Chromatic Aberrations: Yang and Mills Meet Aharonov and Bohm

John Preskill

15 April 1993

This talk was originally scheduled for April 1st. I was delighted. I've always wanted to give a talk on April 1st, and this subject seemed like the ideal one for that date. Then, in a stunning reversal, the talk was rescheduled for April 15th. I was shocked. Suddenly, instead of speaking on the funniest day of the year, I was speaking on the least funny day of the year. I know that everyone is in a somber mood on April 15th, so I have decided that there will be no jokes in this talk. I'm sorry. I hope that you don't find the talk to be overly taxing.

So what is this talk about, anyway. It is about some interesting consequences of non-abelian gauge symmetry: charge with no localized source, and non-abelian generalizations of quantum statistics in two spatial dimensions. What these two things have in common is that they are both consequences of a phenomenon that lies at the heart of gauge theory—the Aharonov-Bohm effect. Since the Aharonov-Bohm effect is really a kind of geometrical effect, we should begin by recalling some concepts from the foundations of geometry.

Geometry and Curvature

The *geometrical* properties of a space are those intrinsic properties that have nothing to do with how we parametrize the space. Suppose, to be definite, that the space is a two dimensional surface. Suppose that a race of people live on this surface. These people know how to parallel transport a tangent vector on this surface—they can carry it from one point to another without rotating it. If one resident of the surface is a physicist skilled at instrumentation, she might devise a two-dimensional harmonic oscillator that is constrained to move tangent to the surface—if there is a gravitational field normal to the surface, an ordinary pendulum works very nicely. She starts the pendulum swinging in some arbitrarily chosen direction, then carries the pendulum along with her on her travels, being very careful at all times not to rotate it. When she arrives at her destination, the direction of the swinging pendulum defines the result of parallel transporting the original tangent vector to the new location.

But she will discover that the result of this procedure is path dependent. Suppose that the surface is a sphere, and that she and her twin live at a point on the equator of the sphere. They both start their pendula swinging in the same direction one day. Then physicist A travels directly from home to the north pole, along a line of longitude. Physicist B first saunters along the equator for a while, then turns north, and eventually joins her sister at the pole. Both have been very careful not to rotate their pendula, but when they are reunited at the pole, they find that their pendula are no longer swinging in the same direction. (The angle between the two tangent vectors is the solid angle on the sphere enclosed by their two paths.) What they have discovered is that the surface on which they reside has intrinsic *curvature*; by determining the extent to which parallel transport of a tangent vector is path dependent, they can measure the curvature.

An example of a curved surface that will be particularly relevant for us is the cone. (At this stage, I need a cone. Did anyone bring one? Oh, yeah. *Remove cone from head, revealing second cone underneath.*) The red line on the cone traces the excursion taken by our physicist, and the black arrows show the direction of the swinging pendulum. We can see that when she returns to her starting point, the pendulum that she left behind swings in a different direction than the one she took along with her, even though they were initially perfectly aligned. She has discovered that her cone is curved. Yet, the cone has the interesting property that it is flat almost everywhere. It doesn't look flat, but if we cut it along an arbitrarily selected path running from the edge to the tip (*cut cone with scissors*), we find that it can be smoothly flattened out on a table. Now it is easy to see that all of the black arrows point in the same direction.

All of the curvature of the cone is concentrated at a single point, the tip. Parallel transport around a closed path results in rotation (by an angle equal to the "deficit angle" of the flattened cone) if the path encloses the tip (once), but not otherwise. We see that a resident of the cone can detect the curvature even if she never visits the tip, and so never experiences it directly.

At the core of geometry is a principle of "orientational democracy" or "local symmetry." Each resident of a surface is free to choose her own conventions for, say, the x direction and the y direction, and to record her conventions by setting two pendula swinging in orthogonal directions at her home. In fact, one day she may tire of her old conventions and establish new ones, by resetting the pendula so that they now swing in different directions. This is a "local symmetry" transformation in the sense that residents at different locations are free to rotate their axes in different ways. But none of the intrinsic geometrical properties, the *observable* properties of the surface, depend on these choices.

The Aharonov-Bohm Effect

Now we know enough about geometry to understand the Aharonov-Bohm effect. The Aharonov-Bohm effect arises in the following situation: Imagine performing a double-slit interference experiment with electrons, and suppose that we place a solenoid carrying magnetic flux in between the two slits. This solenoid is perfectly shielded, so that no electron can penetrate inside and detect the magnetic field directly. Yet we find that the electrons know that the field is there. As the magnetic flux in the solenoid changes, the interference fringes shift. From the amount of the shift, we can infer that there is a field-dependent contribution to the relative phase of electron paths that pass through the top slit and the bottom slit given by

$$e^{ie\Phi/\hbar c} = \exp\left[\frac{ie}{\hbar c}\oint \vec{A}\cdot d\vec{x}\right] \ .$$

This is called the Aharonov-Bohm phase.

