Richard Healey

Statement

and

Readings

Some more recent ideas about the Aharonov-Bohm effect

Richard Healey
University of Arizona, Tucson, AZ, USA

Abstract

I shall critically review some interesting ideas about how to understand the Aharonov–Bohm effect physicists and philosophers have put forward since the acknowledgment that Tonomura's experiments provide convincing if not conclusive evidence of a magnetic A-B effect in regions in which the electromagnetic-field is zero. Some of these are *very* recent. The review will focus on three issues.

- What is the appropriate theoretical framework in which to understand the effect?
- What concepts of locality are threatened, and how should that threat be addressed?
- What physical objects and properties are responsible for the A-B effect?

THE
PHYSICAL REVIEW

A journal of experimental and theoretical physics established by E. L. Nichols in 1893

SECOND SERIES, VOL. 115, NO. 3 AUGUST 1, 1959

Significance of Electromagnetic Potentials in the Quantum Theory

Y. AHARONOV AND D. BOHM
H. H. Wills Physics Laboratory, University of Bristol, Bristol, England
(Received May 28, 1959; revised manuscript received June 16, 1959)

In this paper, we discuss some interesting properties of the electromagnetic potentials in the quantum domain. We shall show that, contrary to the conclusions of classical mechanics, there exist effects of potentials on charged particles, even in the region where all the fields (and therefore the forces on the particles) vanish. We shall then discuss possible experiments to test these conclusions; and, finally, we shall suggest further possible developments in the interpretation of the potentials.

1. INTRODUCTION

IN classical electrodynamics, the vector and scalar potentials were first introduced as a convenient mathematical aid for calculating the fields. It is true that in order to obtain a classical canonical formalism, the potentials are needed. Nevertheless, the fundamental equations of motion can always be expressed directly in terms of the fields alone.

In the quantum mechanics, however, the canonical formalism is necessary, and as a result, the potentials cannot be eliminated from the basic equations. Nevertheless, these equations, as well as the physical quantities, are all gauge invariant; so that it may seem that even in quantum mechanics, the potentials themselves have no independent significance.

In this paper, we shall show that the above conclusions are not correct and that a further interpretation of the potentials is needed in the quantum mechanics.

2. POSSIBLE EXPERIMENTS DEMONSTRATING THE ROLE OF POTENTIALS IN THE QUANTUM THEORY

In this section, we shall discuss several possible experiments which demonstrate the significance of potentials in the quantum theory. We shall begin with a simple example.

Suppose we have a charged particle inside a "Faraday cage" connected to an external generator which causes the potential on the cage to alternate in time. This will add to the Hamiltonian of the particle a term $V(x,t)$ which is, for the region inside the cage, a function of time only. In the nonrelativistic limit (and we shall assume this almost everywhere in the following discussions) we have, for the region inside the cage, $H=H_0+V(t)$ where H_0 is the Hamiltonian when the generator is not functioning, and $V(t)=e\phi(t)$. If $\psi_0(x,t)$ is a solution of the Hamiltonian H_0, then the solution for H will be

$$\psi=\psi_0 e^{-iS/\hbar}, \quad S=\int V(t)dt,$$

which follows from

$$i\hbar\frac{\partial\psi}{\partial t}=\left(i\hbar\frac{\partial\psi_0}{\partial t}+\psi_0\frac{\partial S}{\partial t}\right)e^{-iS/\hbar}=[H_0+V(t)]\psi=H\psi.$$

The new solution differs from the old one just by a phase factor and this corresponds, of course, to no change in any physical result.

Now consider a more complex experiment in which a single coherent electron beam is split into two parts and each part is then allowed to enter a long cylindrical metal tube, as shown in Fig. 1.

After the beams pass through the tubes, they are combined to interfere coherently at F. By means of time-determining electrical "shutters" the beam is chopped into wave packets that are long compared with the wavelength λ, but short compared with the length of the tubes. The potential in each tube is determined by a time delay mechanism in such a way that the potential is zero in region I (until each packet is well inside its tube). The potential then grows as a function of time, but differently in each tube. Finally, it falls back to zero, before the electron comes near the

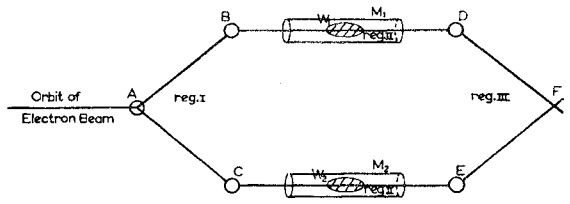


FIG. 1. Schematic experiment to demonstrate interference with time-dependent scalar potential. A, B, C, D, E: suitable devices to separate and divert beams. W_1, W_2: wave packets. M_1, M_2: cylindrical metal tubes. F: interference region.

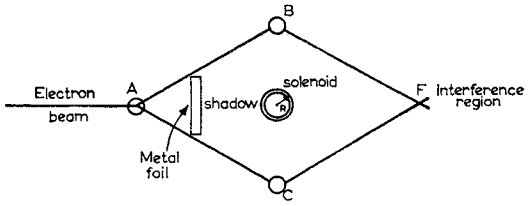


FIG. 2. Schematic experiment to demonstrate interference with time-independent vector potential.

other edge of the tube. Thus the potential is nonzero only while the electrons are well inside the tube (region II). When the electron is in region III, there is again no potential. The purpose of this arrangement is to ensure that the electron is in a time-varying potential without ever being in a field (because the field does not penetrate far from the edges of the tubes, and is nonzero only at times when the electron is far from these edges).

Now let $\psi(x,t)=\psi_1^0(x,t)+\psi_2^0(x,t)$ be the wave function when the potential is absent (ψ_1^0 and ψ_2^0 representing the parts that pass through tubes 1 and 2, respectively). But since V is a function only of t wherever ψ is appreciable, the problem for each tube is essentially the same as that of the Faraday cage. The solution is then

$$\psi=\psi_1^0 e^{-iS_1/\hbar}+\psi_2^0 e^{-iS_2/\hbar},$$

where

$$S_1=e\int\varphi_1 dt, \quad S_2=e\int\varphi_2 dt.$$

It is evident that the interference of the two parts at F will depend on the phase difference $(S_1-S_2)/\hbar$. Thus, there is a physical effect of the potentials even though no force is ever actually exerted on the electron. The effect is evidently essentially quantum-mechanical in nature because it comes in the phenomenon of interference. We are therefore not surprised that it does not appear in classical mechanics.

From relativistic considerations, it is easily seen that the covariance of the above conclusion demands that there should be similar results involving the vector potential, $\mathbf{A}$.

The phase difference, $(S_1-S_2)/\hbar$, can also be expressed as the integral $(e/\hbar)\oint\varphi dt$ around a closed circuit in space-time, where φ is evaluated at the place of the center of the wave packet. The relativistic generalization of the above integral is

$$\frac{e}{\hbar}\oint\left(\varphi dt-\frac{\mathbf{A}}{c}\cdot d\mathbf{x}\right),$$

where the path of integration now goes over any closed circuit in space-time.

As another special case, let us now consider a path in space only (t=constant). The above argument suggests that the associated phase shift of the electron wave function ought to be

$$\Delta S/\hbar=-\frac{e}{c\hbar}\oint\mathbf{A}\cdot d\mathbf{x},$$

where $\oint\mathbf{A}\cdot d\mathbf{x}=\int\mathbf{H}\cdot d\mathbf{s}=\phi$ (the total magnetic flux inside the circuit).

This corresponds to another experimental situation. By means of a current flowing through a very closely wound cylindrical solenoid of radius R, center at the origin and axis in the z direction, we create a magnetic field, $\mathbf{H}$, which is essentially confined within the solenoid. However, the vector potential, $\mathbf{A}$, evidently, cannot be zero everywhere outside the solenoid, because the total flux through every circuit containing the origin is equal to a constant

$$\phi_0=\int\mathbf{H}\cdot d\mathbf{s}=\int\mathbf{A}\cdot d\mathbf{x}.$$

To demonstrate the effects of the total flux, we begin, as before, with a coherent beam of electrons. (But now there is no need to make wave packets.) The beam is split into two parts, each going on opposite sides of the solenoid, but avoiding it. (The solenoid can be shielded from the electron beam by a thin plate which casts a shadow.) As in the former example, the beams are brought together at F (Fig. 2).

The Hamiltonian for this case is

$$H=\frac{[\mathbf{P}-(e/c)\mathbf{A}]^2}{2m}.$$

In singly connected regions, where $\mathbf{H}=\nabla\times\mathbf{A}=0$, we can always obtain a solution for the above Hamiltonian by taking $\psi=\psi_0 e^{-iS/\hbar}$, where ψ_0 is the solution when $\mathbf{A}=0$ and where $\nabla S/\hbar=(e/c)\mathbf{A}$. But, in the experiment discussed above, in which we have a multiply connected region (the region outside the solenoid), $\psi_0 e^{-iS/\hbar}$ is a non-single-valued function[1] and therefore, in general, not a permissible solution of Schrödinger's equation. Nevertheless, in our problem it is still possible to use such solutions because the wave function splits into two parts $\psi=\psi_1+\psi_2$, where ψ_1 represents the beam on

[1] Unless $\phi_0=nhc/e$, where n is an integer.

one side of the solenoid and ψ_2 the beam on the opposite side. Each of these beams stays in a simply connected region. We therefore can write

$$\psi_1=\psi_1^0 e^{-iS_1/\hbar}, \quad \psi_2=\psi_2^0 e^{-iS_2/\hbar},$$

where S_1 and S_2 are equal to $(e/c)\int \mathbf{A}\cdot d\mathbf{x}$ along the paths of the first and second beams, respectively. (In Sec. 4, an exact solution for this Hamiltonian will be given, and it will confirm the above results.)

The interference between the two beams will evidently depend on the phase difference,

$$(S_1-S_2)/\hbar=(e/\hbar c)\int \mathbf{A}\cdot d\mathbf{x}=(e/\hbar c)\phi_0.$$

This effect will exist, even though there are no magnetic forces acting in the places where the electron beam passes.

In order to avoid fully any possible question of contact of the electron with the magnetic field we note that our result would not be changed if we surrounded the solenoid by a potential barrier that reflects the electrons perfectly. (This, too, is confirmed in Sec. 4.)

It is easy to devise hypothetical experiments in which the vector potential may influence not only the interference pattern but also the momentum. To see this, consider a periodic array of solenoids, each of which is shielded from direct contact with the beam by a small plate. This will be essentially a grating. Consider first the diffraction pattern without the magnetic field, which will have a discrete set of directions of strong constructive interference. The effect of the vector potential will be to produce a shift of the relative phase of the wave function in different elements of the gratings. A corresponding shift will take place in the directions, and therefore the momentum of the diffracted beam.

3. A PRACTICABLE EXPERIMENT TO TEST FOR THE EFFECTS OF A POTENTIAL WHERE THERE ARE NO FIELDS

As yet no direct experiments have been carried out which confirm the effect of potentials where there is no field. It would be interesting therefore to test whether such effects actually exist. Such a test is, in fact, within the range of present possibilities.[2] Recent experiments[3,4] have succeeded in obtaining interference from electron beams that have been separated in one case by as much as 0.8 mm.[3] It is quite possible to wind solenoids which are smaller than this, and therefore to place them between the separate beams. Alternatively, we may obtain localized lines of flux of the right magnitude (the magnitude has to be of the order of $\phi_0=2\pi c\hbar/e\sim 4\times 10^{-7}$ gauss cm^2) by means of fine permanently magnetized "whiskers".[5] The solenoid can be used in Marton's device,[3] while the whisker is suitable for another experimental setup[4] where the separation is of the order of microns and the whiskers are even smaller than this.

In principle, we could do the experiment by observing the interference pattern with and without the magnetic flux. But since the main effect of the flux is only to displace the line pattern without changing the interval structure, this would not be a convenient experiment to do. Instead, it would be easier to vary the magnetic flux within the same exposure for the detection of the interference patterns. Such a variation would, according to our previous discussion, alter the sharpness and the general form of the interference bands. This alteration would then constitute a verification of the predicted phenomena.

When the magnetic flux is altered, there will, of course, be an induced electric field outside the solenoid, but the effects of this field can be made negligible. For example, suppose the magnetic flux were suddenly altered in the middle of an exposure. The electric field would then exist only for a very short time, so that only a small part of the beam would be affected by it.

4. EXACT SOLUTION FOR SCATTERING PROBLEMS

We shall now obtain an exact solution for the problem of the scattering of an electron beam by a magnetic field in the limit where the magnetic field region tends to a zero radius, while the total flux remains fixed. This corresponds to the setup described in Sec. 2 and shown in Fig. 2. Only this time we do not split the plane wave into two parts. The wave equation outside the magnetic field region is, in cylindrical coordinates,

$$\left[\frac{\partial^2}{\partial r^2}+\frac{1}{r}\frac{\partial}{\partial r}+\frac{1}{r^2}\left(\frac{\partial}{\partial\theta}+i\alpha\right)^2+k^2\right]\psi=0, \qquad (1)$$

where $\mathbf{k}$ is the wave vector of the incident particle and $\alpha=-e\phi/ch$. We have again chosen the gauge in which $A_r=0$ and $A_\theta=\phi/2\pi r$.

The general solution of the above equation is

$$\psi=\sum_{m+-\infty}^{\infty} e^{im\theta}[a_m J_{m+\alpha}(kr)+b_m J_{-(m+\alpha)}(kr)], \qquad (2)$$

where a_m and b_m are arbitrary constants and $J_{m+\alpha}(kr)$ is a Bessel function, in general of fractional order (dependent on ϕ). The above solution holds only for $r>R$. For $r<R$ (inside the magnetic field) the solution has been worked out.[6] By matching the solutions at $r=R$ it is easily shown that only Bessel functions of positive order will remain, when R approaches zero.

[2] Dr. Chambers is now making a preliminary experimental study of this question at Bristol.

[3] L. Marton, Phys. Rev. **85**, 1057 (1952); **90**, 490 (1953). Marton, Simpson, and Suddeth, Rev. Sci. Instr. **25**, 1099 (1954).

[4] G. Mollenstedt, Naturwissenschaften **42**, 41 (1955); G. Mollenstedt and H. Düker, Z. Physik **145**, 377 (1956).

[5] See, for example, Sidney S. Brenner, Acta Met. 4, 62 (1956).

[6] L. Page, Phys. Rev. **36**, 444 (1930).

This means that the probability of finding the particle inside the magnetic field region approaches zero with R. It follows that the wave function would not be changed if the electron were kept away from the field by a barrier whose radius also went to zero with R.

The general solution in the limit of R tending to zero is therefore

$$\psi=\sum_{m=-\infty}^{\infty} a_m J_{|m+\alpha|}e^{im\theta}. \tag{3}$$

We must then choose a_m so that ψ represents a beam of electrons that is incident from the right ($\theta=0$). It is important, however, to satisfy the initial condition that the current density,

$$\mathbf{j}=\frac{\hbar(\psi^*\nabla\psi-\psi\nabla\psi^*)}{2im}-\frac{e}{mc}\mathbf{A}\psi^*\psi, \tag{4}$$

shall be constant and in the x direction. In the gauge that we are using, we easily see that the correct incident wave is $\psi_{\text{inc}}=e^{-ikx}e^{-i\alpha\theta}$. Of course, this wave function holds only to the right of the origin, so that no problem of multiple-valuedness arises.

We shall show in the course of this calculation that the above conditions will be satisfied by choosing $a_m=(-i)^{|m+\alpha|}$, in which case, we shall have

$$\psi=\sum_{m=-\infty}^{\infty}(-i)^{|m+\alpha|}J_{|m+\alpha|}e^{im\theta}.$$

It is convenient to split ψ into the following three parts: $\psi=\psi_1+\psi_2+\psi_3$, where

$$\begin{aligned}\psi_1&=\sum_{m=1}^{\infty}(-i)^{m+\alpha}J_{m+\alpha}e^{im\theta},\\ \psi_2&=\sum_{m=-\infty}^{-1}(-i)^{m+\alpha}J_{m+\alpha}e^{im\theta},\\ &=\sum_{m=1}^{\infty}(-i)^{m-\alpha}J_{m-\alpha}e^{-im\theta},\\ \psi_3&=(-i)^{|\alpha|}J_{|\alpha|}.\end{aligned} \tag{5}$$

Now ψ_1 satisfies the simple differential equation

$$\begin{aligned}\frac{\partial\psi_1}{\partial r'}&=\sum_{m=1}^{\infty}(-i)^{m+\alpha}J_{m+\alpha}{}'e^{im\theta}\\ &=\sum_{m=1}^{\infty}(-i)^{m+\alpha}\frac{J_{m+\alpha-1}-J_{m+\alpha+1}}{2}e^{im\theta},\quad r'=kr\end{aligned} \tag{6}$$

where we have used the well-known formula for Bessel functions:

$$dJ_\gamma(r)/dr=\tfrac{1}{2}(J_{\gamma-1}-J_{\gamma+1}).$$

As a result, we obtain

$$\begin{aligned}\frac{\partial\psi_1}{\partial r'}&=\frac{1}{2}\sum_{m'=0}^{\infty}(-i)^{m'+\alpha+1}J_{m'+\alpha}e^{i(m'+1)\theta}\\ &\quad-\frac{1}{2}\sum_{m'=2}^{\infty}(-i)^{m'+\alpha-1}J_{m'+\alpha}e^{i(m'-1)\theta}\\ &=\frac{1}{2}\sum_{m'=1}^{\infty}(-i)^{m'+\alpha}J_{m'+\alpha}e^{im'\theta}(-ie^{i\theta}+i^{-1}e^{-i\theta})\\ &\quad+\tfrac{1}{2}(-i)^{\alpha}[J_{\alpha+1}-ie^{i\theta}J_\alpha].\end{aligned} \tag{7}$$

So

$$\partial\psi_1/\partial r'=-i\cos\theta\psi_1+\tfrac{1}{2}(-i)^{\alpha}(J_{\alpha+1}-iJ_\alpha e^{i\theta}).$$

This differential equation can be easily integrated to give

$$\psi_1=A\int_0^{r'}e^{ir'\cos\theta}[J_{\alpha+1}-iJ_\alpha e^{i\theta}]dr', \tag{8}$$

where

$$A=\tfrac{1}{2}(-i)^{\alpha}e^{-ir'\cos\theta}.$$

The lower limit of the integration is determined by the requirement that when r' goes to zero, ψ_1 also goes to to zero because, as we have seen, ψ_1 includes Bessel functions of positive order only.

In order to discuss the asymptotic behavior of ψ_1, let us write it as $\psi_1=A[I_1-I_2]$, where

$$\begin{aligned}I_1&=\int_0^{\infty}e^{ir'\cos\theta}[J_{\alpha+1}-ie^{i\theta}J_\alpha]dr',\\ I_2&=\int_r^{\infty}e^{ir'\cos\theta}[J_{\alpha+1}-ie^{i\theta}J_\alpha]dr'.\end{aligned} \tag{9}$$

The first of these integrals is known[7]:

$$\int_0^{\infty}e^{i\beta r}J_\alpha(kr)=\frac{e^{i[\alpha\arcsin(\beta/k)]}}{(k^2-\beta^2)^{\frac{1}{2}}},\quad 0<\beta<k,\quad -2<\alpha.$$

In our cases, $\beta=\cos\theta$, $k=1$, so that

$$I_1=\left[\frac{e^{i\alpha(\frac{1}{2}\pi-|\theta|)}}{|\sin\theta|}-ie^{i\theta}\frac{e^{i(\alpha+1)(\frac{1}{2}\pi-|\theta|)}}{|\sin\theta|}\right]. \tag{10}$$

Because the integrand is even in θ, we have written the final expression for the above integral as a function of $|\theta|$ and of $|\sin\theta|$. Hence

$$\begin{aligned}I_1&=e^{i\alpha(\frac{1}{2}\pi-|\theta|)}\left[\frac{ie^{-i|\theta|}-ie^{i\theta}}{|\sin\theta|}\right]\\ &=0\quad\text{for }\theta<0,\\ &=e^{-i\alpha\theta}2i^{\alpha}\quad\text{for }\theta>0,\end{aligned} \tag{11}$$

where we have taken θ as going from $-\pi$ to π.

[7] See, for example, W. Gröbner and N. Hofreiter, *Integraltafel* (Springer-Verlag, Berlin, 1949).

We shall see presently that I_1 represents the largest term in the asymptotic expansion of ψ_1. The fact that it is zero for $\theta<0$ shows that this part of ψ_1 passes (asymptotically) only on the upper side of the singularity. To explain this, we note that ψ_1 contains only positive values of m, and therefore of the angular momentum. It is quite natural then that this part of ψ_1 goes on the upper side of the singularity. Similarly, since according to (5)

$$\psi_2(r',\theta,\alpha)=\psi_1(r',-\theta,-\alpha),$$

it follows that ψ_2 will behave oppositely to ψ_1 in this regard, so that together they will make up the correct incident wave.

Now, in the limit of $r'\to\infty$ we are allowed to take in the integrand of I_2 the first asymptotic term of J_α,[8] namely $J_\alpha\to(2/\pi r)^{\frac{1}{2}}\cos(r'-\frac{1}{2}\alpha-\frac{1}{4}\pi)$. We obtain

$$I_2=\int_r^\infty e^{ir'\cos\theta}(J_{\alpha+1}-ie^{i\theta}J_\alpha)dr'\to C+D, \quad (12)$$

where

$$C=\int_r^\infty e^{ir'\cos\theta}\left[\cos(r'-\tfrac{1}{2}(\alpha+1)\pi-\tfrac{1}{4}\pi)\right]\frac{dr'}{(r')^{\frac{1}{2}}}\left(\frac{2}{\pi}\right)^{\frac{1}{2}},$$
$$D=\int_r^\infty e^{ir'\cos\theta}\left[\cos(r'-\tfrac{1}{2}\alpha-\tfrac{1}{4}\pi)\right]\frac{dr'}{(r')^{\frac{1}{2}}}\left(\frac{2}{\pi}\right)^{\frac{1}{2}}(-i)e^{i\theta}. \quad (13)$$

Then

$$C=\int_r^\infty e^{ir'\cos\theta}\left[e^{i[r'-\frac{1}{2}(\alpha+1)\pi-\frac{1}{4}\pi]}+e^{-i[r'-\frac{1}{2}(\alpha+1)\pi-\frac{1}{4}\pi]}\right]\frac{dr'}{(2\pi r')^{\frac{1}{2}}}$$
$$=\left(\frac{2}{\pi}\right)^{\frac{1}{2}}\frac{(-i)^{\alpha+\frac{3}{2}}}{(1+\cos\theta)^{\frac{1}{2}}}\int_{[r'(1+\cos\theta)]^{\frac{1}{2}}}^\infty\exp(+iz^2)dz$$
$$+\left(\frac{2}{\pi}\right)^{\frac{1}{2}}\frac{i^{\alpha+\frac{3}{2}}}{(1-\cos\theta)^{\frac{1}{2}}}\int_{[r'(1-\cos\theta)]^{\frac{1}{2}}}^\infty\exp(-iz^2)dz, \quad (14)$$

where we have put

$$z=[r'(1+\cos\theta)]^{\frac{1}{2}} \quad\text{and}\quad z=[r'(1-\cos\theta)]^{\frac{1}{2}},$$

respectively.

