, of how a black hole can be like a rock. When the black hole forms, the encyclopedia becomes encoded in the quantum state of the Planckian goo. As the black hole radiates and shrinks, the number of accessible states of the Planckian goo declines. Eventually the goo has so few states that information must begin to appear in the outgoing Hawking radiation. As the black hole explodes and disappears, all of the initial information has been transferred to the radiation.

Now, we have a highly attractive candidate for a theory of quantum gravity, superstring theory. And superstring theory has some properties that are very attractive in the present context. In superstring theory, all elementary particles are excitations of an extended object

of Planckian dimensions, a closed loop of string. The strings are not hard pointlike objects, they are soft and squishy. If I try to study the structure of an object by scattering strings off of it, the squishy strings cannot resolve structure on scales small compared to the Planck scale. So it doesn't really make sense to talk about sub-Planckian distances in string theory. And it is therefore plausible that the Planckian goo cannot store information in sub-Planckian quanta.

Furthermore, string theory offers a concrete and suggestive picture of the Planckian goo, which has been developed recently by Lenny Susskind, of Stanford. Susskind envisions virtual string loops in the vicinity of the horizon. Some of these virtual loops intersect the horizon. To the outside observers, each such string loop looks like a finite string segment, with both ends stuck at fixed positions on the horizon, due to the freezing of time at the horizon. These string segments wiggle and wave. There are arguments indicating that the counting of the microscopic states of this string soup agrees with the black hole entropy formula. However, these calculations are not yet fully convincing.

Do black holes destroy information? The big question is, do black holes really destroy information? Is there really trouble on the horizon?

I don't know for sure. But we can say that, if so, then we need to seek a new self-consistent formalism, a new conceptual foundation for all of physics. It is not clear how to proceed with the search for the new formalism. One hopes that we can be guided by experiment (which was so important to Planck and those who followed him). So far, though, there is no experimental evidence for a breakdown of quantum mechanics. Perhaps such evidence will eventually be found, and will guide us toward a new synthesis.

But there is another possibility, which to my mind is more likely...that information is *not* lost in the *right* theory of quantum gravity. Perhaps superstring theory will be shown to have the right properties to allow black hole evaporation to be reconciled with quantum mechanics.

It it indeed turns out that a black hole behaves like a rock,[⋆] then the phenomenon of black hole evaporation provides a remarkable conceptual window on quantum gravity. The signature of physics at the Planck scale is imprinted on the microscopic state of the Hawking

⋆ There is another logical possibility: perhaps quantum mechanics is saved, not because information is encoded in the Hawking radiation, but because the black hole never disappears completely; instead, it settles down to a stable little nugget of information. This is a serious proposal and deserves serious consideration. But I don't think it is the right idea, so I didn't take the time to discuss it in the talk.

radiation, because it is the physics of Planckian goo that determines the details of how the information is encoded in the radiation. This means that black hole evaporation is an exception to what we call the "decoupling principle." The decoupling principle says that physics at a large length scale is very insensitive to the details of physics at much shorter length scales. This principle is important; it is what makes it possible to do physics at all. If we needed to know all about physics at the Planck scale to understand the hydrogen atom, we would be in big trouble. The evaporating black hole is a very rare exception to this rule. Even for a huge black hole—a light year across—we can't have any idea how to read the message in the Hawking radiation unless we understand something about the microscopic properties of Planckian goo.

The remarkable formula

$$\text{Entropy} = \frac{1}{4}\frac{\text{Area}}{L^2_{\text{Planck}}}$$

is a miraculous message from the Planck scale, a powerful hint that can help to guide us toward the right theory of quantum gravity. The entropy presumably counts the number of microscopic degrees of freedom of the Planckian goo, and so tells us something highly nontrivial about Planck scale physics. It provides a stringent consistency test that the right theory of quantum gravity must pass. Soon we may know whether superstring theory can meet this challenge.