What is going on here? The interpretation is that the vector potential $\vec{A}$ is a *connection* that determines a notion of parallel transport. But in this case, what is being transported is not a tangent vector, but rather the phase of the wave function of the electron. The vector potential tells us what happens to an electron if we move it from a point to a neighboring point without rotating its phase. The principle at the core of electrodynamics is a principle of "phase democracy," also called "local symmetry" or "gauge symmetry." According to this principle, everyone is free to choose whatever convention she pleases to define the phase of the electron wave function at each point in space (and at any time). These conventions have no observable consequences, and the "gauge transformations" that modify these conventions have no effect on any observable quantities.

What is observable is the path dependence of this notion of parallel transport. We can carry an electron from $\vec{x}$ to $\vec{y}$, being very careful not to rotate the phase of the electron along the way. But if two electrons that have the same phase at $\vec{x}$ are carried from $\vec{x}$ to $\vec{y}$ along two different paths, they will in general arrive at $\vec{y}$ with different phases. The relative phase of the two electrons is the Aharonov-Bohm phase. It is an observable geometrical, or "gauge invariant," property. We see that the magnetic field can be interpreted as the *curvature* associated with the notion of parallel transport defined by the vector potential.

Note that the Aharonov-Bohm effect is to electrodynamics just as the cone is to Riemannian geometry. As a resident of a cone can infer the existence of the curvature at the tip without ever visiting the tip directly, the electron that propagates in a field-free region can know about the nonvanishing magnetic field inside a perfectly shielded solenoid, even though it never experiences the field directly.

Experiments like this one have been performed since the early 60's. But really good, well-controlled experiments, convincing even to skeptics, were not done until the 80's, thanks in part to advances in microlithography technology. Also in the 80's, a new type of Aharonov-Bohm interference experiement became possible that detected Aharonov-Bohm interference not in the vacumm, but inside matter (in the solid state). This is a gold ring, about a micron across, fabricated by Richard Webb and collaborators at IBM in the mid 80's. The ring has two leads attached to it. They cooled this ring down, placed it in a strong magnetic field, and observed the dependence of the resistance in the circuit on the applied field. An electron traveling from one lead to the other can travel along either of two paths. Whether these two paths interfere constructively or destructively depends on the magnetic flux enlosed by the two paths, so one expects to see oscillations in the resistance with a characteristic period. Indeed, the spacing between successive peaks in the resistance corresponds to a change in the enclosed magnetic flux given by

$$\Delta\Phi = \frac{2\pi\hbar c}{e} \ ,$$

such that the Aharonov-Bohm phase advances once around the unit circle.

What is truly remarkable about this measurement it that the electrons scatter many times off of impurities in the wire as they diffuse through it, yet the scattering does not destroy the phase coherence. The reason is that the scattering at sufficiently low temperature is almost always elastic, and only inelastic scattering destroys coherence. Also notice that, aside from the characteristic Aharonov-Bohm oscillations in the resistance, there are random fluctuations with a smaller frequency. This random noise is extremely interesting. It is associated with the change in the magnetic flux that actually penetrates the wire, rather than the flux enclosed by it. But I don't have time to talk about this here.

Non-abelian Gauge Symmetry

Instead, let's generalize our geometric picture to the case of a non-abelian local symmetry. We'll consider the case of quantum chromodynamics (QCD), the theory of the strong interactions.

Quarks come in three colors, which I'll call red (R), green (G), and blue (B), because those happen to be the colors of the transparency pens that I have. At the heart of QCD lies a principle of "color democracy"—everyone is free to orient her axes in color space however she pleases at each point in space (and at each time). No observable quantity can depend on these color conventions.

So suppose you go out for a walk, quark-watching, and you see a quark. You say, "Oh, look at that beautiful blue quark!" Except it doesn't mean anything. Someone else might see the very same quark and say, "Oh, look at that beautiful *red* quark!"

We can't tolerate this confusion, so we try to do something about it. What we do is establish a Quark Bureau of Standards (QBS) right here in Pasadena. If two quarks are in the same place at the same time we can at least tell whether they have the *same* color or not. So we select three mutually orthogonal directions in color space at the QBS, which we decide to call the R, G, and B directions. We select samples of quarks with each color, and seal them hermetically in bottles. Great care is taken to ensure that nothing rotates the color of these standard reference quarks.

Now if a wild quark is seen out in the field, we know what to do. We capture it, and carry it back to the QBS, being very careful not to rotate its color along the way. Upon arrival at the QBS, we can compare the wild quark to the standard quarks, and unambiguously identify its color.

Except, there is a problem that arises, because of the path dependence of parallel transport. To dramatize the problem, let us suppose (stretching our imaginations only a little) that a revolution comes, and Tommy Lasorda becomes dictator. As his first act in office, he makes a proclamation: "From now on, all quarks shall be Dodger blue!" So Tommy equips each Dodger with a hermetically sealed sample of a blue quark that has been callibrated at the QBS. Every player goes on a quark-hunting expedition, carrying his standard blue quark with him (and being very careful not to rotate its color). Every time he spots a wild quark, he rotates the color of the wild quark so that it lines up perfectly with his standard blue quark.

If color symmetry were a global symmetry, this plan to impose color uniformity might succeed. But because the symmetry is actually local, the forces of truth, justice and color democracy have it within their power to foil the plan. Suppose it's Marge Schott—she doesn't think quarks should be Dodger blue, she thinks they should be Cincinnati red. So she knows what to do. She builds a solenoid, and turns on a carefully selected current of colored quarks in the wire. The flux in the solenoid can be tuned just right so that when one of the Dodgers starts out at the QBS with his blue quark, voyages once around the solenoid, and then returns, he finds that the quark in his bottle has now turned red! Tommy's plan has failed.