Using now the well-known asymptotic behavior of the error function,[9]

$$\int_a^\infty\exp(iz^2)dz\to\frac{i}{2}\frac{\exp(ia^2)}{a},$$
$$\int_a^\infty\exp(-iz^2)dz\to\frac{-i}{2}\frac{\exp(-ia^2)}{a}, \quad (15)$$

we finally obtain

$$C=\left[\frac{(-i)^{\alpha+\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}}\frac{e^{ir'}}{[r'(1+\cos\theta)^2]^{\frac{1}{4}}}+\frac{i^{\alpha+\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}}\frac{e^{-ir'}}{[r'(1-\cos\theta)^2]^{\frac{1}{4}}}\right]e^{ir'\cos\theta}, \quad (16)$$

$$D=\left[\frac{(-i)^{\alpha-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}}\frac{e^{ir'}}{[r'(1+\cos\theta)^2]^{\frac{1}{4}}}+\frac{i^{\alpha-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}}\frac{e^{-ir'}}{[r'(1-\cos\theta)^2]^{\frac{1}{4}}}\right]e^{ir'\cos\theta}(-i)e^{i\theta}. \quad (17)$$

Now adding (16) and (17) together and using (13) and (9), we find that the term of $1/(r')^{\frac{1}{2}}$ in the asymptotic expansion of ψ_1 is

$$\frac{(-i)^{\frac{1}{2}}}{2(2\pi)^{\frac{1}{2}}}\left[(-1)^\alpha\frac{e^{ir'}}{(r')^{\frac{1}{2}}}\frac{1+e^{i\theta}}{1+\cos\theta}+i\frac{e^{-ir'}}{(r')^{\frac{1}{2}}}\frac{1-e^{i\theta}}{1-\cos\theta}\right]. \quad (18)$$

Using again the relation between ψ_1 and ψ_2 we obtain for the corresponding term in ψ_2

$$\frac{(-i)^{\frac{1}{2}}}{2(2\pi)^{\frac{1}{2}}}\left[(-1)^{-\alpha}\frac{e^{ir'}}{(r')^{\frac{1}{2}}}\frac{1+e^{-i\theta}}{1+\cos\theta}+i\frac{e^{-ir'}}{(r')^{\frac{1}{2}}}\frac{1-e^{-i\theta}}{1-\cos\theta}\right]. \quad (19)$$

Adding (18) and (19) and using (11), we finally get

$$\psi_1+\psi_2\to\frac{(-i)^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}}\left[\frac{ie^{-ir'}}{(r')^{\frac{1}{2}}}+\frac{e^{ir'}}{(r')^{\frac{1}{2}}}\frac{\cos(\pi\alpha-\frac{1}{2}\theta)}{\cos(\frac{1}{2}\theta)}\right]+e^{-i(r'\cos\theta+\alpha\theta)}. \quad (20)$$

There remains the contribution of ψ_3, whose asymptotic behavior is [see Eq. (12)]

$$(-i)^{|\alpha|}J_{|\alpha|}(r')\to(-i)^{|\alpha|}\left(\frac{2}{\pi r'}\right)^{\frac{1}{2}}\cos(r'-\tfrac{1}{4}\pi-\tfrac{1}{2}|\alpha|\pi).$$

Collecting all terms, we find

$$\psi=\psi_1+\psi_2+\psi_3\to e^{-i(\alpha\theta+r'\cos\theta)}+\frac{e^{ir'}}{(2\pi ir')^{\frac{1}{2}}}\sin\pi\alpha\frac{e^{-i\theta/2}}{\cos(\theta/2)}, \quad (21)$$

where the $\pm$ sign is chosen according to the sign of α.

The first term in equation (21) represents the incident wave, and the second the scattered wave.[10] The scattering cross section is therefore

$$\sigma=\frac{\sin^2\pi\alpha}{2\pi}\frac{1}{\cos^2(\theta/2)}. \quad (22)$$

[8] E. Jahnke and F. Emde, *Tables of Functions* (Dover Publications, Inc., New York, 1943), fourth edition, p. 138.

[9] Reference 8, p. 24.

[10] In this way, we verify, of course, that our choice of the a_m for Eq. (3) satisfies the correct boundary conditions.

When $\alpha=n$, where n is an integer, then σ vanishes. This is analogous to the Ramsauer effect.[11] σ has a maximum when $\alpha=n+\frac{1}{2}$.

The asymptotic formula (21) holds only when we are not on the line $\theta=\pi$. The exact solution, which is needed on this line, would show that the second term will combine with the first to make a single-valued wave function, despite the non-single-valued character of the two parts, in the neighborhood of $\theta=\pi$. We shall see this in more detail presently for the special case $\alpha=n+\frac{1}{2}$.

In the interference experiment discussed in Sec. 2, diffraction effects, represented in Eq. (21) by the scattered wave, have been neglected. Therefore, in this problem, it is adequate to use the first term of Eq. (21). Here, we see that the phase of the wave function has a different value depending on whether we approach the line $\theta=\pm\pi$ from positive or negative angles, i.e., from the upper or lower side. This confirms the conclusions obtained in the approximate treatment of Sec. 2.

We shall discuss now the two special cases that can be solved exactly. The first is the case where $\alpha=n$. Here, the wave function is $\psi=e^{-ikx}e^{-i\alpha\theta}$, which is evidently single-valued when α is an integer. (It can be seen by direct differentiation that this is a solution.)

The second case is that of $\alpha=n+\frac{1}{2}$. Because $J_{(n+\frac{1}{2})}(r)$ is a closed trigonometric function, the integrals for ψ can be carried out exactly.

The result is

$$\psi=\frac{i^{\frac{1}{2}}}{\sqrt{2}}e^{-i(\frac{1}{2}\theta+r'\cos\theta)}\int_0^{[r'(1+\cos\theta)]^{\frac{1}{2}}}\exp(iz^2)dz. \qquad (23)$$

This function vanishes on the line $\theta=\pi$. It can be seen that its asymptotic behavior is the same as that of Eq. (2) with α set equal to $n+\frac{1}{2}$. In this case, the single-valuedness of ψ is evident. In general, however, the behavior of ψ is not so simple, since ψ does not become zero on the line $\theta=\pi$.

5. DISCUSSION OF SIGNIFICANCE OF RESULTS

The essential result of the previous discussion is that in quantum theory, an electron (for example) can be influenced by the potentials even if all the field regions are excluded from it. In other words, in a field-free multiply-connected region of space, the physical properties of the system still depend on the potentials.

It is true that all these effects of the potentials depend only on the gauge-invariant quantity $\oint\mathbf{A}\cdot d\mathbf{x}=\int\mathbf{H}\cdot d\mathbf{s}$, so that in reality they can be expressed in terms of the fields inside the circuit. However, according to current relativistic notions, all fields must interact only locally. And since the electrons cannot reach the regions where the fields are, we cannot interpret such effects as due to the fields themselves.

In classical mechanics, we recall that potentials cannot have such significance because the equation of motion involves only the field quantities themselves. For this reason, the potentials have been regarded as purely mathematical auxiliaries, while only the field quantities were thought to have a direct physical meaning.

In quantum mechanics, the essential difference is that the equations of motion of a particle are replaced by the Schrödinger equation for a wave. This Schrödinger equation is obtained from a canonical formalism, which cannot be expressed in terms of the fields alone, but which also requires the potentials. Indeed, the potentials play a role, in Schrödinger's equation, which is analogous to that of the index of refration in optics, The Lorentz force $[e\mathbf{E}+(e/c)\mathbf{v}\times\mathbf{H}]$ does not appear anywhere in the fundamental theory, but appears only as an approximation holding in the classical limit. It would therefore seem natural at this point to propose that, in quantum mechanics, the fundamental physical entities are the potentials, while the fields are derived from them by differentiations.

The main objection that could be raised against the above suggestion is grounded in the gauge invariance of the theory. In other words, if the potentials are subject to the transformation $\mathbf{A}_\mu\rightarrow A_\mu'=A_\mu+\partial\psi/\partial x_\mu$, where ψ is a continuous scalar function, then all the known physical quantities are left unchanged. As a result, the same physical behavior is obtained from any two potentials, $A_\mu(x)$ and $A_\mu'(x)$, related by the above transformation. This means that insofar as the potentials are richer in properties than the fields, there is no way to reveal this additional richness. It was therefore concluded that the potentials cannot have any meaning, except insofar as they are used mathematically, to calculate the fields.

We have seen from the examples described in this paper that the above point of view cannot be maintained for the general case. Of course, our discussion does not bring into question the gauge invariance of the theory. But it does show that in a theory involving only local interactions (e.g., Schrödinger's or Dirac's equation, and current quantum-mechanical field theories), the potentials must, in certain cases, be considered as physically effective, even when there are no fields acting on the charged particles.

The above discussion suggests that some further development of the theory is needed. Two possible directions are clear. First, we may try to formulate a nonlocal theory in which, for example, the electron could interact with a field that was a finite distance away. Then there would be no trouble in interpreting these results, but, as is well known, there are severe difficulties in the way of doing this. Secondly, we may retain the present local theory and, instead, we may try to give a further new interpretation to the poten-

[11] See, for example, D. Bohm, *Quantum Theory* (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1951).

tials. In other words, we are led to regard $A_\mu(x)$ as a physical variable. This means that we must be able to define the physical difference between two quantum states which differ only by gauge transformation. It will be shown in a future paper that in a system containing an undefined number of charged particles (i.e., a superposition of states of different total charge), a new Hermitian operator, essentially an angle variable, can be introduced, which is conjugate to the charge density and which may give a meaning to the gauge. Such states have actually been used in connection with recent theories of superconductivity and superfluidity[12] and we shall show their relation to this problem in more detail.

ACKNOWLEDGMENTS

We are indebted to Professor M. H. L. Pryce for many helpful discussions. We wish to thank Dr. Chambers for many discussions connected with the experimental side of the problem.

[12] See, for example, C. G. Kuper, *Advances in Physics*, edited by N. F. Mott (Taylor and Francis, Ltd., London, 1959), Vol. 8, p. 25, Sec. 3, Par. 3.

PHYSICAL REVIEW VOLUME 115, NUMBER 3 AUGUST 1, 1959

Theory of Multiple Scattering: Second Born Approximation and Corrections to Molière's Work

B. P. NIGAM,* M. K. SUNDARESAN,* AND TA-YOU WU
Division of Pure Physics, National Research Council, Ottawa, Canada

(Received March 11, 1959)

The formula given by Molière for the scattering cross section of a charged particle by an atom, on which has been based the formula for the "screening angle" χ_a in his theory of multiple scattering, has been examined and found to contain an inconsistent approximation in all orders of the parameter $\alpha_1 = zZ/137\beta$ except the lowest (the first Born approximation). In the present work, the correct expression of Dalitz is used for the single-scattering cross section of a relativistic Dirac particle by a screened atomic field up to the second Born approximation. It is found that the effect of the deviation from the first Born approximation on the screening angle is much smaller than Molière's expression for this quantity would lead one to believe. This is so because the deviation from the first Born approximation is very small at the small angles that go into the definition of the screening angle. In Molière's work, all the effect of the deviation from the first Born approximation on the distribution function $f(\theta)$ for multiple scattering is contained in the quantity B which depends only on χ_a. In the present work, it is shown that in a consistent treatment of terms of various orders in α_1, there exist additional terms of order $zZ/137$ in the distribution function. These terms, which represent the second Born approximation, become important at large angles. Calculations have been carried out for the scattering of 15.6-Mev electrons by Au and Be. The $1/e$ widths of the distribution function obtained are in good agreement with the experimental result of Hanson *et al.*, whereas Molière's theory gives too great a width compared with the experimental value in the case of Be.

I. INTRODUCTION

THE theory of scattering of fast charged particles by atoms is of importance for the analysis of such experimental results as the scattering of high-energy mesons and electrons in going through sheets of matter. An "exact" theory of multiple scattering has been given by Goudsmit and Saunderson.[1] Its application to a specific scattering problem invokes the knowledge of the law of single scattering by an isolated atom. In a paper in 1947, Molière[2] gives a (nonrelativistic) formula for the scattering of a fast charged particle by a screened Coulomb field, in which an approximation higher than the usual first Born approximation is attempted. In a second paper Molière[3] gives a theory of multiple scattering which has later been shown by Bethe[4] to be obtainable from the theory of Goudsmit and Saunderson by making certain approximations. For the single-scattering law to be used in the theory of multiple scattering, Molière uses the result he obtained in his earlier paper.[2]

Hanson *et al.*[5] have measured the scattering of 15.6-Mev electrons by gold and beryllium foils and compared their experimental results with those calculated according to Molière's theory. The calculated "$1/e$ width" of the distribution has been found to be in excellent agreement with the observed value in the case of gold, but is somewhat too large in the case of beryllium.

In the case of the scattering of μ mesons (in cosmic rays) by matter, the rather scanty data[6] (for large scattering angles) seem to be in agreement with Molière's theory. Here, for high enough energies of the

* National Research Council Postdoctorate Fellows.

[1] S. A. Goudsmit and J. L. Saunderson, Phys. Rev. **57**, 24 (1940), and **58**, 36 (1940).

[2] G. Molière, Z. Naturforsch. **2a**, 133 (1947).

[3] G. Molière, Z. Naturforsch. **3a**, 78 (1948).

[4] H. A. Bethe, Phys. Rev. **89**, 1256 (1953).

[5] Hanson, Lanzl, Lyman, and Scott, Phys. Rev. **84**, 634 (1951).

[6] George, Redding, and Trent, Proc. Phys. Soc. (London) **A66**, 533 (1953); I. B. McDiarmid, Phil. Mag. **45**, 933 (1954); **46**, 177 (1955).

Phase Change during a Cyclic Quantum Evolution

Y. Aharonov and J. Anandan
Department of Physics and Astronomy, University of South Carolina, Columbia, South Carolina 29208
(Received 29 December 1986)

A new geometric phase factor is defined for any cyclic evolution of a quantum system. This is independent of the phase factor relating the initial- and final-state vectors and the Hamiltonian, for a given projection of the evolution on the projective space of rays of the Hilbert space. Some applications, including the Aharonov-Bohm effect, are considered. For the special case of adiabatic evolution, this phase factor is a gauge-invariant generalization of the one found by Berry.

PACS numbers: 03.65.−w

A type of evolution of a physical system which is often of interest in physics is one in which the state of the system returns to its original state after an evolution. We shall call this a cyclic evolution. An example is periodic motion, such as the precession of a particle with intrinsic spin and magnetic moment in a constant magnetic field. Another example is the adiabatic evolution of a quantum system whose Hamiltonian H returns to its original value and the state evolves as an eigenstate of the Hamiltonian and returns to its original state. A third example is the splitting and recombination of a beam so that the system may be regarded as going backwards in time along one beam and returning along the other beam to its original state at the same time.

Now, in quantum mechanics, the initial- and final-state vectors of a cyclic evolution are related by a phase factor $e^{i\phi}$, which can have observable consequences. An example, which belongs to the second category mentioned above, is the rotation of a fermion wave function by 2π rad by adiabatic rotation of a magnetic field[1] through 2π rad so that $\phi=\pm\pi$. Recently, Berry[2] has shown that when H, which is a function of a set of parameters R^i, undergoes adiabatic evolution along a closed curve Γ in the parameter space, then a state that remains an eigenstate of $H(\mathbf{R})$ corresponding to a simple eigenvalue $E_n(\mathbf{R})$ develops a geometrical phase γ_n which depends only on Γ. Simon[3] has given an interpretation of this phase as due to holonomy in a line bundle over the parameter space. Anandan and Stodolsky[4] have shown how the Berry phases for the various eigenspaces can be obtained from the holonomy in a vector bundle. For the adiabatic motion of spin, this is determined by a rotation angle α, due to the parallel transport of a Cartesian frame with one axis along the spin direction, which contains the above-mentioned rotation by 2π radians as a special case. The result of a recent experiment[5] to observe Berry's phase for light can also be understood as a rotation of the plane of polarization by this angle α.

In this Letter, we consider the phase change for *all* cyclic evolutions which contain the three examples above as special cases. We show the existence of a phase associated with cyclic evolution, which is universal in the sense that it is the same for the infinite number of possible motions along the curves in the Hilbert space $\mathcal{H}$ which project to a given closed curve $\hat{C}$ in the projective Hilbert space $\mathcal{P}$ of rays of $\mathcal{H}$ and the possible Hamiltonians $H(t)$ which propagate the state along these curves. This phase tends to the Berry phase in the adiabatic limit if $H(t)\equiv H[\mathbf{R}(t)]$ is chosen accordingly. For an electrically charged system, we formulate this phase gauge invariantly and show that the Aharonov-Bohm (AB) phase[6] due to the electromagnetic field may be regarded as a special case. This generalizes the gauge-noninvariant result of Berry that the AB phase due to a static magnetic field is a special case of his phase. This also removes the mystery of why the AB phase, even in this special case, should emerge from Berry's expression even though the former is independent of this adiabatic approximation.

Suppose that the normalized state $|\psi(t)\rangle\in\mathcal{H}$ evolves according to the Schrödinger equation

$$H(t)|\psi(t)\rangle=i\hbar(d/dt)|\psi(t)\rangle, \tag{1}$$

such that $|\psi(\tau)\rangle=e^{i\phi}|\psi(0)\rangle$, ϕ real. Let $\Pi:\mathcal{H}\rightarrow\mathcal{P}$ be the projection map defined by $\Pi(|\psi\rangle)=\{|\psi'\rangle:|\psi'\rangle=c|\psi\rangle$, c is a complex number$\}$. Then $|\psi(t)\rangle$ defines a curve C: $[0,\tau]\rightarrow\mathcal{H}$ with $\hat{C}\equiv\Pi(C)$ being a closed curve in $\mathcal{P}$. Conversely given any such curve C, we can define a Hamiltonian function $H(t)$ so that (1) is satisfied for the corresponding normalized $|\psi(t)\rangle$. Now define $|\tilde{\psi}(t)\rangle=e^{-if(t)}|\psi(t)\rangle$ such that $f(\tau)-f(0)=\phi$. Then $|\tilde{\psi}(\tau)\rangle=|\tilde{\psi}(0)\rangle$ and from (1),

$$-\frac{df}{dt}=\frac{1}{\hbar}\langle\psi(t)|H|\psi(t)\rangle-\langle\tilde{\psi}(t)|i\frac{d}{dt}|\tilde{\psi}(t)\rangle. \tag{2}$$

Hence, if we remove the dynamical part from the phase ϕ by defining

$$\beta\equiv\phi+\hbar^{-1}\int_0^\tau\langle\psi(t)|H|\psi(t)\rangle dt, \tag{3}$$

it follows from (2) that

$$\beta=\int_0^\tau\langle\tilde{\psi}|i(d|\tilde{\psi}\rangle/dt)dt. \tag{4}$$

Now, clearly, the same $|\tilde{\psi}(t)\rangle$ can be chosen for every curve C for which $\Pi(C)=\hat{C}$, by appropriate choice of

$f(t)$. Hence β, defined by (3), is independent of ϕ and H for a given closed curve $\hat{C}$. Indeed, for a given $\hat{C}$, $H(t)$ can be chosen so that the second term in (3) is zero, which may be regarded as an alternative definition of β. Also, from (4), β is independent of the parameter t of $\hat{C}$, and is uniquely defined up to $2\pi n$ (n=integer). Hence $e^{i\beta}$ is a geometric property of the unparametrized image of $\hat{C}$ in $\mathcal{P}$ only.

Consider now a slowly varying $H(t)$, with $H(t)|n(t)\rangle = E_n(t)|n(t)\rangle$, for a complete set $\{|n(t)\rangle\}$. If we write

$$|\psi(t)\rangle = \sum_n a_n(t) \exp\left(-\frac{i}{\hbar}\int E_n\,dt\right)|n(t)\rangle,$$

and use (1) and the time derivative of the eigenvector equation,[7] we have

$$\dot{a}_m = -a_m\langle m|\dot{m}\rangle - \sum_{n\neq m} a_n \frac{\langle m|\dot{H}|n\rangle}{E_n - E_m}\exp\left[\frac{i}{\hbar}\int (E_m - E_n)dt\right], \tag{5}$$

where the dot denotes time derivative. Suppose that

$$\sum_{n\neq m}\left|\frac{\hbar\langle m|\dot{H}|n\rangle}{(E_n-E_m)^2}\right| \ll 1. \tag{6}$$

Then if $a_n(0)=\delta_{nm}$, the last term in (5) is negligible and the system would therefore continue as an eigenstate of $H(t)$, to a good approximation.