The essence of non-abelian gauge symmetry, then, is a notion of parallel transport for a colored object, and the physical content of this notion is encoded in the path dependence of parallel transport. Suppose a quark is initially at a point x_0, and then we carry that quark around a closed path C that returns to x_0, being careful to preserve its color all along. In general, we find that when the quark returns to where it started, its color has been rotated by some element of the "local symmetry" or "gauge" group, $SU(3)$ in the case of QCD. This group element $U(x_0, C)$ can be said to characterize the "flux" enclosed by the path.

But now we encounter a puzzle. A blue quark and a red quark have different long-range gluon fields. The difference has a well-defined physical meaning. We can measure the force exerted by the gluon field on standard R, G, and B quarks. There will be no problem in imposing a uniform color convention on a large surface, as long as any color magnetic fields fall off rapidly with distance (no color magnetic monopoles). Suppose we bend Marge Schott's solenoid into a closed loop, with the color magnetic flux sealed inside. Now we take our blue quark, carry it once through the loop, and bring it home. It is now a red quark. But is doesn't make any sense, is not consistent with causality, for the long range color field measured on a distant surface to be changed by a well localized process in which a quark winds around a solenoid. So it seems that, after the winding, there is some color hiding somewhere, that compensates for the exchange of a blue quark for a red one. Where did the missing color go?

Part of what makes this puzzle tricky to think about is the existence of massless colored gluons in QCD. (Of course, it has been implicit in all of the discussion so far that our measurements are carried out on distance scales that are small compared to the scale of color confinement in the theory.) When we move quarks around, we can't prevent emission of very soft gluons that can carry color away. It will be easier to think about the puzzle of the disappearing color if we consider a simpler model that retains some of the essential features of QCD, but does not contain any massless charged particles. To find a model with these features, we may imagine that all or most of the gluons have acquired mass via the Higgs phenomenon. But that means that we need to understand how the Higgs phenomenon works.

The Higgs Phenomenon

The prototype of the Higgs phenomenon is superconductivity. The ground state of a superconductor can be viewed as a "pair condensate" a coherent state containing an indefinite number of Cooper pairs, each carrying charge $2e$, where e is the electron charge. It is instructive to consider what happens if we open up a hole in the superconductor, and thrust a solenoid into the hole. If the flux in the solenoid takes a generic value, then a Cooper pair that voyages around the hole acquires a nontrivial Aharonov-Bohm phase, given by

$$\exp\left[i\frac{(2e)}{\hbar c}\Phi\right] .$$

The pairs don't like that—the phase raises their energy. But the pairs have it within their power to do something about that undesirable phase. A supercurrent can flow around the solenoid, generating an additional magnetic field that augments the applied field inside the solenoid. The current can assume a value so that the *total* flux due to the applied field and the current together is such that pairs deep inside the superconductor experience no Aharonov-Bohm phase. The total flux must then be an integer multiple of a fundamental quantum of flux

$$\Phi_0 = \frac{2\pi\hbar c}{2e} = 2\times 10^{-7}\ \mathrm{flux-cm}^2 .$$

This is the phenomenon of flux quantization in superconductors.

In the case of a type-II superconductor, a sufficiently powerful magnetic field will penetrate through the superconductor. But it is energetically favorable for the field to break up into narrow strings of flux, each carrying the flux quantum Φ_0. Thus the field doesn't disturb the pairs in the bulk, far from the nearest string.

We see that in a superconductor, electrodynamics becomes a short-range interaction. The magnetic field outside a string decays like e^{-r/r_0}, where r_0 is the characteristic "penetration depth" of the superconductor. We may say that the "photon" has acquired a mass given by

$$m_\gamma = \frac{\hbar}{c}r_0^{-1} .$$

This is the Higgs mechanism, in the case of an abelian gauge theory.

In a non-abelian gauge theory, the ground state (or vacuum) may be a condensate containing an indefinite number of *colored* particles. Now suppose that a non-abelian magnetic field turns on in this vacuum. We can distinguish two types of magnetic field. The first type does not rotate the color of the condensate at all. This type of magnetic field is free to spread out, without raising the vacuum energy. The associated gauge interactions are long range, and the gauge fields of this type remain massless.

The second type of magnetic field *does* rotate the condensate. This type of magnetic flux will collapse to strings that carry a quantized flux (as in a superconductor), so that a condensate particle that winds around the string does not change its color. The associated gauge interactions are short range, and the gauge fields of this type acquire masses. This is the non-abelian Higgs mechanism.

The Alice String

Now we'll consider an instructive example of this non-abelian Higgs mechanism. For definiteness and ease of visualization, let's suppose that the underlying local symmetry group is $S0(3)$, the three-dimensional rotation group (instead of color $SU(3)$, which is a bit harder to think about). And let's imagine that the condensate is a "spin-two" object—we can think of it as an arrow pointing in a particular direction in the three-dimensional (internal) space on which $S0(3)$ acts, except that an arrow pointing up is identified with an arrow pointing down.

We see that there are two types of magnetic flux. If we call the direction in which the condensate points the $\hat{z}$ direction, then rotations about the $\hat{z}$ axis leave the condensate unchanged. The associated gauge field is a massless "photon;" it couples to an electric charge operator Q, the generator of rotations about the $\hat{z}$ axis.

Rotations about the $\hat{x}$ or $\hat{y}$ axes, on the other hand, move the condensate. The associated guage field are heavy—the magnetic flux is confined to strings. An object that is transported around the string gets rotated by $180°$ about an axis in the $x-y$ plane, for such a rotation leaves the condensate unchanged.

Now notice something remarkable. If R is such a rotation, then

$$RQR^{-1} = -Q \ ;$$

if we flip the $\hat{z}$-axis, rotate by θ, and then flip the axis back, the result is the same as a rotation by $-\theta$. That means that, in this model, charge conjugation is a *local* symmetry.

If you spot a charged particle, it has no invariant meaning to say that the particle is an electron—someone else could just as well identify it as a positron.

We can, of course, establish a Charge Bureau of Standards, by arbitrarily calling one charged particle positive (+), and its antiparticle negative (-), and storing samples of each charge in hermetically sealed bottles. Then if we capture a wild charge and carry it back to the CBS, we can determine its charge—it is either attracted or repelled by the standard positive charge.

But the outcome of this charge measurement is path dependent. Suppose we encounter a string on our way back to the CBS. If we pass to the left of the string, we find that the particle is $+$ when we reach the CBS. But if we pass to the right, the particle is identified as $-$ at the CBS. Because of this feature, this type of string was dubbed the "Alice" string by Albert Schwarz, who first discussed its properties about ten years ago—one who voyages around the string is reflected in the charge-conjugation looking glass.

The Alice string is to electrodynamics as the Möbius strip is to Riemannian geometry. We can establish a local convention to identify a left hand and a right hand on a Möbius strip, but these conventions have no *global* meaning, because a left hand that voyages around the strip becomes a right hand. Similarly, we can establish a charge convention at the CBS, but our convention can not be imposed globally, since a charge that voyages around an Alice string returns with its charge flipped in sign.

Disappearing Charge

In the context of the Alice string model, our puzzle of the disappearing color becomes a puzzle of disappearing charge, which is easier to think about. Using test charges that have been callibrated at the CBS, we can measure the electric field on a large surface, and use Gauss's law to determine the total charge enclosed by the surface. Now suppose that there is a closed loop of Alice string deep inside this surface. A charged particle, initially measured to be $+$, winds through the string loop and becomes $-$. Yet, this localized process cannot have changed the total electic charge as measured on the distant surface. Where did the missing charge go?

To resolve this puzzle, we need to consider in more detail how the electric field behaves as the charge moves around the string. For this purpose, it is convenient to adopt a particular convention for measuring the electric field. To measure the field, we measure the force on a positive test charge, and we can verify that the test charge is positive by carrying it to the CBS, and comparing with the standard + charge there. The trouble with this procedure, of course, is that the calibration of the test charge is path dependent—whether it is + or − depends on which way we carry it around the string.

To avoid this ambiguity, we can take a membrane, stretch it tightly across the string loop, and declare that no test charge is to cross the membrane when it is carried to the CBS for callibration. With this convention, the sign of the electric field at each point is unambiguously determined. But the electric field, measured this way, has an unusual property—it is discontinuous across the membrane. If our positive test charge were to pop through the membrane, it would become a negative test charge according to our conventions. Since the force on the test charge *does* behave smoothly across the membrane, the electric field that we measure must change sign at the membrane. This discontinuity is purely an artifact of our conventions; no physically observable quantity is discontinuous at the membrane (just as no geometrical quantity is really discontinous at the arbitrarily selected curve where we chose to cut open our cone).

We may think of the electric field as a two-valued function, and our convention picks out a single sheet of this two-valued function, with a branch cut (at the location of the membrane) joining it to the second sheet. Now look at what happens as a charged particle passes through the string loop. As the (+) charge approaches the loop, the electric field lines cannot penetrate the string, and they bend back. When the particle reaches the membrane, it "ducks under the cut" just as its image (-) charge on the second sheet pops out from under the cut. As the (-) charge pulls away from the loop, all of the flux emanating from it returns through the cut to the second sheet, while all of the flux emanating from the (+) charge (now on the second sheet) returns to the first sheet through the cut. Once the particle is far away, the cut appears to be a source of positive electric flux equal to twice the charge of the particle. The total electric charge has remained unchanged, and two units of charge have been transfered from the particle to the string loop.

We see that a loop of string can carry electric charge. But this electric charge is peculiar. There is no question about the reality of this charge—with suitably calibrated test charges, we can measure the electric flux through a large surface enclosing the loop. Yet if we venture inside the surface seeking the source of this electric field, no source can be found; the locally measured electric field has vanishing divergence everywhere. The branch cut that appears, with our conventions, to be the source, has no invariant significance. In keeping with the

Alice motif, and insofar as charge without a source is like a smile without a cat, this type of charge should be called "Cheshire charge."