In this adiabatic approximation, (5) yields

$$a_m(t) \simeq \exp\left(-\int\langle m|\dot{m}\rangle dt\right)a_m(0).$$

For a cyclic adiabatic evolution, the phase $i\int_0^T\langle m|\dot{m}\rangle dt$ is independent of the chosen $|m(t)\rangle$ and Berry[2] regarded this as a geometrical property of the parameter space of which H is a function. But this phase is the same as (4) on our choosing $|\tilde{\psi}(t)\rangle \simeq |m(t)\rangle$ in the present approximation. But β, defined by (3), does not depend on any approximation; so (4) is exactly valid. Moreover, $|\psi(t)\rangle$ need not be an eigenstate of $H(t)$, unlike in the limiting case studied by Berry. Also, the two examples below will show respectively that it is neither necessary nor sufficient to go around a (nontrivial) closed curve in parameter space in order to have a cyclic evolution, with our associated geometric phase β. For these reasons, we regard β as a geometric phase associated with a closed curve in the projective Hilbert space and not the parameter space, even in the special case considered by Berry. But given a cyclic evolution, an $H(t)$ which generated this evolution can be found so that the adiabatic approximation is valid. Then β can be computed with the use of the expression given by Berry in terms of the eigenstates of this Hamiltonian.

We now consider two examples in which the phase β emerges naturally and is observable, in principle, even though the adiabatic approximation is not valid. Suppose that a spin-$\frac{1}{2}$ particle with a magnetic moment is in a homogeneous magnetic field $\mathbf{B}$ along the z axis. Then the Hamiltonian in the rest frame is $H_1 = -\mu B\sigma_z$, where

$$\sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Also,

$$|\psi(0)\rangle = \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2) \end{pmatrix}$$

so that

$$|\psi(t)\rangle = \exp(i\mu Bt\sigma_z/\hbar)|\psi(0)\rangle = \begin{pmatrix} \exp(i\mu Bt/\hbar)\cos(\theta/2) \\ \exp(-i\mu Bt/\hbar)\sin(\theta/2) \end{pmatrix},$$

which corresponds to the spin direction being always at an angle θ to the z axis. This evolution is periodic with period $\tau = \pi\hbar/\mu B$. Then from (3), for each cycle, $\beta = \pi(1-\cos\theta)$, up to the ambiguity of adding $2\pi n$. Hence, β is $\frac{1}{2}$ of the solid angle subtended by a curve traced on a sphere, by the direction of the spin state, at the center. This is like the Berry phase except that in the latter case (1) the solid angle is subtended by a curve traced by the magnetic field $\mathbf{B}'(t)$ which is large [i.e., $\mu B'/\hbar \gg \omega$, the frequency of the orbit of $\mathbf{B}'(t)$] so that the adiabatic approximation is valid, and (2) $|\psi(t)\rangle$ is assumed to be an eigenstate of this Hamiltonian. Indeed, we may substitute such a Hamiltonian for the above H_1 or add it to H_1 with $\omega = 2\mu B/\hbar$, without changing β, in this approximation. The spin state will also move through the same closed curve in the projective Hilbert space as above if the magnetic field $\mathbf{B} = (B_0\cos\omega t,\ B_0\sin\omega t,\ B_3)$ with $\cot\theta = (B_3 - \hbar\omega/2\mu)/B_0$, where $B_0 \neq 0$.[8] And β is the same for all such Hamiltonians. This illustrates the statement earlier that β is the same for all curves C in H with the same $\hat{C} \equiv \Pi(C)$. Also, β may be interpreted as arising from the holonomy transformation, around the closed curve on the above sphere traced by the direction of the spin state, due to the curvature on this sphere,[4] which is a rotation. By varying appropriately a magnetic field applied to the two arms of a neutron interferometer with polarized neutrons, it is possible to make the dynamical part of β [the last term in (3)] the same for the two beams.[2,4] Then the phase difference between the two beams is just the geometrical phase, which is observable in principle, from the interference pattern, even when the magnetic field is varied nonadiabatically. In particular, a phase difference of $\pm\pi$ rad would correspond to a 2π-rad rotation of the fermion wave function, which is thus observable.

As our second example, suppose that the magnetic field is $\mathbf{B}(t) = \mathbf{B}_0 + \mathbf{B}_1(t)$, where $\mathbf{B}_0$ is constant and $\mathbf{B}_1(t)$ rotates slowly in a plane containing $\mathbf{B}_0$ with $|\mathbf{B}_1(t)|$

$=|\mathbf{B}_0|$. Suppose that at time t the angle between $\mathbf{B}_1$ and $\mathbf{B}_0$ is $\pi-\theta(t)$ and the spin state $|\psi(t)\rangle$ is in an approximate eigenstate of $H(t)=\mu\mathbf{B}\cdot\boldsymbol{\sigma}$, where σ^i are the Pauli spin matrices. For $0\leq\theta\ll 1$, the adiabatic condition (6) gives $0\leq -\hbar\dot{\theta}/\mu B_0\theta\ll 1$, assuming $\dot{\theta}\leq 0$. Hence $\theta\gg\theta_0\exp(-\mu B_0 t/\hbar)>0$. So θ can never become zero. That is, if $\mathbf{B}(T)=\mathbf{0}$ for some T then the adiabatic approximation, as defined above, cannot be satisfied, regardless of how slowly $\mathbf{B}_1(t)$ rotates. However, because of conservation of angular momentum, $|\psi(t)\rangle$ remains an eigenstate of $H(t)$ even at $t=T$. But if θ changes monotonically then a level crossing occurs at the point of degeneracy ($\mathbf{B}=\mathbf{0}$) so that the energy eigenvalue corresponding to $|\psi(t)\rangle$ changes sign at $t=T$. For each rotation of $\mathbf{B}_1$ by 2π rad, $|\psi\rangle$ rotates by π rad, so that the system returns to its original state after two rotations of $\mathbf{B}(t)$. For this cyclic evolution, our $\beta=\pi$ which can be seen from the fact that a spin-$\frac{1}{2}$ particle acquires a phase π during a rotation, or that the curve $\hat{C}$ on the projective Hilbert space, which is a sphere, is a great circle, subtending a solid angle 2π at the center. This example is similar to Berry's phase in that $|\psi(t)\rangle$ is always an eigenstate of $H(t)$, even though Berry's prescription cannot be applied here because of the crossing of the point of degeneracy at which the adiabatic approximation breaks down.

Consider now a system with electric charge q for which $H=H_k(\mathbf{p}-(q/c)\hat{\mathbf{A}}(t),R_i)+q\hat{A}_0(t)$ in (1). Here, $\langle\mathbf{x}|\hat{A}_\mu(t)|\psi(t')\rangle=A_\mu(\mathbf{x},t)\psi(\mathbf{x},t')$, where $A_\mu(\mathbf{x},t)$ is the usual electromagnetic four-potential, and R_i are some parameters. Under a gauge transformation,

$$|\psi(t)\rangle\rightarrow\exp[i(q/c)\hat{\Lambda}(t)]\,|\psi(t)\rangle,$$

$$\hat{A}_0(t)\rightarrow\hat{A}_0(t)-c^{-1}\partial\hat{\Lambda}(t)/\partial t,$$

and

$$H_k(t)\rightarrow\exp[i(q/c)\hat{\Lambda}(t)]H_k(t)\exp[-i(q/c)\hat{\Lambda}(t)].$$

As before, define $|\tilde{\psi}(t)\rangle=e^{-if(t)}|\psi(t)\rangle$. If we require that $|\tilde{\psi}\rangle$ undergo the same gauge transformation as $|\psi(t)\rangle$, $f(t)$ is gauge invariant. Then, from (1),

$$\frac{df}{dt}(t)=\langle\tilde{\psi}(t)|\frac{d}{dt}-\frac{q}{\hbar}\hat{A}_0(t)|\tilde{\psi}(t)\rangle-\frac{1}{\hbar}\langle\psi(t)|H_k(t)|\psi(t)\rangle. \tag{7}$$

We consider now a cyclic evolution so that

$$|\psi(\tau)\rangle=e^{i\phi}\exp\left(-\frac{iq}{\hbar}\int_0^\tau\hat{A}_0\,dt\right)|\psi(0)\rangle.$$

Choose $f(t)$ so that $\phi=f(\tau)-f(0)$. Then

$$|\tilde{\psi}(\tau)\rangle=\exp\left(-i\frac{q}{\hbar}\int_0^\tau\hat{A}_0\,dt\right)|\tilde{\psi}(0)\rangle.$$

So we now define the gauge-invariant generalization of (3) as

$$\beta\equiv\phi+\frac{1}{\hbar}\int_0^\tau\langle\psi(t)|H_k(t)|\psi(t)\rangle dt, \tag{8}$$

which on use of (7) gives

$$\beta=\int_0^\tau\langle\tilde{\psi}(t)|i\frac{d}{dt}-\frac{q}{\hbar}\hat{A}_0(t)|\tilde{\psi}(t)\rangle dt. \tag{9}$$

Here, $|\tilde{\psi}(\tau)\rangle$ is obtained by parallel transport of $|\tilde{\psi}(0)\rangle$, with respect to the electromagnetic connection, along the congruence of lines parallel to the time axis. We could have chosen, instead, any other congruence of paths from $t=0$ to $t=\tau$ in our definition of ϕ and therefore $|\tilde{\psi}(\tau)\rangle$. This would correspondingly change β, which therefore depends on the chosen congruence. But, again, β is independent of ϕ and $H(t)$ for all the motions in $\mathcal{H}$ that project to the same closed curve $\hat{C}$ in $\mathcal{P}$, for a given chosen congruence. Both β and ϕ, which satisfies

$$e^{-i\phi}=\langle\psi(\tau)|\exp\left(-\frac{iq}{c}\int_0^\tau\hat{A}_0\,dt\right)|\psi(0)\rangle,$$

are gauge invariant. In the adiabatic limit, $|\tilde{\psi}(t)\rangle$ can be chosen to be an eigenstate of $H_k(t)$ and (9) is then a gauge-invariant generalization of the Berry phase.

We illustrate this by means of the AB effect.[6] Berry has obtained the AB phase from the gauge-noninvariant expression (4) with $|\tilde{\psi}(t)\rangle$ an eigenstate of $H(t)$, for a stationary magnetic field, in a special gauge.[9] But a gauge can be chosen so that the AB phase is included in the dynamical phase instead of the geometrical phase (4). Also, in general, there is no cyclic evolution in an AB experiment. But our β defined by Eq. (8) or (9) is gauge invariant and includes the AB phase in the special case to be described now.

Suppose that a charged-particle beam is split into two beams at $t=0$ which, after traveling in field-free regions, are recombined so that they have the same state at $t=\tau$. It is assumed here that the splitting and the subsequent evolution of the two beams occur under the action of two separate Hamiltonians. This is possible if we restrict ourselves to the Hilbert space of a subset of the degrees of freedom of a given system, as in the example considered by Aharonov and Vardi.[10] This belongs to the third example of a cyclic evolution mentioned at the beginning of this Letter. The wave function of each beam

at $t=\tau$, assuming that it has a fairly well defined momentum, is

$$\psi_i(\mathbf{x},\tau)=\exp\left(-\frac{i}{\hbar}\int_0^\tau E_i\,dt\right)\exp\left(-\frac{iq}{c}\int_{\gamma_i}A_\mu dx^\mu\right)\exp\left(\frac{i}{\hbar}\int_{\gamma_i}\mathbf{p}\cdot d\mathbf{x}\right)\psi(\mathbf{x},0),\quad i=1 \text{ or } 2,$$

where γ_i is a space-time curve through the beam and $\mathbf{p}$ represents the approximate kinetic momentum of the beam. Hence on using (8), we have

$$\beta=-\frac{q}{c}\oint_\gamma A_\mu dx^\mu+\frac{1}{\hbar}\oint_\gamma \mathbf{p}\cdot d\mathbf{x}, \tag{10}$$

where γ is the closed curve formed from γ_1 and γ_2. But this is only an approximate treatment and a more careful investigation of this problem is needed.

In conclusion, we note that $\mathcal{H}^*=\mathcal{H}-\{0\}$ is a principal fiber bundle over $\mathcal{P}$ with structure group C^* (the group of nonzero complex numbers), and the disjoint union of the rays in $\mathcal{H}$ is the natural line bundle over $\mathcal{P}$ whose fiber above any $p\in\mathcal{P}$ is p itself. Then, clearly, β, given by (4), arises from the holonomy due to a connection in either bundle such that $|\psi(t)\rangle$ is parallel transported if

$$\langle\psi(t)|(d/dt)|\psi(t)\rangle=0, \tag{11}$$

i.e., the horizontal spaces are perpendicular to the fibers with respect to the Hilbert space inner product. Condition (11) was used by Simon[3] to define a connection on a line bundle over parameter space, which is different from the above bundles. The real part of (11) says that $\langle\psi(t)|\psi(t)\rangle$ is constant during parallel transport. Since this is true also during any time evolution determined by (1), we may restrict consideration to the subbundle $\mathcal{F}=\{|\psi\rangle\in\mathcal{H}:\langle\psi|\psi\rangle=1\}$ of $\mathcal{H}^*$. This $\mathcal{F}$ is the Hopf bundle[11] over $\mathcal{P}$. Then the imaginary part of (11) defines the horizontal spaces in $\mathcal{F}$ which determine a connection. This is the usual connection in $\mathcal{F}$ and $e^{i\beta}$ is the holonomy transformation associated with it. If $\mathcal{H}$ has finite dimension N then $\mathcal{P}$ has dimension $N-1$. For $N=2$, $\mathcal{P}$ is the complex projective space $P_1(C)$ which is a sphere with the Fubini-study metric[11] on $\mathcal{P}$ being the usual metric on the sphere. Opposite points on this sphere represent rays containing orthogonal states. Our geometric phase can then be obtained from the holonomy angle α associated with parallel transport around a closed curve on this sphere like in Ref. 4.

It is a pleasure to thank Don Page for suggesting the relevance of the Hopf bundle and the Fubini-Study metric to this work.

[1]Y. Aharonov and L. Susskind, Phys. Rev. **158**, 1237 (1967).

[2]M. V. Berry, Proc. Roy. Soc. London, Ser. A **392**, 45 (1984).

[3]B. Simon, Phys. Rev. Lett. **51**, 2167 (1983).

[4]J. Anandan and L. Stodolsky, Phys. Rev. D **35** 2597 (1987).

[5]R. Y. Chiao and Y.-S. Wu, Phys. Rev. Lett. **57**, 933 (1986); A. Tomita and R. Y. Chiao, Phys. Rev. Lett. **57**, 937 (1986).

[6]Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485 (1959).

[7]See, for example, L. I. Schiff, *Quantum Mechanics* (McGraw-Hill, New York, 1968), pp. 289–291.

[8]An experiment of this type has been done to measure Berry's phase ($\omega\rightarrow 0$) using nuclear magnetic resonance by D. Suter, G. Chingas, R. A. Harris, and A. Pines, to be published. One of us (J.A.) wishes to thank A. Pines for a discussion during which it was realized that the same type of experiment can be used to measure the geometric phase β introduced in the present Letter for nonadiabatic cyclic evolutions as well.

[9]In this proof, in Ref. 2, the eigenfunctions, in the absence of the electromagnetic field, are in effect assumed to be real, in order that Eq. (34) is valid. Since the coefficients of the stationary Schrödinger equation are then real, it is always possible to find real solutions. Then, for any eigenfunction belonging to a given eigenvalue to be necessarily a real function multiplied by $e^{i\lambda}$ ($\lambda=$const), it is necessary and sufficient that the eigenvalue is simple. But in our treatment of the AB effect, it is not necessary to make this assumption.

[10]Y. Aharonov and M. Vardi, Phys. Rev. D **20**, 3213 (1979).

[11]See, S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1969), Vol. 2.

Deflating the Aharonov-Bohm Effect - DRAFT (do not quote or circulate without permission)

David Wallace*

January 30, 2014

Abstract

I argue that the metaphysical import of the Aharonov-Bohm effect has been overstated: correctly understood, it does not require either rejection of gauge invariance or any novel form of nonlocality. The conclusion that it does require one or the other follows from a failure to keep track, in the analysis, of the complex scalar field to which the magnetic vector potential is coupled. Once this is recognised, the way is clear to a local account of the ontology of electrodynamics (or at least, to an account no more nonlocal than quantum theory in general requires); I sketch a possible such account.

1 Introduction

In classical electromagnetism, the magnetic field can be represented either by the field strength $\mathbf{B}$, or by a vector field $\mathbf{A}$ such that $\nabla \times \mathbf{A} = \mathbf{B}$, where in the latter case $\mathbf{A}$ is determined only up to a family of transformations known as *gauge transformations*. Prior to the discovery — and empirical confirmation — of the Aharonov-Bohm (A-B) effect, it was possible to believe (and, I think, widely *was* believed) that $\mathbf{A}$ had only mathematical significance and that a true description of the magnetic field required only $\mathbf{B}$. The A-B effect demonstrated — as uncontroversially as anything in the foundations of physics — that there are features of electromagnetism that transcend the local action of the magnetic field strength on charged matter: electrons can move through a region of space in which $\mathbf{B} = 0$ but which surrounds a region of nonzero $\mathbf{B}$, and their behaviour is dependent upon the value of $\mathbf{B}$ in that latter region. Mathematically speaking these results are possible because the quantum mechanics of electromagnetism involves the interaction of a complex field ψ with the $\mathbf{A}$-field, and the equations that govern that interaction — though gauge-independent — cannot be rewritten in a local way via $\mathbf{B}$ alone.

But just what the conceptual import is remains controversial. In foundational discussions of late it has been argued — and widely accepted — that the

*Balliol College, Oxford; email: david.wallace@balliol.ox.ac.uk

effect requires either that we accept some new form of non-locality beyond that already encountered in quantum mechanics, or that we abandon the principle that gauge transformations simply redescribe the same physical goings on. In particular, the A-B effect rests on the fact that the values of **B** within a spatial region need not determine the field **A** in that region even up to gauge transformations — but that the residual gauge-invariant features of **A** not captured by **B** have an inherently local character to them.

In this paper I argue that much of this debate[1] rests upon a mistake: that of considering the **A**-field in isolation rather than in conjunction with the ψ-field. After reviewing the A-B effect and the contemporary foundational literature in section 2, I demonstrate this in section 3 by considering the gauge-invariant features of ψ and **A** jointly, which are not exhausted by the gauge-invariant features of ψ and **A** separately. I demonstrate that those joint features can in general be given an entirely local characterisation, blocking the concern that some gauge-invariant features are inherently non-local. In section 4 I show in a different way how this apparent nonlocality arises in the study of **A** alone and how it is blocked when we allow for **A** and ψ jointly.

In section 5 I attempt an interpretation of these results: my proposal is that we should not think of ψ and **A** as representing separate entities but as representing, jointly and redundantly, features of a single entity, with the redundancy being localisable either to ψ or to **A** as a matter of pure convention; I illustrate this proposal via brief consideration of the Higgs mechanism.

In sections 6-7 I address two possible concerns with the account I give, and in doing so explore further the extent to which we can give a properly local account of the physical goings on around the solenoid in the A-B effect. Section 8 is the conclusion.

2 The A-B Effect Reviewed

The classical theory of a point electric charge moving under the influence of a background magnetic field is straightforward. The particle is represented mathematically by a vector function $\mathbf{q}(t)$ of time, and the field by a vector field $\mathbf{B}(\mathbf{x}, \mathbf{t})$. The field satisfies two of Maxwell's equations,

$$\nabla \cdot \mathbf{B}(\mathbf{x}, t) = 0 \quad \text{and} \quad \nabla \times \mathbf{B}(\mathbf{x}, t) = 4\pi \mathbf{J}(\mathbf{x}, t), \tag{1}$$

where **J** is the electric current density, and the force on it is given by the Lorentz force law,

$$\mathbf{F}(t) = e\dot{\mathbf{q}}(t) \times \mathbf{B}(\mathbf{q}, t), \tag{2}$$

where e is the particle's charge. (I use Gaussian units with $c = 1$.) In general we will be working in the background-field regime, where the back-reaction of the particle on the field is ignored.

Mathematically, it is always possible to express **B** as the curl of another vector field **A**, the *vector potential*: $\mathbf{B} = \nabla \times \mathbf{A}$. In many cases in classical

[1] Including some parts to which I contributed: cf Wallace and Timpson (2007).

magnetostatics, doing so can be mathematically convenient. For instance, since the divergence of a curl is always zero, the first equation in (1) is automatically satisfied if $\mathbf{B}$ is defined in terms of $\mathbf{A}$. More relevantly for our purposes, the standard way to put the Lorentz force law into Hamiltonian form uses the Hamiltonian

$$H(\mathbf{q}, \mathbf{p}) = \frac{1}{2m}(\mathbf{p} + e\mathbf{A}(\mathbf{q}))^2. \tag{3}$$

That is: it is expressed in terms of the vector potential, rather than the field strength.

At least in classical electromagnetism, the standard assumption is that $\mathbf{A}$ is merely a mathematical convenience, and that $\mathbf{B}$ fully represents the physical features of the magnetic field. There are two interrelated reasons for this:

1. The definition of $\mathbf{A}$ in terms of $\mathbf{B}$ specifies $\mathbf{A}$ only up to the gradient of an arbitrary smooth function Λ: if we replace $\mathbf{A}$ with $\mathbf{A'} = \mathbf{A} + \nabla\Lambda$, then $\nabla \times \mathbf{A'} = \nabla \times \mathbf{A}$.

2. Only $\mathbf{B}$ appears to be physically detectable.

In the Maxwell equations and the Lorentz force law, the dependence of the physics on $\mathbf{B}$ alone rather than $\mathbf{A}$ is manifest. It is only tacit in the Hamiltonian formulation of the theory (there is no straightforward way to write a Hamiltonian form of the Lorentz law in terms of $\mathbf{B}$ alone), but it is strongly suggested by the fact that the *classical gauge transformation*

$$\mathbf{A} \longrightarrow \mathbf{A} + \nabla\Lambda; \quad \mathbf{q} \to \mathbf{q}; \quad \mathbf{p} \to \mathbf{p} - e\nabla\Lambda \tag{4}$$

is a symmetry of the Hamiltonian, and furthermore, a symmetry that leaves the trajectory of the particle unchanged.