When a right hand winds around a Möbius strip, it becomes a left hand, but we can't say exactly when it changed from left to right. Similarly, when a + charge winds through a loop of Alice string, it transfers two units of + charge to the string loop. We can't say exactly when this transfer of charge happens, but it definitely happens.

Cheshire Charge in the Laboratory

Cheshire charge is amusing, but does it really have anything to do with physics? Are there systems that can be studied in the laboratory and that exhibit these peculiar Alice properties?

In fact, similar phenomena *do* occur in nematic liquid crystals. These materials contain long, narrow, rod-like molecules. In the low temperature (nematic) phase, the rods tend to line up with one another. This phase flows like a liquid, but has long-range orientational order. The order parameter in this case is the "director" field, a vector that indicates how the rods are aligned. But the rods have no preferred direction, so a vector pointing up must be identified with a vector pointing down. Thus, the symmetry breaking pattern in a nematic is precisely the same as in the Alice model. The only difference is that the spontaneously broken symmetry, in the nematic, is a global symmetry rather than a gauge symmetry.

In the nematic, there is a line defect that could be described as a "global Alice string." The director field rotates by 180° on a closed path that encircles the core of the string, just like the condensate outside the core of an Alice string. In the Alice model, though I haven't mentioned them so far, there are also point defects, magnetic monopoles. In the nematic, there is a corresponding "global monopole" or "hedgehog" in which the director field points radially outward. A closed loop of global string can carry magnetic charge which is analogous to Cheshire charge. As a function of position along the string, the plane of the director field can twist in space. The magnetic charge carried by the loop turns out to be the number of times this plane twists around before the loop closes. This is easiest to visualize in the case of a loop with a single twist (shown here in cross section). On a surface containing this loop, the director points radially outward, so there is a magnetic charge inside the surface.

In the Alice model, a magnetic monopole that winds around a string loop transfers magnetic charge to the loop, just as an electrically charged particle transfers electric charge.

So it is in the nematic. The magnetic charge inside a surface can be inferred from the "winding number" of the director on that surface. A global monopole that passes through a loop of global Alice string becomes an antimonopole, with the opposite value for the winding number. The total magnetic charge, defined by the behavior of the director on a large surface that encloses the loop and monopole, is not changed by this process. What happens is that the passage of the monopole through the loop twists the string, so that the magnetic charge acquired by the loop compensates for the charge lost by the monopole. It sounds odd that a discrete quantity, the winding number, which takes an integer value, can be changed in a continuous process, the passage of the monopole number through the loop. But that is what Cheshire charge is all about. When the monopole is in the vicinity of the string loop, there is no unambiguous way of assigning a winding number to the string; we can resolve the ambiguity only by adopting a particular convention for measuring the winding. One cannot say exactly *when* the transfer of charge happens, but it definitely happens.

Static properties of defects in liquid crystals have been studied for a long time, but dynamical properties of defects in nematics have been studied only recently, in particular by Yurke and collaborators at AT&T, and Bowick and collaborators at Syracuse. They use a commercially available material called K15, which has a nematic-isotropic first-order phase transition at the convenient temperature $35°C$ (1 atmosphere pressure). One puts a drop of this glop on a microscope slide and heats it with a light bulb. Then one allows it to air cool. It passes through the critical temperature, and the transition to the nematic phase then proceeds via bubble nucleation. Bubble walls collide, producing defects. The evolution of the defects can then be viewed and recorded on videotape. Because of the similarity of this scenario with cosmological phases transitions that have been much discussed in connection with the formation of large scale structure in the universe, Yurke et al. called their paper "Cosmology in the Laboratory."

I have a brief sample tape that was provided by Mark Bowick. Unfortunately, I don't have a picture of a monopole-string interaction, but we can watch both twisted and untwisted string loops as they shrink to a point and annihilate. The untwisted loops can disappear completely, but the annihilation of a twisted loop leaves behind a pointlike defect, a monopole, that won't go away.

Exotic Quantum Statistics

Now I would like to turn to a different topic related to the Aharonov-Bohm effect, exotic generalizations of quantum statistics. In three spatial dimensions, indistinguishable particles are either bosons or fermions, but more general possibilities exist in two dimensions. As a concrete example, imagine a charge q bound to a flux Φ. If two such objects are interchanged, there is an Aharonov-Bohm interaction between the charge of each object and the flux of the other object. As a result, under the interchange, the two-body wave function acquires the Aharonov-Bohm phase

$$e^{iq\Phi} = e^{i\theta} \ .$$

Since there is no restriction on the allowed flux, this exchange phase θ can take any value. Thus, Frank Wilczek suggested the name "anyon" for such objects.

What is special about two spatial dimensions? To understand the difference between two and three dimensions, imagine performing two exchanges in succession. A double exhange is equivalent to winding one object around the other, and returning it to its original position. In three dimensions, carrying one particle around a closed path that encloses the other particle is really the same as doing nothing at all, for we can smoothly deform the path to a trivial path in which the particle doesn't go anywhere. Thus, a double exchange leaves the two-body wave function invariant, and the exchange phase is restricted to take the values 1 and -1. But in two dimensions, carrying one particle around another is not the same as doing nothing at all—the path has a winding number with an invariant topological meaning. A history in which one particle winds around another cannot be smoothly deformed to a history in which no winding occurs, unless the particles meet at some point in spacetime. So the phase $e^{2i\theta}$ acquired by the two body wave function under the double exchange need not be trivial.

Nor do anyons exhaust the possibilities for unusual quantum statistics in two dimensions; other possibilities can arise in non-abelian gauge theories. Before, we considered what happens when a colored particle is transported around a non-abelian flux. Now let's ask what happens when two fluxes are interchanged.

As we've seen, a flux can be labeled by an element of the local symmetry group. We can establish a Color Bureau of Standards, and when our standardized colored objects are carried around an isolated flux and returned to the CBS, the color has been rotated by a transformation U. For the purpose of describing the exchange of two fluxes, it is convenient to adopt a particular convention for measuring color on the background of a flux. The color of an object can be identified by carrying it back to the Color Bureau of Standards, where it can be compared to the standard colored objects that are stored there. To avoid any ambiguities, as in our discussion of Alice strings, we choose a "cut" terminating on the flux, and agree that no colored object will be allowed to cross the cut when it is brought back to the CBS. A way to characterize the flux, then, is to say that the color of an object, measured using this convention, jumps discontinuously when the object crosses the cut. We have

$$|\psi\rangle_{\text{below cut}} = U|\psi\rangle_{\text{above cut}} \ ,$$

where $|\psi\rangle$ is the wave function of a colored object, and U is the element of the local symmetry group that is associated with the flux. Of course, the cut is purely an artifact of our conventions, and has no physical significance of its own.

Now if we exchange two fluxes labeled by group elements U and V, we must drag their cuts along during the exchange. We see that after the exchange (shown here) the effect of carrying a colored object around the flux that was originally labeled by the group element U is still the color rotation U. But a path that winds around the flux that was originally labeled by V must cross the U cut both before and after crossing the V cut; thus the effect is the color rotation UVU^{-1}. After the exchange, the two flux state is modified according to

$$|U,V\rangle \to |UVU^{-1},U\rangle \ ;$$

the fluxes differ from the original values if U and V do not commute.

As a concrete example, suppose that the color $SU(3)$ group of QCD is spontaneously broken down to the subgroup S_3, the group that permutes the three colors R, G, and B. In this Higgs phase, flux is confined to strings, such that the effect of voyaging around the string is a transformation in this unbroken group, a transformation that preserves the value of the Higgs condensate. In two spatial dimensions, these strings are pointlike particles, which we will call "vortices."

Let's consider the three vortices whose flux corresponds to a transposition of two colors: (RG)—which we'll call vortex A, (GB)—vortex B, and (BR)—vortex C. We can establish a Vortex Bureau of Standards, where standard samples are preserved of vortices A, B, and C. A wild vortex in the field can be carried to the VBS, where its flux can be identified. Of course, there is the usual ambiguity: if many vortices are present, the outcome of the flux measurement depends on just how we weave the vortex to be measured through the others on its way to the VBS.

Now suppose we are interested in the amplitude for a process in which two vortices propagate from specified positions at an initial time to specified positions at a final time, and suppose that the initial vortices are A and B. We may compute the amplitude by summing over all possible two-vortex trajectories. In the contribution to the amplitude due to histories in which no exchange takes place, the final vortices are also A and B. If a single exchange takes place, then, according to the rule found above, the final vortices are C and A. If there is a double exchange, the final vortices are B and C, and if there is a triple exchange, the final vortices are A and B. We see that the single and double exchange processes do not interfere with the no-exchange process, because the final quantum numbers of the vortex pair is different in these cases. The triple exchange process *does* add coherently to the no-exchange process in the amplitude for AB to go to AB. But notice that in a triple exchange, the two vortices have changed places—the A vortex has become a B vortex and the B vortex has become an A vortex.

The existence of an exchange contribution to the amplitude is the hallmark of identical particles in quantum mechanics; when we say that two particles are indistinguishable, we mean that it is not always possible to keep track of "who's who." But our non-abelian vortices A and B are indistinguishable particles with an unusual feature—they have different quantum numbers! It seems that we can take vortex A or vortex B to the VBS, and verify that they are different types of objects. They *are* different, but they are also indistinguishable. Distinct objects that are indistinguishable are the essence of non-abelian statistics.

Non-abelian Statistics in the Laboratory?

That's interesting. But do objects like this, obeying non-abelian quantum statistics, really exist in nature?

There is no (firm) evidence that they have ever been seen. But life is more exciting when you take risks, so I will make the reckless prediction that evidence for non-abelian indistinguishable particles will eventually be found in condensed matter systems. Why would I make such a rash statement? It is because our experience with condensed matter physics has shown that that a frustrated strongly-correlated electron system will go to *extraordinary* lengths to relieve its frustration.

The most spectacular realization of this principle is the fractional quantum Hall effect (FQHE). The FQHE is one of the most amazing phenomena ever discovered in condensed matter, comparable in its conceptual implications, I believe, to the discovery of superconductivity or superfluidity. The effect arises when a two-dimensional electron gas, confined to the interface between two semiconductors, is cooled and placed in a strong magnetic field normal to the interface. What two-dimensional electrons in a magnetic field love above all else is to fill a Landau level. In the filled Landau level, one electron executing quantized cyclotron motion sits atop each quantum of magnetic flux. A felicitous property of this state is that it is incompressible—squeezing it an infinitesimal amount costs a finite amount of energy, because an electron must get bumped up to the next Landau level.