In applications of the vector potential in electromagnetism, it is common to impose some additional condition — a *choice of gauge* — such that exactly one $\mathbf{A}$-field is compatible with any given set of empirical data. A common choice, for instance, is the Coulomb gauge, defined by the condition that $\nabla\cdot\mathbf{A} = 0$. If $\mathbf{A}$ and $\mathbf{A'}$ are two gauge-equivalent vector potentials related by a gauge transformation Λ and both satisfying the Coulomb gauge condition, then $\nabla^2\Lambda = 0$, which together with appropriate boundary conditions on the theory entails that Λ is constant and hence that $\mathbf{A} = \mathbf{A'}$.

The quantum mechanics of a particle interacting with a background magnetic field is obtained in the standard way by replacing $\mathbf{q}$ and $\mathbf{p}$ in the classical Hamiltonian with the quantum-mechanical position and momentum operators. The resultant Schrödinger equation (in units where $\hbar = 1$) in the position representation is

$$\frac{\partial\psi}{\partial t}(\mathbf{x}, t) = \frac{i}{2m}\left(\nabla - ie\mathbf{A}(\mathbf{x}, t)\right)^2 \psi(\mathbf{x}, t). \tag{5}$$

The Schrödinger equation is invariant under a quantum-mechanical version of the classical gauge transformation. Since momentum in configuration-space

wave mechanics is given by the gradient of the phase of the wave-function, we would expect that the classical momentum transformation becomes a phase change, and so it does: the form of the transformation is

$$\mathbf{A} \longrightarrow \mathbf{A} + \nabla\Lambda; \quad \psi \longrightarrow \mathrm{e}^{ie\Lambda}\psi, \tag{6}$$

again for an arbitrary smooth function Λ. And just as the classical transformation left particle trajectories unchanged, the quantum version leaves unchanged the probability of finding the particle in any given location.

The gauge-invariance of the Schrödinger equation might suggest that, in quantum just as in classical mechanics, it is the **B**-field rather than the **A**-field that is of physical significnance. The Aharonov-Bohm effect calls this into question: in its simplest form, it works as follows.

1. A beam of charged particles is separated into two; the two beams flow round opposite sides of a solenoid and are then allowed to re-interfere.

2. In the absence of any current through the solenoid (and hence of any induced magnetic field), there will be a set of interference fringes produced by the reinterference of the two beams.

3. When the solenoid is turned on, there will be a shift in the interference pattern. The magnitude of the shift will be proportional to the difference of the integrals of the **A**-field along the paths traversed by the left and right beams respectively. That is, the shift Δ will be proportional to the integral of **A** around the loop Γ formed by the two halves of the beam:

$$\Delta \propto \oint_{\Gamma} \mathbf{A} \cdot \mathrm{d}\mathbf{x} \tag{7}$$

4. By Stokes' theorem, the line integral of a vector field **V** around a closed loop in a simply-connected region (that is: a region in which any closed loop can be continuously deformed to a point without moving any part of it out of the region) is equal to the surface integral of the curl of **V** over any surface bounded by the loop. Since $\nabla \times \mathbf{A} = \mathbf{B}$, this means that Δ is proportional to the integral of the magnetic field over the interior of the region enclosed by the beam, or in other words that it is proportional to the magnetic flux through that region.[2]

The conceptual problem is that a sufficiently well-constructed and well-shielded solenoid will result both in negligible magnetic field *outside* the solenoid, and negligible wavefunction *inside* the solenoid. So the electron is moving (almost) entirely through a region in which the magnetic field is zero — and yet, its evolution is still detectably different from what would occur if the solenoid were turned off.

[2]Of course, the electron will be quite delocalised, and indeed this delocalisation is central to the observation of interference fringes, so "the" path taken by the electron is not really well-defined. But since **B** vanishes outside the solenoid, by Stokes' theorem any two paths which pass the solenoid on the same side will have the same line integral of **A**.

If we hold on to the idea that the magnetic field is completely represented by the field strength $\mathbf{B}$ (what Healey (2007, p.54) calls a 'no new EM properties' view), this means action at a distance: the passage of the electron around the solenoid is affected by the magnetic flux within the solenoid directly, without any mediating field to transmit its influence. This is doubly embarrassing because the equations governing the electron's motion certainly *look* as if they involve local action — but between ψ and $\mathbf{A}$, not ψ and $\mathbf{B}$.[3]

This suggests a natural alternative(called the "new localized EM properties" view by Healey (2007, p.55)): take the $\mathbf{A}$-field as a physical field. The problem, of course, is gauge invariance: since two gauge-equivalent $\mathbf{A}$-fields (that is, two $\mathbf{A}$-fields related by a gauge transformation) are empirically indistinguishable, how is it to be determined which is the true $\mathbf{A}$ field? This can be thought of as giving rise both to a problem of empirical inaccessibility of the present electromagnetic state (no amount of evidence can tell us which of the various gauge-equivalent $\mathbf{A}$-fields is correct) and a problem of indeterminism (the equations of electromagnetism determine a system's evolution only up to gauge transformations, so if $\Lambda(\mathbf{x}, t) = 0$ for $t < 0$, they fail to tell us whether a given set of $t < 0$ initial conditions will evolve into $\mathbf{A}$ or $\mathbf{A} + \nabla\Lambda$).

It is possible to remove the underdetermination by imposing a particular gauge condition (what Maudlin (1998) calls a "one true gauge" strategy[4]). But given the gauge symmetry, there seem to be few grounds beyond aesthetic preference for selecting one gauge rather than another, and the problems of empirical inaccessibility and indeterminism are replaced by a problem of underdetermination of theory by data. One need not be a crude verificationist to find this level of underdetermination unattractive.

These concerns suggest looking for a gauge-invariant representation of the theory. Our slogan might be: "the physical facts about the fields are represented by the gauge-invariant features of $\mathbf{A}$. One of those gauge-invariant features is $\mathbf{B} = \nabla \times \mathbf{A}$, but the A-B effect shows us that there are others." As stated, this is a mathematical problem: find a complete characterisation of $\mathbf{A}$, up to gauge transformations, in any given region R. And there is a well-known answer: $\mathbf{A}$ is characterised completely and gauge-invariantly by its line integral around every loop in R (called the *holonomies* of the loops).

For future purposes, it will be useful to explain this a little further. Given some functional f from $\mathbf{A}$-fields to some other space, f can be said to characterise the gauge-invariant features of the $\mathbf{A}$-fields provided that $f(\mathbf{A}) = f(\mathbf{A}')$ iff $\mathbf{A}$ and $\mathbf{A}'$ are related by a gauge transformation. To see that this is the case

[3]There is a subtler problem: the problems of interpretation of the vector potential in electromagnetism generalise to so-called 'non-Abelian gauge-theories', but the no new properties view does not generalise readily to these more exotic cases. See Healey (2007, p.84) and references therein for details.

[4]In discussion I have found that Maudlin is often understood as advocating this strategy; my own more minimal reading is that he is simply pointing out that it is possible as part of a case to undermine analogies between the A-B effect and Bell's inequality.

for holonomies, suppose that $\mathbf{A}$ and $\mathbf{A}'$ satisfy

$$\oint_\Gamma \mathbf{A} \cdot \mathrm{d}\mathbf{x} = \oint_\Gamma \mathbf{A}' \cdot \mathrm{d}\mathbf{x} \tag{8}$$

for any loop Γ. Then the integral of $(\mathbf{A} - \mathbf{A}')$ around any closed loop is zero, or put another way, the integral of $(\mathbf{A} - \mathbf{A}')$ between $\mathbf{x}_0$ and $\mathbf{x}$ depends only on $\mathbf{x}_0$ and $\mathbf{x}$ and not on the path connecting them. If we then choose arbitrary $\mathbf{x}_0$ and define

$$\Lambda(\mathbf{x}) = \int_{\mathbf{x}_0}^{\mathbf{x}} (\mathbf{A} - \mathbf{A}') \cdot \mathrm{d}\mathbf{x}, \tag{9}$$

then $\nabla\Lambda = (\mathbf{A} - \mathbf{A}')$ and so $\mathbf{A}, \mathbf{A}'$ are gauge-equivalent. Conversely, if they are gauge-equivalent then (since the integral of $\nabla\Lambda$ around a closed loop always vanishes) they have the same holonomies.

This suggests Healey's own preferred interpretation of the magnetic field's ontology, the "new non-localized EM properties" view: the magnetic field is represented by a map from loops to real numbers. By the definition of the curl, the integral of $\mathbf{A}$ around an infinitesimal loop at point $\mathbf{x}$ is equal to $\mathbf{B} \cdot \mathbf{n}\delta S$, where $\mathbf{n}$ is normal to the surface enclosed by the loop and δS is the area of that surface. So among the components of Healey's ontology (in effect) is the magnetic field. But that ontology is much larger than just the field.

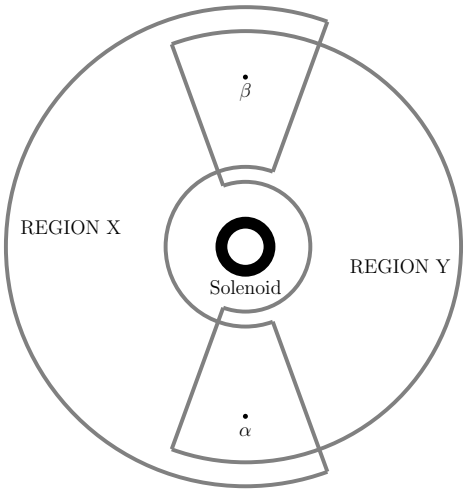
Healey's loop ontology faces three main objections. Firstly, just as with the **B**-field ontology there is no natural way to write the equations of motion of the theory in terms of the loop properties directly; the **A**-field remains indispensable mathematically. Secondly, the ontology is very redundant: loops can be decomposed into smaller loops, and the real number assigned to the larger loop must be the sum of those assigned to its components. (If a region R is simply connected, any loop can be decomposed into infinitesimal loops, and the **B** field of R actually completely determines the values of all the loops in R.) Not only is this awkward, it is difficult to explain naturally *except* by defining the values of each loop as the integral of some vector field around the loop.

Most strikingly, Healey's ontology is non-separable: if X and Y are simply connected spatial regions whose union is not simply connected, then fully specifying the values assigned to each loop in X and Y separately leaves some loops in $X \cup Y$ unspecified. The A-B effect itself offers an illustration: consider X and Y to be as given in diagram 1. Since X and Y are each simply connected, and since in each $\mathbf{B} = 0$, each is magnetically trivial: each loop integral is equal to zero. Insofar as the magnetic field in a region is supposed to be represented by the gauge-invariant facts about X in that region, in both X and Y the magnetic field is the same as in empty space (there is a gauge transformation that transforms it to zero). But the field in $X \cup Y$ is *not* the same as in empty space: the value of loops that enclose the solenoid is non-zero.

So the A-B effect appears to present us with a trilemma. We would like an understanding of electromagnetism that is separable, gauge-invariant, and has no action at a distance. It appears that one of these has to be rejected.

Before going on I should note that while this discussion has been carried out at a relatively elementary level, many proposed ways of understanding the

Figure 1: Regions of space around the solenoid



ontology of electromagnetism in the light of the A-B effect are much more sophisticated, and in particular, involve extensive appeal to the mathematics of fibre bundles[5]. It is perhaps worth making clear that whatever the virtues of these approaches, they cannot avoid the basic trilemma. For the **A**-field in region X is gauge-equivalent to what it would be if the solenoid where absent, and so is the **A**-field in region Y, but the **A**-field in region $X \cup Y$ is not. So any representation of the field that is gauge-invariant must violate either separability (by assigning a nontrivial electromagnetic state to region $X \cup Y$) or local action (by assigning a trivial electromagnetic state to the region in which the electron moves).

Here ends my summary of the A-B effect.

3 The A-B effect and the complex field

The A-B effect arises because of certain features of the mathematical theory of a complex scalar field ψ coupled to a real vector field **A**. It is therefore in

[5]See, for instance, Nounou (2003) or Leeds (1999).

hindsight a little odd that the literature on the A-B effect has been almost wholly concerned with the $\mathbf{A}$ field and hardly at all with the ψ field. In particular, the line of reasoning that leads to the loop ontology — and to the argument that any gauge-invariant representation of the magnetic field is non-separable — is concerned purely with the gauge-invariant features of $\mathbf{A}$ and not with ψ at all. Let us attempt to rectify this.

Prima facie, there are two obvious ideas as to how to think about the gauge-invariant features of ψ:

1. Representing the gauge-invariant features of $\mathbf{A}$ by loop holonomies already takes care of the gauge freedom. Any two complex fields ψ, ψ' can thus be thought of as representing different physical possibilities. The physical states of the theory are thus represented by a complex field and a set of loop holonomies.

2. Since there is a gauge transformation relating any two fields ψ, ψ' satisfying $|\psi(\mathbf{x})| = |\psi'(\mathbf{x})|$, the only gauge-invariant feature of ψ is its magnitude. The physical states of the theory are thus represented by a real field $|\psi|$ and a set of loop holonomies.

Neither is satisfactory, for neither provides a complete characterisation of the gauge-invariant features of the theory. To see why, suppose that $(\psi, \mathbf{A})$ and $(\psi', \mathbf{A}')$ are two possible pairs of fields. A given function f of the fields characterises them completely up to gauge transformations provided that $f(\psi, \mathbf{A}) = f(\psi', \mathbf{A}')$ just if for some Λ, $\psi' = \mathrm{e}^{i\Lambda}\psi$ and $\mathbf{A}' = \mathbf{A} + \nabla\Lambda$.

For the first suggested characterisation, f takes ψ to itself and $\mathbf{A}$ to the loop holonomies. But here the only gauge transformations that leave ψ invariant are those for which $\Lambda(\mathbf{x}) \neq 0$ only when $\psi(\mathbf{x}) = 0$. So in general this representation is not itself gauge invariant. For the second suggestion, f takes ψ to its magnitude and $\mathbf{A}$ to its holonomies, and this clearly *is* gauge invariant. But consider the pairs $(\psi, \mathbf{A})$ and $(\mathrm{e}^{ie\sigma}\psi, \mathbf{A})$, for some arbitrary function σ. These have the same holonomies and the same $|\psi|$. But they are gauge-equivalent only if, for some Λ,

$$\mathrm{e}^{ie\sigma}\psi = \mathrm{e}^{ie\Lambda}\psi \quad \text{and} \quad \mathbf{A} = \mathbf{A} + \nabla\Lambda. \tag{10}$$

This pretty clearly requires (i) Λ to be constant (at least on the connected parts of the region of space we are considering) and (ii) $\Lambda(\mathbf{x}) = \sigma(\mathbf{x}) + 2n\pi/e$ on any connected region where $\psi \neq 0$. In general (that is, for any choice of σ which is not constant on any connected region where $\psi \neq 0$) this cannot be satisfied. So the second suggested characterisation erroneously represents gauge-inequivalent pairs of fields as physically equivalent. (And, in case it's not obvious, these gauge-inequivalent fields are definitely physically inequivalent: two pairs of fields which at time t are gauge-inequivalent but agree on the magnitude of the wavefunction and on the holonomies will not in general so agree at later times, and $|\psi|$ is empirically accessible.)

Our two suggestions share a common flaw. They attempt to characterise the gauge-invariant features of the fields by separately representing the gauge-invariant features of ψ and $\mathbf{A}$. But the gauge transformations act *jointly* on

the two fields, and there are joint features of the pair of fields that are gauge-invariant but do not derive directly from gauge-invariant features of the field considered separately.

In particular, consider the quantity $|\nabla\psi - ei\mathbf{A}\psi|$. This is gauge-invariant — indeed, the fact that it is gauge invariant is the central heuristic of the gauge principle in particle physics[6] — but its gauge invariance does not derive from gauge-invariant features of ψ and $\mathbf{A}$ separately but rather from the cancellation of terms in the gauge transformations of both.

This suggests that a gauge-invariant characterisation of $(\psi, \mathbf{A})$ will need to consider joint features. A helpful way to get at such a characterisation starts by decomposing ψ into its magnitude and phase:

$$\psi(\mathbf{x}, t) = \rho(\mathbf{x}, t) \exp(ie\theta(\mathbf{x}, t)). \tag{11}$$

(This decomposition is unique, up to an overall constant $2n\pi/e$ in θ, provided that $\psi(\mathbf{x}, t)$ is everywhere nonzero; I return to the $\psi = 0$ case later.)

Clearly, ρ is a gauge-invariant feature of ψ alone, and hence of $(\psi, \mathbf{A})$ jointly. More interestingly, consider the gauge-invariant quantity $\psi^*(\nabla - ei\mathbf{A})\psi$. Expressed in terms of ρ and θ, it is

$$\psi^*(\nabla - i\mathbf{A})\psi = \rho\nabla\rho + ie\rho^2(\nabla\theta - \mathbf{A}). \tag{12}$$

Since ρ^2 and $\rho\nabla\rho$ are gauge-invariant, so is $\mathcal{D}\theta \equiv \nabla\theta - \mathbf{A}$, the gauge-covariant derivative of θ (something that can also be verified directly).

So: we now have two gauge-invariant features of the theory: the scalar field $\rho = |\psi|^2$, and the vector field $\mathcal{D}\theta$. In fact, no others are needed. For suppose that $\psi' = \rho' \mathrm{e}^{ie\theta'}$ and $\mathbf{A}'$ satisfy

$$\rho' = \rho \quad \text{and} \quad \nabla\theta' - \mathbf{A}' = \nabla\theta - \mathbf{A}. \tag{13}$$

Then it is easy to verify that

$$\psi' = \psi \mathrm{e}^{ie(\theta' - \theta)} \quad \text{and} \quad \mathbf{A}' = \mathbf{A} + \nabla(\theta' - \theta). \tag{14}$$

In other words, $(\theta' - \theta)$ defines a gauge transformation from $(\psi, \mathbf{A})$ to $(\psi', \mathbf{A}')$. In particular, the holonomies can be recovered from the covariant derivatives of the phase:

$$\oint \mathcal{D}\theta \cdot \mathrm{d}\mathbf{x} = \oint \nabla\theta \cdot \mathrm{d}\mathbf{x} + \oint \mathbf{A} \cdot \mathrm{d}\mathbf{x} = \oint \mathbf{A} \cdot \mathrm{d}\mathbf{x}\,, \tag{15}$$

since the integral of a gradient around a closed loop is zero.

The alert reader will have noticed something rather striking about this representation. Both ρ and $\mathcal{D}\theta$ are *local* features of the theory: their values at a point $\mathbf{x}$ depend only on ψ and $\mathbf{A}$. The $\mathbf{A}$-field alone may admit of no description which is both separable and gauge-invariant, but the ψ and $\mathbf{A}$ fields jointly admit of both.

[6] Slightly more accurately, the central heuristic is that $(\nabla\psi - ei\mathbf{A}\psi)$ transforms under the gauge group in the same way as does ψ itself.

Indeed, we can rewrite the Schödinger equation in a local and gauge-invariant way in terms of these quantites; since the method of doing so is instructive for later, I spell it out here. Firstly, let us make a choice of gauge: the *unitary gauge*, in which ψ is always real. (This may seem unfamiliar: gauge conditions are usually specified via a constraint on $\mathbf{A}$ rather than ψ. But mathematically a gauge condition is just a condition which picks a unique element out of each equivalence class of gauge-equivalent fields, and — again on the assumption that $\psi \neq 0$ — the unitary gauge does that just fine. I return to its conceptual significance later.)

In the unitary gauge we can write $\psi = \rho$; the Schrödinger equation becomes

$$\frac{1}{2m}\left(\nabla^2\rho - \mathbf{A}\cdot\mathbf{A}\rho - 2i\mathbf{A}\cdot\nabla\rho - i(\nabla\cdot\mathbf{A})\rho\right) = i\dot{\rho}. \tag{16}$$

Separating real and imaginary parts, we get

$$(\nabla^2 - \mathbf{A}\cdot\mathbf{A})\rho = 0; \tag{17}$$

$$2\mathbf{A}\cdot\nabla\rho + (\nabla\cdot\mathbf{A})\rho = 2m\dot{\rho}. \tag{18}$$

Combined with the condition that the magnetic field strength $\mathbf{B}$ vanishes,

$$\nabla\times\mathbf{A} = 0, \tag{19}$$

this is a complete and deterministic set of equations for ρ and $\mathbf{A}$ in the unitary gauge.

(If you are wondering how the Schrödinger equation, which is supposed to determine the evolution of the *particle*, has given rise to a joint equation for the particle probability density and the vector potential, recall that in the unitary gauge, phase information about the particle is carried by $\mathbf{A}$. If this makes you start to worry that we don't have a clean separation any more between matter and magnetic degrees of freedom, hold that thought!)

To get a gauge-invariant set of equations, we just note that in the unitary gauge, $\nabla\theta = 0$ and so $\mathcal{D}\theta = \mathbf{A}$. So in this gauge, we can replace $\mathbf{A}$ with $\mathcal{D}\theta$ to get

$$(\nabla^2 - (\mathcal{D}\theta)^2)\rho = 0;\ 2\mathcal{D}\theta\cdot\nabla\rho + (\nabla\mathcal{D}\theta)\rho = 2m\dot{\rho};\ \nabla\times\mathcal{D}\theta = 0. \tag{20}$$

But this equation, being expressed entirely in terms of gauge-invariant quantities, does not depend on the unitary gauge. We have obtained a set of local, deterministic, gauge-invariant differential equations for the A-B effect.

All this ought to suggest that the apparent nonlocal-action/ gauge-dependence/ non-separability trilemma of the A-B effect is just an artefact of our failure to consider ψ as well as $\mathbf{A}$. Indeed, I think this suggestion is correct. Before exploring the suggestion further, though, it will be helpful to get clear just how that trilemma arises and how the introduction of matter blocks it.

4 Origins of non-separability

Recall the definition of non-separability: the state of a region of space $X \cup Y$ is nonseparable if specification of all properties of regions X and Y separately does not completely specify the properties of $X \cup Y$. In the case of electromagnetic gauge theory under the assumption that all physical properties are gauge-invariant, the properties of a region are supposed to be in some way represented by gauge-invariant features of the fields, with two regions having the same physical properties iff the fields on those regions are gauge-equivalent.