But suppose, for a given value of the magnetic field, that the density of electrons is too small to fill a Landau level; specifically, suppose the number of electrons per flux quantum is

$$\nu \equiv \frac{\text{electrons}}{\text{flux quanta}} = \frac{1}{2m+1} ,$$

where m is an integer. The electrons are not very happy, but it is within their power to improve the situation. If the magnetic field is large enough, the temperature is low enough, and the mobility of the sample is high enough (small density of impurities), the interactions among the electrons drive the formation of a remarkable collective state. Very loosely speaking, each electron forms a bound state with $2m$ flux quanta, so that the abundance of these charged composites is just right to fill a Landau level in the remaining, unbound, magnetic field. When two of these electron-flux composites are exchanged, there is an additional Aharonov-Bohm contribution to the exchange phase; roughly speaking the exchange phase becomes $e^{i\pi(2m+1)}$—the phase winds m times around the unit circle before advancing to -1. These bound "superfermions" can dissociate into $2m+1$ constituents, each carrying electric charge $e/(2m+1)$, and with exchange phase $e^{i\theta}$, where $\theta = \pi/(2m+1)$. These fractionally charged, anyonic quasi-particles are responsible for carrying the fractional Hall current.

Another way to say what has happened is that the electrons have manufactured a fictitious "statistical" magnetic flux that partially cancels the applied flux, so that the electrons can fill a Landau level with respect to the combined total flux. Any way you look at it, what happens is remarkable. From their short range mutual interactions, the electrons manage to create an Aharonov-Bohm "gauge interaction" between distantly separated quasi-particles. It is the incompressibility of the resulting collective state that makes this trick possible.

This trick works only for special "magic" values of the filling factor. My speculation is that, when the filling factor is unfavorable, this electron system, having exhausted the anyonic tricks at its disposal, may turn to the option of manufacturing a fictitious *non-abelian* magnetic flux, in order to establish a felicitous collective state. Quasi-particle excitations in this state could obey non-abelian statistics. A particulary promising value of the filling factor at which to look for such exotica would be $\nu = \frac{1}{2}$. Similar tricks might turn up in other types of frustrated quantum many-body systems, for example in frustrated anti-ferromagnets.

An important question is, if a condensed matter system were produced that supports non-abelian quasi-particles, how would we know? What qualitative features of, say, the bulk transport properties would signal that non-abelian statistics had been discovered? I don't know—I understand little about the many-body physics of particles that obey non–abelian statistics. Perhaps an important clue is that two or more such objects can carry unlocalized charges analogous to Cheshire charge. The many-body physics remains an intriguing open question.

Some History

Before concluding, I would like to mention some aspects of the history of the theory of the Aharonov-Bohm effect.

The idea that the electrodynamic vector potential A_μ is a connection that determines a notion of parallel transport was actually first put forward in 1918, by the mathematician Hermann Weyl. General relativity had been proposed only a few years earlier, and Weyl felt, as Einstein did, that if there is a geometrical theory of gravity, there should also be a geometric theory of the other interaction that was known at that time, electromagnetism.

Weyl did not have exactly the right idea in 1918. The trouble was that he was too far ahead of his time. Quantum mechanics had not been invented yet, so he could hardly have anticipated that the relevant notion was transport of the phase of the electron wave function. But he made a clever suggestion. Weyl said that, just as the Riemannian connection is needed to tell us how to transport a tangent vector from one point to another without changing its *direction*, so another connection is also needed to tell us how to transport a vector from one point to another without changing its *length*. He proposed that A_μ is this new connection.

When Weyl's paper was published, it was followed by a comment by A. Einstein. Einstein remarked that Weyl's idea was very interesting, but it could not be right. According to Weyl, if a clock or a meter stick is carried around a closed path in a magnetic field, the length of the stick or the interval between ticks of the clock is rescaled by the factor

$$\exp\left[c \oint \vec{A} \cdot d\vec{x}\right] ,$$

where c is some constant. We can't do physics that way, said Einstein, for the way each clock kept time would depend not just on where it was, but also on where it had been.

The right interpretation, that the vector potential defines parallel transport of the *phase* of the wave function, was proposed not long after the emergence of quantum mechanics, by London and Fock, independently, in 1927. Weyl returned to the idea in 1929, and gave the first modern formulation of the notion of a gauge transformation and of gauge invariance.

The next important step was taken by Dirac, in his famous 1931 paper on magnetic monopoles in quantum theory. Dirac proposed that a magnetic monopole could be envisaged as a semi-infinitely long, infinitesimally thin string of magnetic flux. The end of the string, where the flux spills out, appears to be a magnetic charge. But for this picture to make sense, the string would have to be completely invisible. Dirac was familiar with Weyl's ideas, and quoted Weyl's 1929 paper. He pointed out that an electron would be able to detect the string unless the phase $e^{ie\Phi/\hbar c}$ acquired by an electron that circumnaviagates the string is trivial. The flux Φ in the string, and thus the charge of the monopole, must therefore be an integer multiple of the magnetic charge quantum

$$\Phi_0 = \frac{2\pi\hbar c}{e} .$$

Dirac was so fond of his idea that he concluded his paper by saying, "One would be surprised if Nature made no use of it." So Dirac was a pioneer of a style of doing theoretical particle physics that remains popular today.

Though Dirac came remarkably close, neither Weyl or Dirac explicitly described what later came to be known as the Aharonov-Bohm interference experiment, in which the change in the flux of a shielded solenoid causes a shift in the fringes of an electron interference pattern. The first explicit discussion that I know of was due to Ehrenberg and Siday in 1949. In a crucial passage (accompanied by a diagram) in the conclusion of their paper, they remarked, "One might therefore expect wave-optical phenomena to arise which are due to the presence of a magnetic field, but not due to the magnetic field itself, *i.e.*, which arise whilst the rays are in field-free regions only." Yet Ehrenberg and Siday did not make much of a fuss about this effect. (It is not mentioned in the abstract or introduction of the paper.)

Their paper seems not to have attracted much attention at the time, and it remains little known and rarely quoted even today.

(In 1950, London, familiar with the notion of parallel transport of a phase since 1927, first predicted the phenomenon of flux quantization in superconductors.)

The Aharonov-Bohm paper appeared in 1959. In spite of all the anticipations, their paper is justly hailed as a great classic. Much more clearly and comprehensively than previous authors, they stressed the special role of the electromagnetic potentials in quantum theory, and that non-local gauge invariant quantities can have observable effects. They also emphasized the experimental implications, and indeed, experiments confirming the effect were already done within a year after the appearance of their paper.

Meanwhile, in another important development, Yang and Mills invented non-abelian gauge theory in 1954. Curiously, they were completely ignorant, at the time, of Weyl's geometrical ideas. Yang had learned about gauge invariance from articles by Pauli, and Pauli had very deliberately stripped away all of the geometrical motivation, which he did not like. What had deeply impressed Yang was that local symmetry principles can powerfully constain the dynamics of a theory, as gauge invariance determines the form of the coupling of the electron to the photon in electrodynamics. It was this feature that he and Mills sought to generalize. Yang says that it was not until the late 60's that he recognized that Yang-Mills theory has a geometrical interpretation, and that the field strength is analogous to the Riemannian curvature. The geometrical viewpoint re-entered the physics literature in a 1975 paper by Wu and Yang, which introduced the language of fibre bundles to many physicists. Since the mid 70's, geometrical ideas have played a decisive role in the development of theoretical particle physics.

Conclusions

I have done work related to what I've discussed here with a variety of collaborators: Mark Alford, Martin Bucher,* Sidney Coleman, Lawrence Krauss, Kai-Ming Lee,* Hoi-Kwong Lo,* John March-Russell, Patrick McGraw,* Robert Navin,* and Frank Wilczek. Those marked with an asterisk (*) are current or former Caltech graduate students.

The main pedagogical goal of this talk has been to explain what non-abelian gauge invariance is all about. We have also seen that when Yang and Mills meet Aharonov and Bohm, novel phenomena arise. I have particularly emphasized the existence of "Cheshire" charge with no localized source, and of non-abelian quantum statistics in two spatial dimensions (whereby distinct objects must be regarded as indistinguishable).

Where is the *physics* in all this? There are a number of interesting implications that I have not had time to discuss. For example, it is amusing to think about the consequences if strings with Alice properties were produced in a phase transition in the early universe. By thinking about the Aharonov-Bohm interactions of black holes with strings, one gains insight into some of the quantum-mechanical properties of black holes. The non-abelian Aharonov-Bohm effect provides a powerful tool for classifying the different possible Higgs phases of a gauge theory.

But the most likely ways of making contact between the ideas I've discussed and natural phenomena come from condensed matter systems. I've described the analog of magnetic Cheshire charge carried by the topological defects associated with a spontaneously broken global symmetry. I didn't have time to describe an analog of electric Cheshire charge that can arise in systems with global symmetries as a a consequence of the Berry phase. The greatest potential for a truly deep connection with phenomenology comes from the possibility that strongly-correlated electron systems, or other frustrated quantum many-body systems, may contain quasi-particles that obey non-abelian statistics.

That was a pretty good idea that Hermann Weyl had in 1918. His principle, gauge invariance, turned out to be the key to understanding the electroweak and strong interactions, as well as electromagnetism. Why did this turn out to be so?

One can think of various explanations, but the truth is that we don't really know. This appears to be a good example of what Wigner called "the unreasonable effectiveness of mathematics in the physical sciences." At a fundamental level, geometry seems to govern physics. That conclusion, I think, would have been satisfying to Weyl, and to Einstein. They both died in 1955, and as far as I know, neither one of them ever learned anything about Yang-Mills theory. I like to believe that they would have both been pleased by Yang-Mills theory, and especially pleased by the role it has assumed in the description of nature.

Weyl's principle of gauge invariance, and its implications, are still not completely understood. I expect that the geometrical principles at the heart of gauge theory will continue to make contact with physical phenomena in surprising and highly enlightening ways.