We can now express the presence or absence of non-separability mathematically: fields $\psi, \mathbf{A}$ defined on $X \cup Y$ give rise to non-separability iff there exist other fields $\psi', \mathbf{A}'$ defined on $X \cup Y$ such that

(i) $\psi'|_X, \mathbf{A}'|_X$ (the restrictions of ψ' and $\mathbf{A}'$ to X) are gauge-equivalent to $\psi|_X, \mathbf{A}|_X$;

(ii) likewise $\psi'|_Y, \mathbf{A}'|_Y$ and $\psi|_Y, \mathbf{A}|_Y$ are gauge-equivalent; but

(iii) $\psi', \mathbf{A}'$ and $\psi, \mathbf{A}$ are *not* gauge-equivalent.

For any possible state of $X \cup Y$ must be represented by some pair of fields on $X \cup Y$, and non-separability is the possibility of two such non-gauge-equivalent pairs $\psi, \mathbf{A}$ and $\psi', \mathbf{A}'$ whose restrictions to X and to Y are gauge-equivalent.

Suppose (i) and (ii) are the case. Then there exist functions Λ_X, Λ_Y on X and Y respectively such that

1. On X, $\psi' = \mathrm{e}^{ie\Lambda_X}\psi$ and $\mathbf{A}' = \mathbf{A} + \nabla\Lambda_X$.
2. On Y, $\psi' = \mathrm{e}^{ie\Lambda_Y}\psi$ and $\mathbf{A}' = \mathbf{A} + \nabla\Lambda_Y$.

It follows that on the intersection region $X \cap Y$,

1. $\mathrm{e}^{ie(\Lambda_X - \Lambda_Y)}\psi = \psi$;
2. $\nabla(\Lambda_X - \Lambda_Y) = 0$.

So $\Lambda_X - \Lambda_Y$ is a real function on $X \cap Y$ which (1) is equal to zero except where $\psi = 0$ and (2) has vanishing gradient everywhere. These are strict conditions. The first can be satisfied by $\Lambda_X - \Lambda_Y \neq 2n\pi/e$ only in regions where $\psi = 0$. The second entails that if x and y are points in $X \cap Y$ connected by a path lying within $X \cap Y$, then $(\Lambda_X - \Lambda_Y)(x) = (\Lambda_X - \Lambda_Y)(y)$. Jointly, then, the conditions can be satisfied by a function with non-vanishing gradient only if $X \cap Y$ is path-disconnected (if there are regions of $X \cap Y$ that cannot be joined by any path lying within $X \cap Y$) and if ψ is zero on at least one of the connected components.

If these conditions are not satisfied, then Λ_X and Λ_Y agree (up to a removable $2n\pi/e$ term) on the intersection of X and Y. We can then defined a single function Λ consistently by declaring it equal to Λ_X on X and to Λ_Y on Y; this function generates a gauge transformation between $\psi, \mathbf{A}$ and $\psi', \mathbf{A}'$, so that (iii) is not satisfied.

Conversely, if they *are* satisfied then we can choose arbitrary functions Λ_X, Λ_Y which are constant on each connected component of $X \cap Y$ but which are not equal to each other on at least one such component. The fields obtained by applying a gauge transformation generated by Λ_X to the restriction of $\psi, \mathbf{A}$ to X, and likewise for Y, agree on $X \cap Y$ and so can be consistently combined into a pair of fields on $X \cup Y$, but they are not gauge-equivalent.

So we have found a necessary and sufficient condition for non-separability in gauge theory: it can occur with respect to regions X, Y when their intersection is disconnected and when the matter field vanishes on at least one connected component. (In fact, the result generalises straightforwardly to more general gauge theories: what is required there is not per se that ψ vanishes on a connected component but that there is some element of the gauge group g such that $g\psi = \psi$ on that region. This generally requires ψ to remain strictly confined to some small subspace of the internal vector space.)

The first of these conditions is purely topological. A necessary (though not sufficient) condition for it to occur is that $X \cup Y$ is not simply connected;[7] note that this is satisfied by the region outside the solenoid in the A-B effect, and recall that we have seen that non-separability occurs in the loop ontology only where non-simply-connected regions are considered.

The second condition, however — the vanishing of ψ on an open set — is implausibly, indeed unphysically, stringent. Notice that there is no 'give' in the condition at all: even if $|\psi| = 10^{-1000}$, there is no prospect of non-separability. (The local facts about X and Y separately that determine the joint properties of $X \cup Y$ might be extremely difficult to ascertain, but that is a limit of practice, not principle.) In one-particle quantum mechanics, it is a theorem[8] that ψ is never exactly zero on an open set in space*time*, so that the condition can hold, if at all, only for an instant. And in quantum field theory the most perspicuous way (in this context) to think of the system is as a superposition of different field configurations, in which the weight given to the configuration where ψ is *exactly* zero will itself be exactly zero. (I consider the quantum-field-theoretic case more carefully in section 7). I conclude that we can set aside this case. Once set aside, there is no obstacle to a fully local, but fully gauge-invariant, understanding of the theory.

[7] Proof sketch: suppose $X \cup Y$ is simply connected and let f be any smooth function which is constant on each connected component of $X \cap Y$. Then for arbitrary a, b, there is a well-defined vector field v on $X \cup Y$ such that $v|_X = a\nabla f$ and $v|_Y = b\nabla f$. For arbitrary $p, q \in X \cap Y$, let γ_X and γ_Y be paths in X and Y respectively from p to q. Then the integral of v along the loop from x to y along γ_X and back along γ_Y is $(a-b)(f(q)-f(p))$. But since $\nabla \times v = 0$, by Stokes' theorem this integral must vanish. So $f(p) = f(q)$, i. e. any function constant on the connected components is constant.

[8] The result is proved under rather general conditions by Hegerfeldt (1998a, 1998b); see also the discussion in Halvorson and Clifton (2002). To see intuitively why it is correct, just notice that to confine a particle exactly to a finite region requires it to have arbitrarily high-momentum Fourier components, corresponding to arbitrarily high momenta, and so to components of the wavefunction that will spread out at arbitrarily high speeds.

5 The interlinking of A and ψ

I have shown formally that the gauge-invariant features of ψ and $\mathbf{A}$ can generically be jointly represented in a fully local (i. e. , non-separable) way. But it is still reasonable to ask what those gauge-invariant features are actually supposed to represent: that is, what kind of ontology is compatible with the theory?

It is tempting to think that the question can be innocently rephrased as: what kind of ontologies for the electromagnetic field, and for the matter field, are compatible with the theory? Tempting, but mistaken — and this is one of the main points of the paper. For since the gauge transformation thoroughly mixes the two together, there is simply no justification — as long as we wish our ontology to depend only on gauge-independent features of the theory — in regarding the two mathematically-defined fields as representing two *separate* but interacting entities, rather than as (somewhat redundantly) representing aspects of a *single* entity.

To press the point, let us consider again the question of a choice of gauge. Most gauge choices encountered in electromagnetism impose a constraint on the $\mathbf{A}$-field, and leave the ψ-field unconstrained: thus the Coulomb gauge, $\nabla \cdot \mathbf{A} = 0$, for instance, or the London gauge $\mathbf{A}_z = 0$, each place one constraint on $\mathbf{A}$ per point of space. Hence the temptation to see the $\mathbf{A}$-field, with its apparent three degrees of freedom per space point, as really having two once gauge redundancy is allowed for, and likewise to see the ψ field as genuinely having two degrees of freedom per space point.

But this is pure convention. Consider again the unitary gauge, in which we require that the phase of ψ vanishes (i. e. , that ψ is real). In this gauge, ψ has only one degree of freedom, but there is no residual gauge invariance of $\mathbf{A}$ — each of its three apparent degrees of freedom are physical. So do we have one degree of freedom for matter and three for electromagnetism, or two for each? The question is only meaningful if we persist in supposing that two distinct entities are present.

To be sure, from the perspective of quantum field theory there is no conventionality about the *particles* that are associated with the fields: whatever gauge we choose, we will discover a particle spectrum consisting of a massless vector boson (two degrees of freedom) and a charged scalar boson (one degrees of freedom, but with both matter and antimatter versions[9]). But the particle spectrum of a theory represents the expansion of the theory's Hamiltonian in normal modes around a (possibly local) minimum of energy, and is by its nature holistic: the particle spectrum of the theory is a dynamical and not a metaphysical matter, and should not be thought to require the existence of metaphysically distinct matter and electromagnetic fields.

Indeed, it need not always be the case that a complex-scalar-field-plus-vector-potential field theory even has that particular particle spectrum. If the gauge symmetry is spontaneously broken (that is, if the minimum-energy configuration has a non-zero expected value of $|\psi|$) then the particle spectrum consists

[9]For more on the curious way in which complex classical degrees of freedom give rise to antimatter, see Wallace (2009) and Baker and Halvorson (2009).

instead of a massive vector field and a real scalar field (indeed, this is one of the main applications of the unitary gauge). In popularisations of the Higgs mechanism, this phenomenon is sometimes described as the electromagnetic field "eating" one of the degrees of freedom of the scalar field and thus gaining mass, a metaphor that has been sharply criticised by Earman (2003) (see also Struyve (2011)). But once we realise that the electromagnetic and scalar fields cannot be thought of as separate entities, there need be no residual surprise that the normal-mode expansion of the physical system that they jointly describe is best analysed in different ways in different regimes.

But how are we to think about this "jointly described" entity? We know that it can be characterised entirely by the magnitude of ψ (a scalar field) and by its covariant derivative (a vector field, or more precisely a one-form field). It is important to remember that these are conceptually and mathematically very different entities. A scalar field, mathematically, is just an assignment of a real number to every point of space, and can easily enough be thought of as ascribing properties to *points* of space. A one-form field is not so simple and cannot be so represented: to speak loosely, it is more like an assignation of properties to infinitesimally small diffences between points of space. Or put another way, if a vector is thought of loosely as an infinitesimal arrow from one space point to a neighboring one, a one-form field assigns a real number to each such infinitesimal arrow. A one-form is then something more like a set of relations between (infinitesimally close) points of space.

That suggests that there are indeed two components of the ontology of the system: a collection of properties of points of space, and a collection of relations between infinitesimally close points of space. In certain circumstances (mathematically, when the holonomy vanishes) integrating the infinitesimal relations from x to y along a given path gives a result which is in fact independent of the path; in these situations we can consistently define a relation between those finitely-separated points and call it the *phase difference*, and then the system can be represented by a complex field with no remaining redundancy save for a single choice of phase. Conversely, the holonomy — the integral of the infinitesimal relations around a closed loop — provides a measure of the extent to which this representation of the systems is blocked, and the holonomy in turn is *mostly* determined by the integral of the relations around infinitesimal closed loops — the curvature.

The extent to which this somewhat loose talk of 'infinitesimal relations' can be made more precise lies beyond the scope of this paper; it is perhaps worth remembering, though, that in any case the empirical success of (classical or quantum) electrodynamics provides no licence whatever to regard the theory as a reliable description of the physical world on *arbitrarily short* lengthscales, so that thinking about the relations between extremely but finitely close points of space may actually be a more reliable way of approaching the theory's ontology than appeal to vector bundles or to actual infinitesimals.[10]

[10] For more consideration of the metaphysics of vector fields, see Butterfield (2006b, 2006a) and references therein.

Nonlocality and the Aharonov-Bohm Effect*

Richard Healey†‡
Department of Philosophy, University of Arizona

At first sight the Aharonov-Bohm effect appears nonlocal, though not in the way EPR/Bell correlations are generally acknowledged to be nonlocal. This paper applies an analysis of nonlocality to the Aharonov-Bohm effect to show that its peculiarities may be blamed either on a failure of a principle of local action or on a failure of a principle of separability. Different interpretations of quantum mechanics disagree on how blame should be allocated. The parallel between the Aharonov-Bohm effect and violations of Bell inequalities turns out to be so close that a balanced assessment of the nature and significance of quantum nonlocality requires a detailed study of both effects.

1. Introduction. Aharonov and Bohm (1959) drew attention to the quantum mechanical prediction that an interference pattern due to a beam of charged particles could be produced or altered by the presence of a constant magnetic field in a region from which the particles were excluded. This effect was first experimentally detected by Chambers (1960), and since then has been repeatedly and more convincingly demonstrated in a series of experiments including the elegant experiments of Tonomura et al. (1986).

At first sight, the Aharonov-Bohm effect seems to manifest nonlocality. It seems clear that the (electro)magnetic field acts on the particles since it affects the interference pattern they produce, and this must be action at a distance since the particles pass through a region from which that field is absent. There have been numerous attempts to avoid this

*Received December 1995.

†Send requests for reprints to the author, Department of Philosophy, University of Arizona, 213 Social Sciences Bldg. #27, Tucson, AZ 85721.

‡I wish to acknowledge the support of the National Science Foundation, under whose grant SBER94-22185 this work was completed. Thanks also to a number of colleagues, especially Paul Teller and one other anonymous reviewer, for encouragement and advice.

Philosophy of Science, 64 (March 1997) pp. 18–41. 0031-8248/97/6401-0002$2.00

conclusion. But despite the fact that no interpretation has succeeded in portraying the Aharonov-Bohm effect as completely local, we have much to learn from these attempts. For different interpretations of the Aharonov-Bohm effect portray it as nonlocal in different senses. By examining these interpretations we can gain a fresh perspective on the nature of quantum nonlocality.

Such a perspective is sorely needed. The intense scrutiny of EPR-type correlated systems and the associated violations of Bell inequalities has produced a kind of tunnel vision that has made it hard to achieve a balanced assessment of the nature and significance of quantum nonlocality. There is, for example, a widespread belief that quantum nonlocality is manifested only by compound systems in "entangled" states. The nonlocality of the Aharonov-Bohm effect refutes that belief. More importantly, while it is necessary to distinguish a number of different senses in which the Aharonov-Bohm effect may be judged not to be local, the central senses of locality (conformity to principles of local action and separability) are just those that also help to define what is most at stake when it comes to violations of Bell inequalities.

2. The Aharonov-Bohm Effect. As noted by Aharonov and Bohm (1959), quantum mechanics predicts that the interference pattern produced by a beam of charged particles may be altered by the presence of a constant magnetic field, even though that field is confined to a region from which the particles are excluded.[1] This has since been confirmed experimentally.[2] A simple example of the effect is depicted in Figure 1.

If no current flows through the solenoid behind the two slits, then the familiar two-slit interference pattern will be detected on the screen. But if a current passes through the solenoid, generating a constant magnetic field ***B*** *confined to its interior* in the z-direction parallel to the two slits, the whole two-slit interference pattern is shifted by an amount

$$\Delta x = \frac{l\lambda}{2\pi d}\frac{e}{\hbar}\Phi \tag{1}$$

1. The effect had been noted previously by Ehrenberg and Siday (1949), but it was Aharonov and Bohm's work that brought it into prominence in the literature. Lorentz covariance implies the existence of a corresponding effect involving electric rather than magnetic fields, which is harder to investigate experimentally and raises no new issues for nonlocality.

2. The first experimental confirmation by Chambers (1960) has more recently been duplicated much more convincingly. For a recent review, see Peshkin and Tonomura 1989.

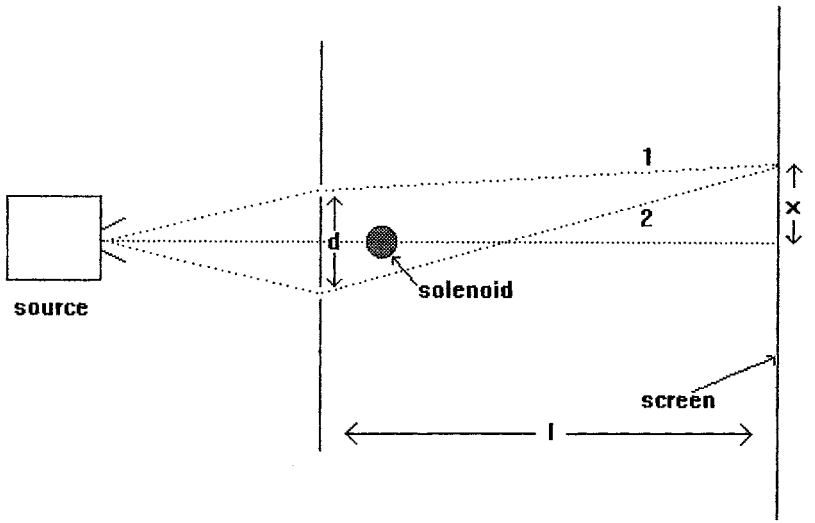


Figure 1.

where d is the slit separation, l is the distance to the screen, λ is the de Broglie wavelength of the electrons in the beam, and Φ is the magnetic flux through the solenoid.

How is this phenomenon to be explained? At first sight, it appears that the magnetic field inside the solenoid must have some kind of nonlocal effect on the electrons, since $\boldsymbol{B}$ is zero everywhere outside the solenoid, in the region through which they must pass on their way from the slits to the screen. But Aharonov and Bohm (1959) denied that the effect was nonlocal, claiming instead that it arose from a purely local interaction with the magnetic vector potential $\boldsymbol{A}$ (or more generally the electromagnetic potential A^{μ}). They concluded that while in classical mechanics this potential could be regarded as just a mathematical device for conveniently representing the physically real (electro)magnetic field, quantum mechanics shows that it is itself a physically real field. This view was endorsed and widely promulgated by Feynman in his famous Lectures (Volume 2).

Aharonov and Bohm (1959) first presented the effect as a theoretical consequence of quantum mechanics prior to any experimental demonstration. They derived this consequence by solving the Schrödinger equation for scattering of an electron beam by an infinitely long and infinitely thin solenoid. A simplified QM derivation for the setup pictured in Fig. 1 is as follows. Consider two paths by which an electron might arrive at the same point on the screen, one passing through the

upper slit, one through the lower slit. If the difference in path lengths is a, then there will be a corresponding phase difference δ given by

$$\delta = 2\pi a/\lambda \tag{2}$$

where λ is the electrons' de Broglie wavelength. For x much less than l,

$$a \approx xd/l, \text{ and so } \delta \approx 2\pi xd/\lambda l. \tag{3}$$

If no current passes through the solenoid, then we have the ordinary two-slit interference experiment. The condition for constructive interference between these paths is $\delta = 2n\pi$, and so an interference maximum will appear on the screen at a distance $x \approx n\lambda l/d$ from the axis of symmetry, for each value of $n = 0,1,2$, etc.

Passing a constant current through the solenoid produces a magnetic field inside it (directed towards you) and a magnetic vector potential $\boldsymbol{A}$ both inside and outside. This produces an additional phase difference of $-e/\hbar \boldsymbol{A}.d\boldsymbol{r}$ in the electrons' wave function between point $\boldsymbol{r}$ and point $\boldsymbol{r} + d\boldsymbol{r}$ (assuming the electron's charge is $-e$). The total additional phase change over a path is then

$$\delta = -\frac{e}{\hbar}\int \boldsymbol{A}.d\boldsymbol{r} \tag{4}$$

This will introduce an additional phase difference between two paths from source to screen of

$$\Delta(\delta) \equiv \delta_1 - \delta_2 = \left(-\frac{e}{\hbar}\int_1 \boldsymbol{A}.d\boldsymbol{r}\right) - \left(-\frac{e}{\hbar}\int_2 \boldsymbol{A}.d\boldsymbol{r}\right) \tag{5}$$

Now if the solenoid is close to the slits and very small, then the direct path from source to screen through the top slit will go around the top of the solenoid, and the direct path from source to screen through the bottom slit will go around the bottom of the solenoid. Hence the additional phase difference between such paths will be given by

$$\Delta(\delta) = -\frac{e}{\hbar}\oint \boldsymbol{A}.d\boldsymbol{r} \tag{6}$$

where the integral is now taken around the closed curve formed by tracing a path from source to screen via the upper slit, and then returning from screen to source via the lower slit—a path that encloses the solenoid. It follows that

$$\Delta(\delta) = -\frac{e}{\hbar}\oint \boldsymbol{A}.d\boldsymbol{r} = -\frac{e}{\hbar}\int curl\boldsymbol{A}.d\boldsymbol{S} = \frac{e}{\hbar}\Phi \tag{7}$$

This additional phase difference is independent of x, and so the entire interference pattern is shifted upward by the same amount, namely

$$\Delta x = \frac{l\lambda}{2\pi d}\Delta(\delta) = \frac{l\lambda}{2\pi d}\frac{e}{\hbar}\Phi \tag{8}$$

On this account, the shift in the interference pattern is produced by a direct local interaction between electrons and the magnetic vector potential $\boldsymbol{A}$ outside the solenoid. $\boldsymbol{A}$ is no longer zero either inside *or outside* the solenoid because of the current flowing through it, even though $\boldsymbol{B} = \text{curl}\boldsymbol{A} = 0$ everywhere outside the solenoid. However, this QM derivation shows that the A-B effect is local only if $\boldsymbol{A}$ is a physically real field, capable of acting on the electrons directly. But there is reason to doubt that the magnetic vector potential is a physically real field, since $\boldsymbol{A}$ is not gauge-invariant, unlike the magnetic field $\boldsymbol{B}$ and the phase-shift $\Delta(\delta)$. That is to say, both $\boldsymbol{A}$ and

$$\boldsymbol{A}' = \boldsymbol{A} + \nabla\chi \tag{9}$$

are to be regarded as specifying the same physical state, for an arbitrary (but suitably differentiable) function χ. If one nevertheless maintains that in some way A represents a physically real field, the following argument appears to establish that its gauge-dependence excludes a local account of the A-B effect.

With no current flowing, $\boldsymbol{A}$ is zero everywhere outside the solenoid; or more precisely, there exists a function χ such that $\boldsymbol{A}$ can be set to zero everywhere outside the solenoid by the transformation $\boldsymbol{A} \rightarrow \boldsymbol{A}'$ defined by Eq. 9. But even *with* a current flowing, this transformation permits one to set $\boldsymbol{A}$ equal to zero over a very wide region outside the solenoid! By a suitable choice of χ one can, for example, set $\boldsymbol{A}$ equal to zero outside the solenoid everywhere except within a segment, of arbitrarily small angle α, of a solid cylinder of infinite thickness whose inner radius coincides with the outside of the solenoid. One could thus set $\boldsymbol{A}$ equal to zero everywhere along path 1, or everywhere along path 2 (but not both at once).[3] Now for the shift in the interference pattern to be produced by a direct local interaction between each individual electron and the magnetic vector potential $\boldsymbol{A}$ outside the solenoid, that interaction would have to be different when a current is passing through the solenoid. However, the potential is defined only up to a gauge-transformation, and for any continuous path from source to

3. Indeed, if one generalizes the concept of a choice of gauge along the lines of Wu and Yang 1975, it is even possible to choose a global gauge according to which $\boldsymbol{A}$ is zero *everywhere* outside the solenoid (though this will not be a global gauge in which there is a single value of $\boldsymbol{A}$ in each region within the solenoid).

screen that does not enclose the solenoid there is a gauge-transformation that equates the value of A at every point on that path when a current is flowing to its value when no current is flowing. The shift in the interference pattern cannot therefore be produced by a direct local interaction between individual electrons following such continuous paths and the magnetic vector potential A outside the solenoid. Thus accepting the physical reality of the vector potential fails to render the AB effect local: while denying its physical reality leaves one without any local explanation of the effect.

3. Locality. To see more clearly what is at stake in the A-B effect, we need a better grasp on the notion of locality. Although various explications of locality have been offered by those investigating violations of Bell inequalities, many of these seem inapplicable in the present context. I think the right way to view an explication like Bell's (1964) original "locality" condition is as a purported consequence of a general conception of locality of wider applicability. Einstein formulated just such a conception as follows.

> . . . it appears to be essential for [the] arrangement of the things introduced in physics that, at a specific time, these things claim an existence independent of one another, insofar as these things 'lie in different parts of space'. . . .
>
> Field theory has carried this principle to the extreme, in that it localizes within infinitely small (four-dimensional) space elements the elementary things existing independently of one another that it takes as basic, as well as the elementary laws it postulates for them. For the relative independence of spatially distant things (A and B), this idea is characteristic: an external influence on A has no immediate effect on B; this is known as the 'principle of local action', which is applied consistently only in field theory.
>
> (1948, 322–323)

As has now been widely recognized,[4] one can find two distinct ideas in this and similar passages from Einstein's writings. I shall call these the principle of *Local Action* and the principle of *Separability,* and state them as follows.

> *Local Action*
> If A and B are spatially distant things, then an external influence on A has no immediate effect on B.

4. See, for example, Howard 1985, Redhead 1987, Healey 1991, 1994.

Separability
Any physical process occurring in spacetime region R is supervenient upon an assignment of qualitative intrinsic physical properties at spacetime points in R.

I explain and defend my formulation of the principle of *Separability* in the next section. The remainder of this section focuses on the principle of *Local Action.*

The idea behind *Local Action* is that if an external influence on A is to have any effect on B, that effect must *propagate* from A to B via some continuous physical process. Any such mediation between A and B must occur via some (invariantly) temporally ordered and continuous sequence of stages of this process. Non-relativistically, such mediation could not be instantaneous, and so an effect on B could not occur at the same time as the external influence on A. Thus although in the non-relativistic case the term 'immediate' in *Local Action* may be read as ambiguous between 'unmediated' and 'instantaneous', that ambiguity seems relatively harmless, in so far as any instantaneous effect would have to be unmediated.

Applied to the Bell case, *Local Action* entails that a measurement on a particle A in one wing of an Aspect-type device has no immediate effect either on a particle B on which a measurement is performed in a different wing of the device, or on the apparatus which performs that measurement. Given *Local Action,* a measurement on particle A in one wing of an Aspect-type device can affect either particle B or the apparatus which performs a measurement on it only if some continuous process mediates that effect. But the experimental conditions are designed precisely so as to rule out the possibility that any process could mediate between the two measurement events.[5] This supports the conclusion that the results of the two measurements are causally independent. It is this consequent condition of causal independence that is taken (explicitly or implicitly) to justify more specific "locality" conditions appealed to in derivations of Bell-type inequalities.

Applied now to the Aharonov-Bohm case, *Local Action* entails that a change in the current passing through the solenoid has no immediate effect on the behavior of any electron outside the solenoid. Here the force of the term 'immediate' is to require that any effect of the field inside the solenoid on the behavior of electrons outside the solenoid be

5. In fact, these conditions at most exclude the possibility of mediation via a separable process (cf. Section 4). The measurement events may still be connected by a nonseparable process (as in Healey 1994) in which case the question of causal dependence is reopened. Healey (1992) argues that our concept of causation may then not be sufficiently univocal to permit this question to be decisively answered.

mediated by some influence which acts directly where these electrons are—somewhere outside the solenoid.

Now in some of Aspect's experiments revealing violations of Bell inequalities, the measurement events occurred at spacelike separation. And in these circumstances it is common to justify specific "locality" conditions that figure essentially in derivations of Bell inequalities by appeal to a principle of relativistic locality, to the effect that that there can be no direct causal connection between spacelike separated events. This principle is both logically independent of *Local Action* and not in question in the Aharonov-Bohm effect. Taking *Relativistic Locality* to offer at least as legitimate an explication of the root notion of locality as *Local Action,* one might conclude that experimental violations of Bell inequalities have different, and indeed more significant, implications for locality than experimental demonstrations of the Aharonov-Bohm effect.

There are two reasons for rejecting this conclusion. First, since the condition of relativistic locality cannot be applied in Aspect-type experiments when the measurement events are timelike (or null) separated, an independent justification for excluding the possibility of direct causal connections would have to be given in those cases. Such a justification would naturally appeal to the properties of all known interactions, importantly including the fact that these conform to the principle of *Local Action,* and it would apply equally to the case in which the measurement events are spacelike separated. Hence the principle of *Local Action* at least figures in an important independent line of argument against the possibility of direct causal connections between measurement events in the Bell case. Secondly, it is not at all clear that relativistic locality is in fact a direct consequence of relativity theory. Relativity theory is naturally taken to be a theory governing the structure of spacetime and the sorts of physical processes that can occur within it. As such, it contains no explicit reference to causal notions (despite the common, but potentially misleading, practice of employing causal terminology, as in "the causal structure of spacetime"). Now there are arguments seeking to derive a contradiction from supposed violations of relativistic locality—so called causal paradoxes. But these all import explicitly causal assumptions from outside relativity theory itself, frequently concerning our ability to set up or control various devices and/or physical processes. And these causal assumptions are themselves open to question in situations such as those involved in Aspect-type experiments. Indeed, as I have argued elsewhere (1994), on some models of these experiments there is a coherent conception of causation in accordance with which there is indeed a direct causal connection between spacelike separated events in the relevant Aspect-type

experiments, while the principle of *Local Action* holds, insofar as this connection is mediated by a continuous (albeit nonseparable) causal process!

4. Separability. The previous section offered a formulation of a principle of *Separability* which requires some explanation, especially because it differs from formulations offered by other authors.

On one common understanding, any "entangled" quantum systems are nonseparable in so far as they must be described quantum mechanically by a tensor-product state-vector which does not factorize into a vector in the Hilbert space of each individual system

$$\Psi_{12\ldots n} \neq \Psi_1 \otimes \Psi_2 \otimes \ldots \Psi_n \tag{10}$$

This is related to a more general understanding of nonseparability, according to which any two spatially separated systems possess their own separate real states. For if the state vector of a quantum system gives its real state, then any spatially separated quantum systems described by the "entangled" state vector of Eq. 10 will count as nonseparable, on this more general understanding.[6] But even if spatially separated systems do possess their own separate real states, the system they compose may still fail to be separable if its real state does not supervene on theirs. This suggests the following formulation of separability:

> *Spatial Separability*
> The qualitative intrinsic physical properties of a compound system are supervenient on those of its spatially separated component systems together with the spatial relations among these component systems.

Here I take the real state of a system to be given by its qualitative intrinsic physical properties.

Now while the condition of *Spatial Separability* is naturally applied to an n-particle system that figures in an Aspect-type experiment, it is less clear how it is relevant to an Aharonov-Bohm experiment. But, as shown in Healey (1991), *Spatial Separability* is itself a consequence of a yet more general principle (there called *Spatiotemporal Separability*) which is immediately applicable to the Aharonov-Bohm case. Here is a condensed statement of that principle:

6. Note that this conclusion would not follow if, like Einstein, one were to *deny* that a quantum system's real state is given by its state vector. Indeed, as Howard (1985) and others have pointed out, a number of Einstein's reasons for this denial assumed that the real state (unlike the quantum state) must be separable in the sense just noted.

Separability
Any physical process occurring in spacetime region R is supervenient upon an assignment of qualitative intrinsic physical properties at spacetime points in R.

According to this principle, whether a process is nonseparable depends on what qualitative, intrinsic properties there are. Deciding this involves both conceptual and scientific difficulties. The conceptual difficulty is to say just what it means for a property to be qualitative and intrinsic.

Intuitively, a property of an object is *intrinsic* just in case the object has that property in and of itself, and without regard to any other thing. This contrasts with *extrinsic* properties, which an object has in virtue of its relations, or lack of relations, to other things. Jupiter is intrinsically a body of mass 1.899×10^{27} kilograms, but only extrinsically the heaviest planet in the solar system.[7] Unfortunately, philosophers have been unable to agree on any precise definition of the distinction between intrinsic and extrinsic properties, or even on whether such a definition can or should be given.[8] This is true also of the distinction between qualitative and individual properties, where a property is *qualitative* (as opposed to *individual*) if it does not depend on the existence of any particular individual. Having a mass of 1.899×10^{27} kilograms is a qualitative property of Jupiter, while both the property of being Jupiter and the property of orbiting our sun are individual properties of Jupiter.[9]

After such an inconclusive resolution of the conceptual difficulty, it may seem premature to consider the scientific difficulty of discovering

7. Note that I follow philosophers' usage rather than physicists' here. I take Jupiter's mass to be intrinsic to it even though Jupiter's mass may vary, or indeed might have always been different, from 1.899×10^{27} kilograms. Physicists tend to use the term 'intrinsic' differently, to refer only to unchanging, or even essential, properties (where an essential property is one which an object could not have lacked while remaining that very object).

8. David Lewis (1986a, 61–69), for example, tentatively offers two possible definitions but argues that the distinction is both possible and necessary even if it cannot be defined in terms of anything more basic.

9. Note that while the latter individual property is also an extrinsic property of Jupiter, the former appears to be an intrinsic property. But both would count as extrinsic on a slight broadening of the notion of an extrinsic property. For one might argue that the property of *being Jupiter* should after all be counted as an extrinsic property of that planet, in so far as Jupiter has that property purely by virtue of being related *to itself* in a particular way (namely, through the identity relation). If such a broadening is accepted, then it may turn out that all individual properties are extrinsic, in which case to speak of an intrinsic property as qualitative would be redundant.

what qualitative, intrinsic properties there in fact are. But this is not so. Whatever a qualitative, intrinsic property is *in general,* it seems clear that science, and in particular physics, is very much in the business of finding just such properties.

Physics proceeds by first analyzing the phenomena with which it deals into various kinds of systems, and then ascribing states to such systems. To classify an object as a certain kind of physical system is to ascribe to it certain, relatively stable, qualitative intrinsic properties: and to further specify the state of a physical system is to ascribe to it additional, more transitory, qualitative intrinsic properties. Fundamental physics is concerned with the basic kinds of physical systems, and it seeks to characterize the states of these systems so completely as to determine all the physical properties of all the systems these constitute. A physical property of an object will then be both qualitative and intrinsic just in case its possession by that object is wholly determined by the underlying physical states and physical relations of all the basic systems that compose that object. Of course, physics has yet to achieve, and indeed may never achieve, true descriptive completeness in this sense. But to the extent that it is successful, it simultaneously defines and discovers an important class of qualitative intrinsic properties.[10]

What is meant by a process being supervenient upon an assignment of qualitative intrinsic physical properties at spacetime points in a spacetime region *R?* The idea is familiar. It is that there is no difference in that process without some difference in the assignment of qualitative intrinsic physical properties at spacetime points in *R.* I take the geometric structure of *R* itself to be already fully specified by means of the spatiotemporal properties of and relations between its constituent points.[11] The supervenience claim is that if one adds to this geometric structure an assignment of qualitative intrinsic physical properties at spacetime points in *R,* then there is physically only one way in which that process can occur.

What is a physical process? While neither ordinary nor scientific usage can be expected to determine a unique, precise answer to this question, I offer the following rough, preliminary analysis. A particular physical process consists of a suitably continuous set of stages, typically involving one or more enduring systems. The stages occur in some definite sequence, and may be seen as conforming to some character-

10. Compare Quine's (1969) view of natural kinds as those which science seeks to define. And note also that Lewis 1986a gives a preliminary analysis of intrinsic properties in terms of natural properties.

11. If ***R*** is closed, it may be necessary to add information on how points in ***R*** are related to points just outside ***R.***

istic pattern, and/or tending toward some characteristic end. We use such features to group processes into types. A process may be spatially localized or, like continental drift or star formation, it may occur over an extended spatial region.[12]

Spatial Separability (and hence also *Separability* itself) is in question in Aspect-type experiments to the extent that the intrinsic properties of an "entangled" compound quantum system fail to supervene on those of its components. A violation of *Separability* of this kind would be associated with a kind of physical holism (see Healey 1991). Note that since *Local Action* does not presuppose *Separability,* a local but nonseparable account of these experiments is possible (see Healey 1994).[13]

The Aharonov-Bohm effect challenges *Separability* in a different way. In this case what is at issue is whether either the process by which each electron passes through the region outside the solenoid, or the (electro)magnetic potential there throughout the time of its passage, supervenes on qualitative intrinsic physical properties of (objects at) points in that region at moments during that time.[14]

12. Healey 1994 contains a more detailed analysis of various general features of processes.

13. I wish to make it clear that by saying this I am endorsing neither Howard's (1989) identification of locality with *Parameter Independence* and of separability with *Outcome Independence;* nor the suggestion, traceable to Jarrett 1984, that the analysis of failure of Bell inequalities singles out these two independence conditions as being of special physical significance. Indeed, one main goal of the present paper is to divert attention away from probabilistic independence conditions of limited applicability and toward conceptions of locality and separability that are at once more general and of greater physical and philosophical significance. The principles of *Local Action* and *Separability* stated here express just such general conceptions. Moreover, I believe that, assuming both principles, one can give a compelling derivation of Bell inequalities, though space limitations prevent me from offering it here; perhaps the reader will find room in the margin to add it!

14. In the light of my distinction between locality and separability, it is ironic that Aharonov (1984) himself argues that the Aharonov-Bohm effect is a non-local phenomenon—ironic, but not paradoxical. For consider what he means by 'non-local':

> Let us, first of all, say quite generally what we mean by a "non-local" property of a physical system. Suppose, we have a system which occupies two separate regions of space (the system might consist, for example, of two objects, one in each region; or, if it is a quantum system, it may consist of a single object whose wave-function is non-zero in these regions, but zero elsewhere). The essential difference between local and non-local properties of the system is that in the former case all possible information can be obtained by independent measurements made in the two regions, while in the latter case this is not true. (p. 12)

Apart from what I regard as the regrettably operationalist flavor inserted by the concluding reference to measurements, what Aharonov here describes as a 'non-local' property of a system in the region of space surrounding the solenoid at a time is just what

5. The Two-Slit Experiment. Since the Aharonov-Bohm experiment as depicted in Fig. 1 is a variation on the familiar two-slit experiment, it will be useful to begin by considering how the principles of *Local Action* and *Separability* may be applied there.

The two-slit experiment is conveniently modelled in nonrelativistic quantum mechanics by assigning a wave-function to the ensemble of electrons passing through the apparatus. Orthodox interpretations of quantum mechanics take this quantum-mechanical description of the electrons to be complete. But what does this mean? It may be understood as the radical claim that the wave-function has no descriptive significance—that it has the purely instrumental role of permitting statistical predictions of the results of measurements on the electrons, and it wholly fulfills that role (in the sense that no supplementary characterization of the electrons would permit more definite predictions of such results). This understanding goes along with a strong version of the Copenhagen interpretation, according to which quantum mechanics simply has nothing to say about a system when it is not being observed. Those who adhere to this version of the Copenhagen interpretation will not ascribe even a nonlocalized position to an electron in the two slit experiment until its position is observed at the detection screen.

But there is another way of understanding the completeness claim, which goes along with a weaker version of the Copenhagen interpretation. On this version, an individual system may be described by a wave-function somewhat as follows: if the wave function at some moment is non-negligible only for some set Δ of possible values of dynamical variable Q, then the electron has the dynamical property *Q is restricted to Δ*. For example, even though a hydrogen atom in a superposition of its ground and first excited states has no precise energy, it does have the property *energy is not greater than -3.4ev*. Applied to position, this interpretation implies that an electron may have an imprecise location, being localized only within a region in which its wave-function is non-negligible.[15] This does not, of course, imply that

---

I have described as the nonseparability of that system in that region at that time. His argument that the AB effect manifests "nonlocality" in fact supports the conclusion that the effect manifests nonseparability.

15. One may try to make this view more precise by imposing the so-called eigenvalue-eigenstate rule, according to which a system has a quantum mechanical dynamical property at a time if and only if its wave-function then assigns probability 1 to that property. But since no wave-function assigns probability 1 to any dynamical property of the form *is located at position r*, an electron never has an absolutely precise position, on this interpretation. At most, an electron may have some dynamical property of the

an electron has any component parts or internal spatial structure; it may still be a so-called "point-particle."

On this understanding, quantum mechanics describes the process involved in the passage of a single electron through the apparatus as follows. Some time after the electron is emitted from the source, its wave function approximates to a plane wave parallel to the barrier with the two slits in it. The electron then has no precise position. Indeed, its position is then highly nonlocalized: it is restricted to no region in the plane of the barrier of dimension comparable to the separation of the slits. At a later time the electron is located on the other side of the barrier. Its position is then more localized, but not in the neighborhood of just one of the two slits: rather, its position is then restricted to the union of two such neighborhoods, one close to each slit. One might take this to mean that the electron passes through both slits at once, as long as this is not understood to imply that the electron is a composite object, with different components going through different slits.[16] Subsequently, the position of the electron becomes less localized, until it is detected at a screen some distance behind the slits. Whether one believes that its position then becomes more narrowly localized depends on what account one accepts of the measurement process at the screen.

There are of course also unorthodox interpretations which reject the completeness claim, including views which assign a precise position to an electron at all times. Such a view appears committed to an account of the two-slit experiment that conflicts with *Local Action.* For if each electron passed through just one slit, then it would seem that to explain the interference pattern one would have to assume that opening or shutting the other slit produces an immediate effect on the distant electron, in violation of *Local Action.* But Bohm's (1952) view, which does assign a precise position to an electron at all times, in fact avoids this

---

form *is localized in region R,* for some compact *R:* but since a typical wave-function does not have compact support, this rule would imply that electrons are rarely if ever localized in any compact region! Despite its inherent lack of precision, the view is important in the context of the Aharonov-Bohm experiment. For if one follows orthodoxy in denying that an electron has a well-defined trajectory through the apparatus, an interpretation along these lines seems required to make sense of the claim that electrons are excluded from the region in which there is a nonzero magnetic field.

16. There is no such implication since there is no reason to suppose that the electron is composed at each moment of at least two enduring objects, one of which goes through one slit while the other goes through the other slit. It is interesting that Tonomura himself says after describing the single-electron buildup of an interference pattern in an analogous set-up: "Therefore, we must conclude that a single electron passes through both sides of the electron biprism and forms the probability amplitude of the biprism interference pattern." (Peshkin and Tonomura 1989, 139, caption to his Fig. 5.29)

implication by treating the wave-function itself as a physically real field which mediates the effect on the electron caused by opening or closing the slit through which it does not pass.[17]

Now on the more orthodox account of what happens in the two-slit experiment, all processes and interactions involved conform to *Local Action.* For that account *denies* that an electron is ever confined to the region of one slit. It implies that no electron has a trajectory which keeps it spatially distant from any external influence applied to either slit. Opening or shutting either slit produces a direct and immediate effect on an electron within the region to which its position is localized.

Instead, on this account, the two-slit experiment manifests a violation of *Separability.* The passage of an electron through the apparatus constitutes a nonseparable process. At each moment the electron will have physical properties (including its location) which do not supervene on an ascription of intrinsic physical properties at spacetime points within the region to which it is then confined.

6. Locality and Separability in the Aharonov-Bohm Effect. When a current is passed through the solenoid, the two-slit experiment depicted in Fig. 1 manifests the Aharonov-Bohm effect. In what sense, if any, is this effect nonlocal? While no interpretation gives a completely local account of the effect, on some interpretations the effect involves a violation of *Local Action,* on others a violation just of *Separability,* and on others violations of both principles. This demonstrates a close analogy between the Aharonov-Bohm effect and violations of Bell inequalities. For the violation of Bell inequalities may also be interpreted in one of these different ways, depending on how one understands the application of quantum mechanics to this phenomenon.

Consider first a strong version of the Copenhagen interpretation

17. One may still question whether there is nonlocality in a Bohmian treatment of the two-slit experiment, either on the grounds that "momentum is nonlocal on Bohm's view" (to quote an anonymous reviewer of an ancestor of this paper), or on the grounds that the electrons' positions are affected nonlocally if, for example, a detector is placed near one slit. Now it is true that on Bohm's view the momentum we "measure" is not an intrinsic property of an electron but depends on the experimental context, and what are called momentum measurements in fact reduce to measurements of position. But this fact alone constitutes a violation of neither *Local Action* nor *Separability.* And certainly, on Bohm's view, placing a detector near one slit will affect the trajectories of all electrons, even those not passing near that slit. But that is because the addition of the detecting system expands the effective configuration space into one appropriate to a compound system, and thereby *introduces* violations of *Local Action* of the kind that are familiar from the application of Bohm's view to Aspect-type experiments; no such violations are inherent in the application of Bohm's view to the unmodified two-slit experiment.

which declines to attribute any location to the electrons between emission and detection. On this view, it is senseless to ask whether the electrons are acted on nonlocally as they pass through the apparatus. But since the causal connection between alterations in the current through the solenoid and changes in the interference pattern cannot therefore be said to be mediated by the passage of electrons around the solenoid, this causal connection does not conform to *Local Action.*

Other interpretations of quantum mechanics do not decline to describe the passage of electrons through the apparatus. Some ascribe a well-defined trajectory to each electron, others describe their passage in some less classical way. Any interpretation that ascribes a nonlocalized position to an electron on its way through the apparatus is committed to a violation of *Separability* already in the two-slit experiment, and *a fortiori* in the Aharonov-Bohm experiment. But is such an interpretation also committed to some violation of *Local Action* in the Aharonov-Bohm effect? On Bohm's view, the simple two-slit experiment involves no violation of *Local Action* or of *Separability.* Can a Bohmian also give a local, separable account of the Aharonov-Bohm experiment? In order to answer such questions, it is necessary to look more closely at the representation of electromagnetism in the Aharonov-Bohm effect and elsewhere.

7. Is Electromagnetism Local? Clearly the Aharonov-Bohm effect involves some kind of interaction between (electro)magnetic fields or potentials and the interfering electrons. If either that interaction, or the fields or potentials themselves, are not local, then nor is the effect itself. Now if the (electro)magnetic field acts directly on the electrons, and if the field is non-zero only inside the solenoid, while the electrons are never located inside the solenoid, then we have a violation of *Local Action.* But the gauge-dependence of the potential makes it hard to see how it could provide the mediation needed to restore conformity to *Local Action,* irrespective of the gauge-invariant nature of the effect itself.

Now following Wu and Yang's (1975) analysis, it has become common to consider electromagnetism to be completely and nonredundantly described in all instances neither by the electromagnetic field, nor by its generating potential, but rather by the so-called

$$\textbf{\textit{Dirac phase factor}} \quad \exp[-(ie/\hbar) \oint_C A^\mu(x^\mu) \cdot dx^\mu]$$

where A^μ is the electromagnetic potential at spacetime point x^μ, and the integral is taken over each closed loop C in spacetime. Applied to the present instance of the Aharonov-Bohm effect, this means that the

constant magnetic field in the solenoid is accompanied by an association of a phase factor $S(C)$ with all closed curves C in space, where $S(C)$ is defined by

$$\exp[-(ie/\hbar) \oint_C \boldsymbol{A(r).dr}]$$

This approach has the advantage that since $S(C)$ is gauge-invariant, it may readily be considered a physically real quantity. Moreover, the effects of electromagnetism outside the solenoid may be attributed to the fact that $S(C)$ is nonvanishing for those closed curves C that enclose the solenoid whenever a current is passing through it. But it is significant that, unlike the magnetic field and its potential, $S(C)$ is not defined at each point of space at each moment of time.

Can $S(C)$ at some time be taken to represent an intrinsic property of a region of space corresponding to the curve C? There are two difficulties with this suggestion. The first is that the presence of the quantity e in the definition of $S(C)$ appears to indicate that $S(C)$ rather codes the effect of electromagnetism on objects with that specific charge. If in fact *all* charges are multiples of some minimal value e, then this would no longer be a problem; the fact that $S(C)$ at some time represents an intrinsic property of a region of space corresponding to the curve C would be a natural reflection of this fact. If not, one could rather take $I(C) = \oint_C \boldsymbol{A}\cdot d\boldsymbol{r}$ to be an intrinsic property of C. The second difficulty is that closed curves do not correspond uniquely to regions of space; e.g., circling the solenoid twice on the same circle will produce a different curve from circling it once. But this does not prevent one from taking $S(C)$ at some time to represent an intrinsic property of the region of space occupied by a nonself-intersecting closed curve C.

Once these difficulties have been handled, it is indeed possible to consider electromagnetism in the Aharonov-Bohm effect as faithfully represented at a time by a set of intrinsic properties of regions of space occupied by nonself-intersecting closed curves. But if one does so, then electromagnetism itself manifests nonseparability! For these intrinsic properties do not supervene on any assignment of qualitative intrinsic physical properties at spacetime points in the region concerned. Whether the current through the solenoid remains constant or changes, the associated electromagnetism constitutes a nonseparable process, and so the Aharonov-Bohm effect violates *Separability*.[18]

18. There is an alternative perspective according to which the electromagnetic potential is represented as a connection one-form on a principle fiber bundle, with Minkowski spacetime (or some region of it) as base space and the group U(1) as fibre. Though mathematically elegant, this does not render electromagnetism separable in the AB

with the earlier result, that if the motion of the electrons through the apparatus is a nonseparable process, then it is possible to account for the AB effect in terms of a purely local interaction between (nonseparable) electromagnetism and this process.

9. Conclusion. The kind of nonlocality manifested by the Aharonov-Bohm effect is much more closely analogous to the kind of nonlocality manifested by violations of Bell inequalities than has been previously acknowledged. Neither effect can be given a completely local explanation. But in both cases one may analyze the residual nonlocality as involving the violation either of a principle of local action, or of a principle of separability, or of both; and in both cases, exactly how one analyzes the nonlocality depends on how one interprets quantum mechanics. The fact that the same general principles of local action and separability come into question in both cases is one reason to take these principles as basic, so that more specialized "locality" principles may be seen to derive from their application to the different circumstances of the two effects.

Another reason to take local action and separability as basic is just that these do, after all, capture the most interesting (and interestingly different) notions from the point of view of natural philosophy. That violations of Bell inequalities manifest action at a distance would be a striking conclusion, even if such action could not be used to transmit superluminal messages. The alternative conclusion, that it is because some systems have nonlocalized properties that Bell inequalities are violated, would be striking in a different way. The conclusion that the Aharonov-Bohm effect manifests action at a distance would be ironic, for (as Einstein noted in the passage quoted in Section 3) postulating a physically real electromagnetic field is generally taken to be a way of avoiding any appeal to action at a distance. By accepting the alternative conclusion, that the Aharonov-Bohm effect arises because electromagnetism acts nonseparably, one might eliminate action at a distance. But this acceptance comes at a price: it involves the denial of the view that ". . . all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another. . . . We have geometry: a system of external relations of spatiotemporal distance between points. . . . And at those points we have local qualities: perfectly natural intrinsic properties which need nothing bigger than a point at which to be instantiated. . . . And that is all. . . . All else supervenes on that." (Lewis 1986, x).[22]

22. I take it to be significant that on a Bohmian interpretation violations *both* of *Local Action and* of *Separability* occur in the experiments of Aspect as well as Tonomura.

Evidence for Aharonov-Bohm Effect with Magnetic Field Completely Shielded from Electron Wave

Akira Tonomura, Nobuyuki Osakabe, Tsuyoshi Matsuda, Takeshi Kawasaki, and Junji Endo
Advanced Research Laboratory, Hitachi Ltd., Kokubunji, Tokyo 185, Japan

and

Shinichiro Yano and Hiroji Yamada
Central Research Laboratory, Hitachi, Ltd., Kokubunji, Tokyo 185, Japan
(Received 4 December 1985)

Evidence for the Aharonov-Bohm effect was obtained with magnetic fields shielded from the electron wave. A toroidal ferromagnet was covered with a superconductor layer to confine the field, and further with a copper layer for complete shielding from the electron wave. The expected relative phase shift was detected with electron holography between two electron beams, one passing through the hole of the toroid, and the other passing outside. The experiment gave direct evidence for flux quantization also.

PACS numbers: 03.65.Bz, 41.80.Dd

The Aharonov-Bohm (AB) effect[1] has recently received much attention as an unusual but important quantum effect.[2] The predicted effect is the production of a relative phase shift between two electron beams enclosing a magnetic flux even if they do not touch the magnetic flux. Such an effect is inconceivable in classical physics and directly demonstrates the gauge principle of electromagnetism.[3]

Although the affirmative experimental test was offered[4] soon after its prediction, Bocchieri *et al.*[5] and Roy[6] questioned the validity of the test, attributing the phase shift to leakage fields. The authors' recent experiment[7] using a toroidal magnet established the existence of the AB effect, under the condition of complete confinement of the magnetic field in the magnet; electron holography confirmed quantitatively the expected relative phase shift between the two beams. Bocchieri, Loinger, and Siragusa[8] still argued that the phase shift could be due to the Lorentz-force effect on the portion of the electron beam going through the magnet.[9]

The present experiment[10] is designed to provide a crucial test of the AB effect. A tiny toroidal magnet covered entirely with a superconductor layer and further with a copper layer is fabricated. The two layers prevent the incident electron wave from penetrating the magnet. In addition, the magnetic field is confined to the toroidal magnet by the Meissner effect of the covering superconductor. Then the relative phase shift between two electron beams, one passing through a region enclosed by the toroid and the other passing outside the toroid, is measured by means of electron holography. The experimental results detected the predicted relative phase shift, giving conclusive evidence for the AB effect. This experiment also demonstrated the flux quantization.[11]

Tiny toroidal samples were fabricated by use of photolithography. A Permalloy (80% Ni and 20% Fe) thin film, 200 Å thick, was prepared by vacuum evaporation on a silicon wafer covered with Al (3000 Å thick), Nb (2500 Å thick), and SiO (500 Å thick); the SiO layer serves to reduce the coercive force of the Permalloy. After evaporation of a 2000-Å-thick layer of SiO on the Permalloy, the toroidal shape was cut out to the depth of the Nb surface. The NbO produced by the lithography processes at the Nb surface had to be removed to ensure a perfect contact with the Nb layer (2500 Å thick) that was subsequently sputtered on the whole structure (see Fig. 1). The superconducting contact was confirmed by another experiment. We note that the thickness of the upper SiO layer decreased to 500 Å after the ion sputtering.

A toroidal sample with a tiny support bridge (see the scanning electron micrograph in Fig. 2) was then cut so that the Permalloy toroid was completely covered

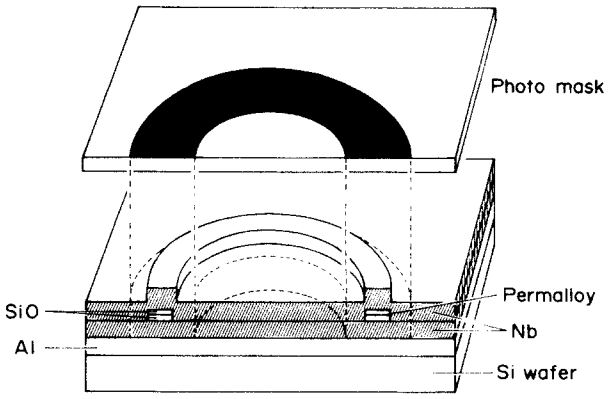


FIG. 1. Schematic diagram for fabrication of the toroidal magnet.

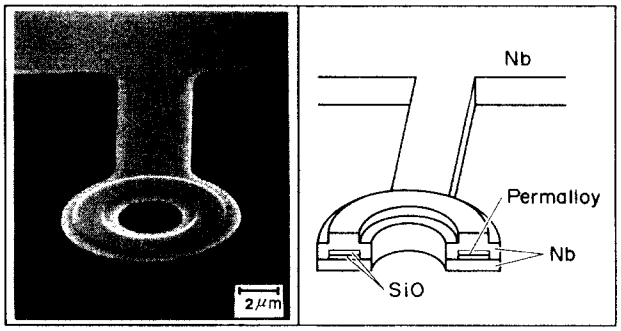


FIG. 2. Toroidal magnet. (a) Scanning electron micrograph; (b) diagram. The toroid is connected to a Nb plate by a tiny bridge for high thermal conductivity.

by the superconducting bulk Nb. The toroidal sample was peeled off the wafer by dissolving the Al in NaOH solution, and was placed on a Cu mesh. Finally, a copper film 500–2000 Å thick was evaporated on all of its surfaces; the film serves to prevent penetration of the electron wave, and to keep the sample from experiencing charge-up and contact-potential effects.

Electron holograms were formed in a 150-kV field-emission electron microscope (wavelength, 0.030 Å) that had a liquid-He–cooled specimen stage attached. The object wave, phase shifted by the sample, and the reference wave were brought together by the electron biprism to form an interference pattern, as shown in Fig. 3. The pattern was enlarged 1000 times by electron lenses and recorded on film as a hologram.

The phase shift due to the sample was reconstructed by means of He-Ne laser light (wavelength, 6328 Å) in the optical system shown in Fig. 4. Two waves, A and B, illuminated the hologram. Each wave produces two diffracted waves, one which reconstructs the phase shift due to the sample, and the other, its conjugate.

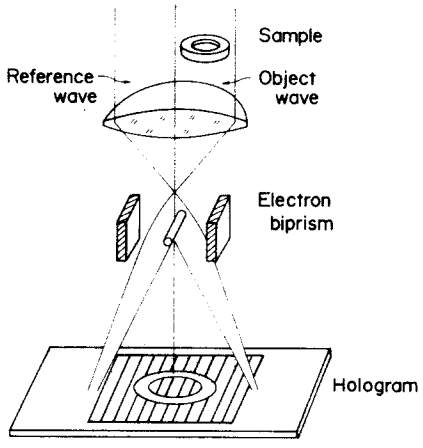


FIG. 3. Electron-optical system for hologram formation.

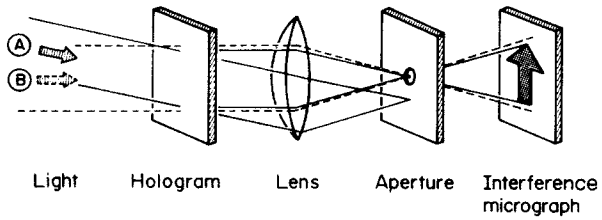


FIG. 4. Optical reconstruction system for interference microscopy.

An interference micrograph is obtained when the reconstructed image of beam A is superposed with beam B after passage through an aperture. Moreover, a twice-phase-amplified interference micrograph[12] is formed when the reconstructed image of beam A and the conjugate image of beam B are superposed by the tilting of beam B.

The leakage fluxes of fabricated samples at room temperature were quantitatively measured[13] by interference electron microscopy, and only samples with flux less than $h/20e$ [14] were selected for this experiment. Figure 5 shows an example of a twice-phase-amplified interference micrograph, which indicates a very large leakage flux of $\sim 2h/e$.

Now, the AB effect is the production of a relative phase shift of $\pi\Phi/(h/2e)$ between two electron beams enclosing magnetic flux Φ. The interference micrograph in Fig. 6(a) is clear evidence for the AB effect. Each interference fringe inside the ring, i.e., the image of the toroidal sample, lies just in the middle of two fringes outside the ring. This shows that there is a relative phase shift $n\pi$ (n odd), as expected from the quantized magnetic flux $nh/2e$ enclosed within the superconducting Nb. That the relative phase shift here is an integral multiple of π can be seen precisely from the twice-phase-amplified micrograph obtained from the same hologram [Fig. 6(b)], in which there are no relative displacements between the fringes inside and outside the ring. We emphasize that the magnetic flux is confined within the superconductor and that the

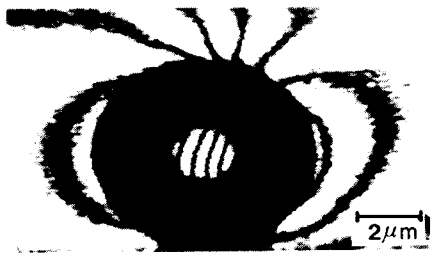


FIG. 5. Leakage fields from a toroidal magnet (phase amplification, 2×). Leakage flux can be quantitatively measured since a constant flux of $h/2e$ flows between two adjacent interference fringes.

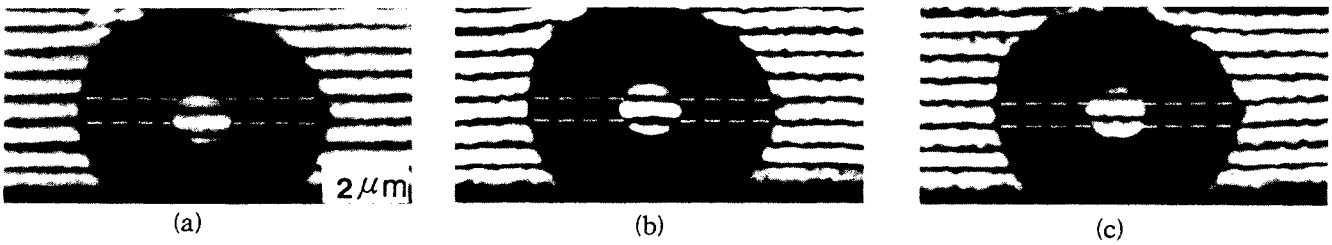


FIG. 6. Interference micrographs of a toroidal magnet at low temperatures. (a) $T = 4.5$ K (phase amplification, $1\times$); (b) $T = 4.5$ K (phase amplification, $2\times$); (c) $T = 15$ K (phase amplification, $2\times$). The enclosed flux is quantized in units of $h/2e$ when $T < T_c$ ($= 9.2$ K). The number of fluxons is odd.

field is shielded from the electron wave by the Cu and Nb covering. It is estimated that the leakage flux is far less than $h/20e$, since the leakage flux at room temperature is less than $h/20e$ and the minimum thickness and penetration depth of Nb are 2500 and 1100 Å, respectively. Only a slight portion, approximately 10^{-6}, of the incoming electron wave is estimated to reach the magnetic field coherently, since a 150-kV electron beam has to penetrate through the Cu (~ 1000 Å) and Nb (2500 Å) layers for it. The sufficient shielding of electron penetration is also supported by the experimental result that the change in the Cu-layer thickness from 500 to 2000 Å had no effect on the interference fringes around the quantized magnetic flux.

If the temperature T of the sample is raised, the interference pattern changes abruptly when T crosses the superconducting critical temperature T_c; the relative phase shift is no longer an integral multiple of π. In the case of Figs. 6(a) and 6(b), it in fact becomes $(0.32 + n)\pi$ as can be seen from Fig. 6(c). The transition was confirmed to be reversible. This behavior is evidence for the effect of the superconductor that confines the magnet flux quanta below T_c.

Of course, there are cases of even n, in which no relative displacements are observed, as shown in Figs. 7(a) and 7(b). With this sample, the relative displacement can be seen only when its temperature is raised above T_c; the displacement in Fig. 7(c) represents a relative phase shift of $(0.25 + n)\pi$ (n even).

When the temperature T of the sample was further raised to room temperature, the relative displacement changed by half the fringe spacing in a twice-phase-amplified interference micrograph; this corresponds to the estimated decrease ($\sim 5\%$) in the magnetization of the Permalloy. This temperature dependence supports our view that the relative phase shift is controlled by the magnetic flux of the Permalloy.

The experimental results described above provide crucial evidence for the existence of the AB effect. Furthermore, the quantization of the flux trapped by a superconductor was directly observed with use of the AB effect of an electron beam.

The most controversial point in the dispute over experimental evidence for the AB effect has been whether or not the phase shift would be observed when both electron intensity and magnetic field were extremely small in the region of overlap. Since experimental realization of absolutely zero field is impossible, the continuity of physical phenomena in the transition from negligibly small field to zero field should be accepted instead of perpetual demands for the ideal; if a discontinuity there is asserted, only a futile agnosticism results.

The authors are grateful for the idea for this experiment, which was proposed by Professor Chen Ning Yang of the State University of New York.[15] Also deserving of thanks are Dr. Ushio Kawabe of Hitachi, Ltd., for his advice and stimulation, Mr. Mikio Hirano of Hitachi, Ltd., for his help in preparing samples, and

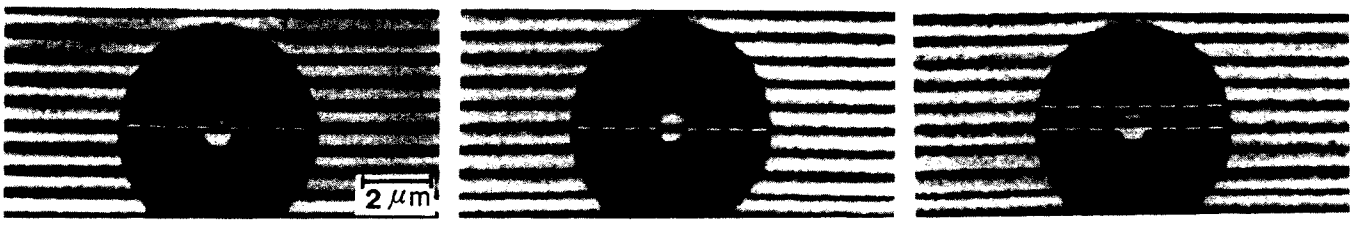


FIG. 7. Interference micrographs of a toroidal magnet at low temperatures. (a) $T = 4.5$ K (phase amplification, $1\times$); (b) $T = 4.5$ K (phase amplification, $2\times$); (c) $T = 15$ K (phase amplification, $2\times$). The number of fluxons is even.

Mr. Shuji Hasegawa of Hitachi, Ltd., for his assistance in the experiment. We also gratefully acknowledge the valuable discussions and advice in preparing this manuscript given by Professor Hiroshi Ezawa of Gakushuin University, and also by Dr. Akira Fukuhara of Hitachi, Ltd. Thanks are due to Professor Ryozo Aoki of Kyushu University for his cooperation in developing a liquid-He–cooled specimen stage.

[1]Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485 (1959).

[2]For example, S. Olariu and I. I. Popescu, Rev. Mod. Phys. **57**, 339 (1985).

[3]T. T. Wu and C. N. Yang, Phys. Rev. D **12**, 3845 (1975).

[4]R. G. Chambers, Phys. Rev. Lett. **5**, 3 (1960); H. A. Fowler, L. Marton, J. A. Simpson, and J. A. Suddeth, J. Appl. Phys. **32**, 1153 (1961); H. Boersch, H. Hamisch, K. Grohmann, and D. Wohlleben, Z. Phys. **165**, 79 (1961); G. Möllenstedt and W. Bayh, Phys. Bl. **18**, 299 (1962).

[5]P. Bocchieri and A. Loinger, Nuovo Cimento Soc. Ital. Fis. **47A**, 475 (1978); P. Bocchieri, A. Loinger, and G. Siragusa, Nuovo Cimento Soc. Ital. Fis. **51A**, 1 (1979); P. Bocchieri and A. Loinger, Lett. Nuovo Cimento Soc. Ital. Fis. **30**, 449 (1981).

[6]S. M. Roy, Phys. Rev. Lett. **44**, 111 (1980).

[7]A. Tonomura *et al.* Phys. Rev. Lett. **48**, 1443 (1982).

[8]P. Bocchieri, A. Loinger, and G. Siragusa, Lett. Nuovo Cimento Soc. Ital. Fis. **35**, 370 (1982).

[9]The phase shift was also detected when the top surface of a toroidal magnet was covered with gold film thick enough to prevent electron penetration. See A. Tonomura *et al.*, in *Proceedings of the International Symposium on Foundations of Quantum Mechanics, Tokyo, 1983*, edited by S. Kamefuchi *et al.* (Physical Society of Japan, Tokyo, 1984), p. 20.

[10]A similar experiment using a hollow toroidal superconductor was proposed by C. G. Kuper, Phys. Lett. **79A**, 413 (1980).

[11]The quantization of the trapped flux in a hollow superconducting cylinder has been detected by electron interferometry. See H. Wahl, Optik **28**, 417 (1968/1969); B. Lischke, Phys. Rev. Lett. **22**, 1366 (1969).

[12]J. Endo, T. Matsuda, and A. Tonomura, Jpn. J. Appl. Phys. **18**, 2291 (1979).

[13]A. Tonomura *et al.*, Phys. Rev. Lett. **44**, 1430 (1980); T. Matsuda *et al.*, J. Appl. Phys. **53**, 5444 (1982); N. Osakabe *et al.*, Appl. Phys. Lett. **42**, 746 (1983).

[14]Holographic interference microscopy is estimated to be as precise as $\frac{1}{30}$ of a wavelength in an ideal case, which corresponds to a magnetic flux of $h/50e$. See A. Tonomura *et al.*, Phys. Rev. Lett. **54**, 60 (1985).

[15]C. N. Yang, in *Proceedings of the International Symposium on Foundations of Quantum Mechanics, Tokyo, 1983*, edited by S. Kamefuchi *et al.* (Physical Society of Japan, Tokyo, 1984), p. 27.

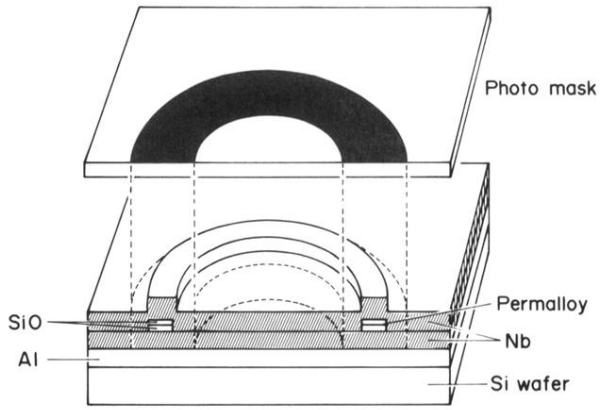


FIG. 1. Schematic diagram for fabrication of the toroidal magnet.

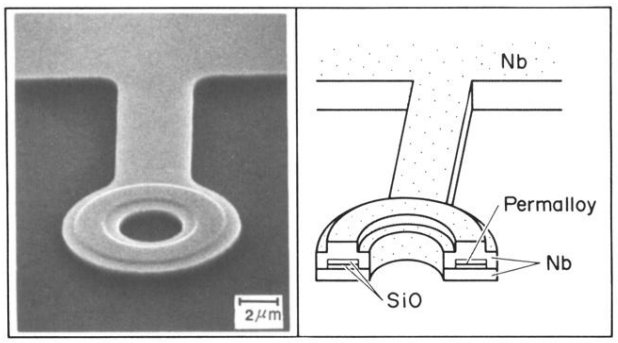


FIG. 2. Toroidal magnet. (a) Scanning electron micrograph; (b) diagram. The toroid is connected to a Nb plate by a tiny bridge for high thermal conductivity.

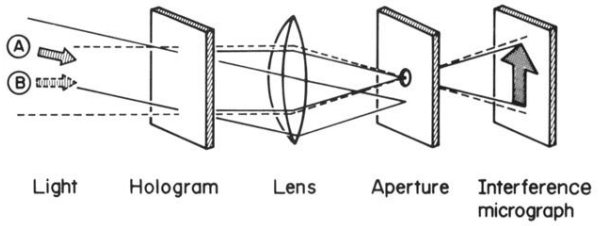


FIG. 4. Optical reconstruction system for interference microscopy.

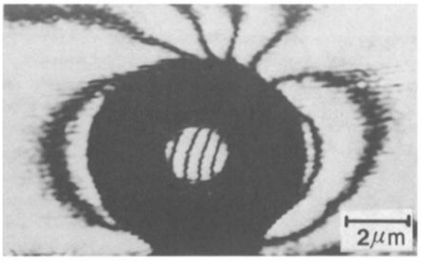


FIG. 5. Leakage fields from a toroidal magnet (phase amplification, 2×). Leakage flux can be quantitatively measured since a constant flux of $h/2e$ flows between two adja cent interference fringes.

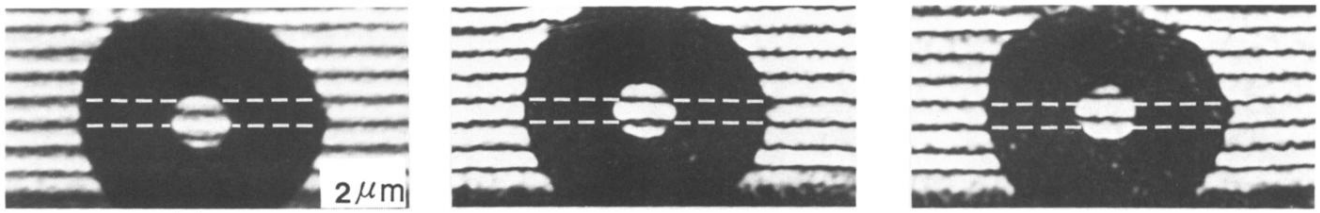
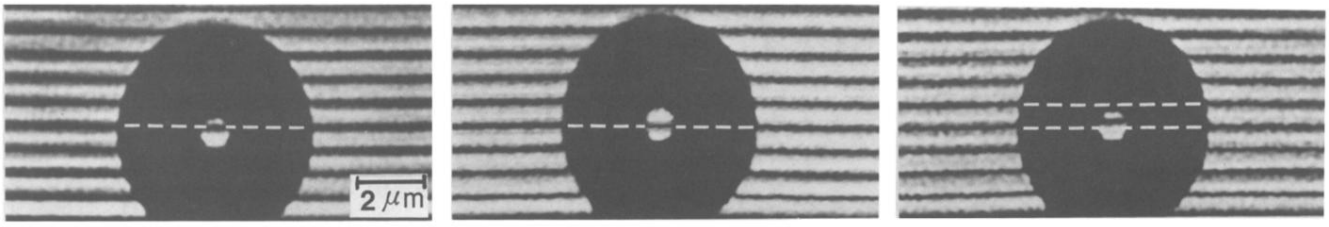


FIG. 6. Interference micrographs of a toroidal magnet at low temperatures. (a) $T = 4.5$ K (phase amplification, $1\times$); (b) $T = 4.5$ K (phase amplification, $2\times$); (c) $T = 15$ K (phase amplification, $2\times$). The enclosed flux is quantized in units of $h/2e$ when $T < T_c$ ($= 9.2$ K). The number of fluxons is odd.



(a) (b) (c)

FIG. 7. Interference micrographs of a toroidal magnet at low temperatures. (a) $T = 4.5$ K (phase amplification, $1\times$); (b) $T = 4.5$ K (phase amplification, $2\times$); (c) $T = 15$ K (phase amplification, $2\times$). The number of fluxons is even.

PHYSICAL REVIEW A **86**, 040101(R) (2012)

Role of potentials in the Aharonov-Bohm effect

Lev Vaidman

Raymond and Beverly Sackler School of Physics and Astronomy, Tel-Aviv University, Tel-Aviv 69978, Israel

(Received 12 October 2011; revised manuscript received 12 January 2012; published 10 October 2012)

There is a consensus today that the the main lesson of the Aharonov-Bohm effect is that a picture of electromagnetism based on the local action of the field strengths is not possible in quantum mechanics. Contrary to this statement, it is argued here that when the source of the electromagnetic potential is treated in the framework of quantum theory, the Aharonov-Bohm effect can be explained without the notion of potentials. It is explained by local action of the field of the electron on the source of the potential. The core of the Aharonov-Bohm effect is the same as the core of quantum entanglement: the quantum wave function describes all systems together.

DOI: 10.1103/PhysRevA.86.040101 PACS number(s): 03.65.Ta, 03.65.Ud, 03.65.Vf

Before the Aharonov-Bohm (AB) effect [1] was discovered, the general consensus was that particles can change their motion only due to fields at their locations, fields which were created by other particles. The main revolutionary aspect of the AB effect was that this is not generally true, and that in certain setups two particles, prepared in identical states, move in the same fields but end up in different final states. In particular, the electromagnetic field can vanish at every place where the electron has been, yet the electron motion is affected by the electromagnetic interaction. The AB effect states that the motion of an electron is completely defined by the potentials in the region of its motion and not just by the fields. The potentials depend on the choice of gauge, which cannot affect the motion of particles, but there are gauge-invariant properties of the potentials (apart from the fields) that specify the motion of particles. The validity and the meaning of the AB effect has been extensively discussed [2–15]. I argue that there is an alternative to the commonly accepted mechanism which leads to the effect, and that we might change our understanding of the nature of physical interactions back to that of the time before the AB effect was discovered. The quantum wave function changes due to local actions of fields.

The discussion will be on the level of gedanken experiments, without questioning the feasibility of such experiments in today's laboratory. Consider a Mach-Zehnder interferometer for electrons tuned in such a way that the electron always ends up in detector B; see Fig. 1. We can change the electric potential in one arm of the interferometer such that there will be no electromagnetic field at the location of the wave packets of the electron but, nevertheless, the electron will change its behavior and sometimes (or it can be arranged that always) will end up in detector A. This is the electric AB effect. Alternatively, in the magnetic AB effect, the interference picture can be changed due to a solenoid inside the interferometer which produces no electromagnetic field at the arms of the interferometer.

Let us start our analysis with the electric AB effect. In the original proposal, the potential was created using conductors, capacitors, etc. While those are closer to a practical realization of the experiment, a precise theoretical description of such devices is difficult. I consider, instead, two charged particles, the fields of which cancel at the location of the electron.

For simplicity of presentation, instead of the Mach-Zehnder interferometer, I shall consider a one-dimensional

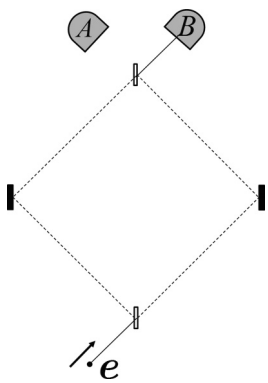


FIG. 1. Mach-Zehnder interferometer with electron as a test bed of the AB effect. Introduction of a relative electric potential between the arms of the interferometer or of a solenoid inside the interferometer spoils the destructive interference in detector A.

interferometer; see Fig. 2. (In fact, for an observer moving with a constant velocity in a perpendicular direction, this interferometer looks very much like the one described in Fig. 1.) The electron wave packet starts moving to the right toward a barrier which transmits and reflects equal-weight wave packets toward mirrors A and B. After reflection from the mirrors, the wave packets split again on the barrier. The interferometer is tuned in such a way that there is a complete destructive interference toward mirror A, and the electron reaches mirror B with certainty.

Another modification (the sole purpose of which is simplicity of the quantitative analysis of the experiment) is design of a special mirror for the electron which makes it spend a long time τ near it. For this purpose we introduce an interaction between the electron and the mirror with potential energy as a function of the electron distance from the mirror shown in Fig. 3. It goes to infinity at the surface of the mirror, smoothly becomes a constant value V at $x \in (0,d)$, and smoothly goes to zero for $x > d$. The energy of the electron is only slightly higher than V. The dimensions of the interferometer are much larger than d and we state that the electron is near the mirror when $x \in (0,d)$.

The source of the AB potential will be two particles of mass M and charge Q placed symmetrically on the perpendicular axis at equal large distances from mirror A. They have equal initial velocities toward the location of mirror A. At equal

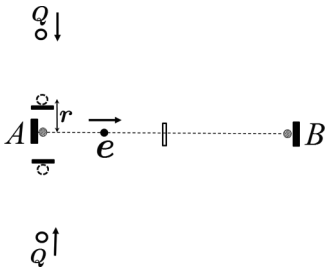


FIG. 2. A realization of the electric AB effect. Identical charges brought symmetrically to the electron wave packet in the left arm of the interferometer create a potential for the electron without creating an electric field at its location.

distance r from the mirror, the charged particles bounce back due to other similarly designed mirrors, which make the charges spend a time T near these mirrors. We choose $T < \tau$, so that the charges Q are near their respective mirrors during the time the electron's wave packets are near their mirrors. We then can approximate the potential that the electron in the left arm experiences as $\frac{-2eQ}{r}$ for the time T. Indeed, when the charges are far away, their potential can be neglected, and the time the charges travel toward and from the mirror is much smaller than T. Thus, the phase difference between the two wave packets of the electron is

$$\phi_{AB} = \frac{-2eQT}{r\hbar}. \tag{1}$$

The electron does not experience an electric field at any place where its wave packet passed, but it exhibits an interference pattern which is different from the pattern obtained in such an experiment by a neutral particle.

How can this result be understood if we consider all particles? The quantum state of the composite system is a superposition of two product states which I name branches. In the first one, the wave packet of the electron is on the left and in the other, it is on the right. The energy in the left branch is equal to the energy in the right branch, so energetic considerations cannot explain the phase difference. The electron does not experience any electric force, so the electron's wave packets are not shifted and thus cannot provide an explanation of the effect. The charges Q, however, do experience different forces in different branches. Thus, their wave packets in the left branch are slightly shifted relative to their wave packets in the right branch.

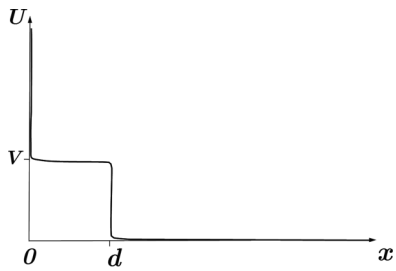


FIG. 3. The potential of the mirror forces. The potential energy of the particle as a function of its distance from the mirror. The particle with an energy slightly higher than V spends long time near the mirror.

Let us calculate the shift of position of the wave packet of one of the two Q charges due to its electromagnetic interaction with the electron. The shift is developed during the time T when this charge is near its mirror. The interaction with the electron leads to a small perturbation in the motion of the charge and, since $d \ll r$, the velocity of the charge during this time, v, can be considered to be constant. The change in the kinetic energy of the charge due to its interaction with the electron allows us to find the change in its velocity and thus the shift δx we are looking for:

$$\frac{-eQ}{r} = \delta\left(\frac{Mv^2}{2}\right) \simeq Mv\delta v \Rightarrow \delta x = \frac{-eQT}{Mvr}. \tag{2}$$

To observe the interference in the AB experiment, this shift should be much smaller than the position uncertainty of the charges. The de Broglie wavelength of the charge $\lambda = \frac{h}{Mv}$. Both charges Q are shifted in the same way, creating the AB phase: $2\frac{\delta x}{\lambda}2\pi = \phi_{AB}$.

The entanglement between the electron and the charges, which could be created if the uncertainty in the velocity of the charges when they are near their mirrors is smaller than δv, disappears when the charges Q travel back. Note, however, that if, contrary to our assumption, the position uncertainty of the charges is smaller than δx, then the entanglement will remain and will lead to decoherence, washing out the AB effect.

Let us turn now to the magnetic AB effect. I will show that the AB effect arises from different shifts of the wave packets of the source which experiences different local electric fields created by the left and the right wave packets of the electron.

Consider the following setup. The solenoid consists of two cylinders of radius r, mass M, large length L, and charges Q and $-Q$ homogenously spread on their surfaces. The cylinders rotate in opposite directions with surface velocity v. The electron encircles the solenoid with velocity u in a superposition of being in the left and in the right sides of the circular trajectory of radius R; see Fig. 4.

The flux in the solenoid due to the two cylinders is

$$\Phi = 2\pi r^2 \frac{4\pi}{c}\frac{Qv}{2\pi rL} = \frac{4\pi Qvr}{cL}. \tag{3}$$

Thus, the AB phase, i.e., the change in the relative phase between the left and the right wave packets due to the electromagnetic interaction, is

$$\phi_{AB} = \frac{e\Phi}{c\hbar} = \frac{4\pi eQvr}{c^2L\hbar}. \tag{4}$$

To simplify the alternative calculation based on direct action of the electromagnetic field, we assume $r \ll R \ll L$. Before entering the circular trajectory, the electron moves toward the axis of the solenoid and thus it provides zero total flux through any cross section of the solenoid. During its motion on the circle, the magnetic flux through a cross section of the solenoid at distance z from the perpendicular drawn from the electron is

$$\Phi(z) = \frac{\pi r^2 euR}{c(R^2+z^2)^{3/2}}. \tag{5}$$

When the electron enters one arm of the circle, it changes the magnetic flux and causes an electromotive force on the charged solenoids which changes their angular velocity. In

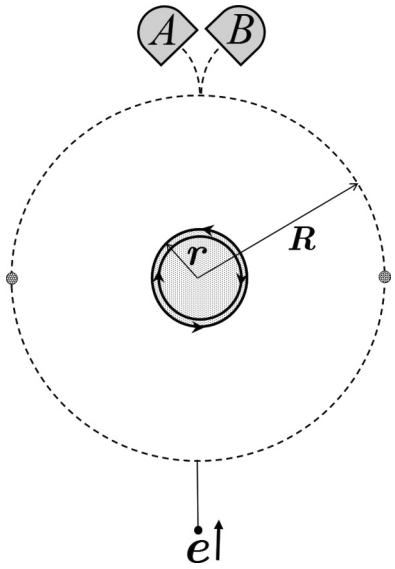


FIG. 4. The magnetic AB effect. The electron wave packet coming directly toward the solenoid splits into a superposition of two wave packets which encircle the solenoid from two sides and come out almost in the same direction, interfering toward detectors A and B.

order to calculate the change in the velocity of the surface of the cylinder we have to integrate the impulse exerted on all thin slices of the cylinder. For simplicity, I consider here the surface motion as a linear motion. The contribution of a slice with an infinitesimal charge dQ to the impulse is $\frac{\Phi(z)dQ}{c2\pi r}$, and integration over the slices yields

$$\delta v = \frac{1}{M}\int_{-L/2}^{L/2} \frac{\pi r^2 euR}{c^2(R^2+z^2)^{3/2}} \frac{1}{2\pi r}\frac{Q}{L}dz \simeq \frac{uQer}{c^2MRL}. \quad (6)$$

Then, the shift of the wave packet of a cylinder during the motion of the electron is

$$\delta x = \delta v \frac{\pi R}{u} = \frac{\pi Qer}{c^2 ML}. \quad (7)$$

The relevant wavelength of the de Broglie wave of each cylinder is $\lambda = \frac{h}{Mv}$. For calculating the AB phase we should take into account that both cylinders are shifted and that they are shifted (in opposite directions) in the two branches. This leads to a factor 4 and provides the correct expression for the AB phase: $4\frac{\delta x}{\lambda}2\pi = \phi_{AB}$.

If the uncertainty in the velocity of the cylinders is smaller than δv, then, during the electron circular motion, the electron and the cylinders become entangled. But when the electron leaves the circular trajectory, it exerts an opposite impulse on the cylinders and this entanglement disappears.

I have explained both electric and magnetic AB effects through actions of local fields on the quantum wave function. The electron in states $|L\rangle_e$ and $|R\rangle_e$ causes, via action of its electromagnetic field, different evolutions for the quantum state of the source: $|\Psi_L\rangle_S$ and $|\Psi_R\rangle_S$. The total wave function of the electron and the source is

$$\frac{1}{\sqrt{2}}\left(|L\rangle_e|\Psi_L\rangle_S + |R\rangle_e|\Psi_R\rangle_S\right). \quad (8)$$

During the evolution, the source states $|\Psi_L\rangle_S$ and $|\Psi_R\rangle_S$ might become orthogonal, or mostly differ only in their phase, but at the end of the process, the states of the source are identical except for the AB phase. Thus, the total wave function becomes

$$\frac{1}{\sqrt{2}}|\Psi\rangle_S(|L\rangle_e + e^{i\phi_{AB}}|R\rangle_e), \quad (9)$$

and the AB phase can be observed in the electron interference experiment.

The celebrated manifestation of a quantum wave function for a combined system is the nonlocal correlations which are generated by entangled states. The AB effect is conceptually different, since it can appear even if in the state (8) there is almost no entanglement at all times.

One might wonder why, instead of performing exact calculations in the framework of quantum mechanics, I consider particles and cylinders pushed by fields in the framework of classical mechanics and then use the correspondence principle to calculate the shifts of the quantum wave packets of particles and cylinders. I have to follow this path because the standard formulation of quantum mechanics, and the Schrödinger equation in particular, are based on potentials. I hope that a general formalism of quantum mechanics based on local fields will be developed. It will provide a solution to the problem of motion of a quantum particle in a force field even if there is no potential from which it can be derived. Meanwhile my assertion provides one useful corollary: If the fields vanish at locations of all particles then these fields yield no observable effect.

Let us test this corollary. Consider a modification of the electric AB effect described above in which the charges Q do not automatically perform their motion toward mirror A and back, but only when the electron on the path A triggers this motion, i.e., only in the left branch. I choose a particular value of the charge of the external particles, $Q = 4e$ for which the total electric field at the location of each particle created by other particles is zero. Neither the electron nor the charges Q experience an electromagnetic field in any of the branches. My assertion is that there will be no AB effect in this setup, in spite of the fact that the electron of the left branch has an electric potential, while the electron of the right branch has not. The original treatment of the AB effect is invalid since we do not have here a motion of an electron in a classical electromagnetic field, but a treatment of the problem using a "private potential" created by induced charges [16] shows that indeed there is no AB effect in this case.

I believe that we can find an explanation of the kind presented above for any model of the AB experiment. However, the pictorial explanation of the creation of a relative phase due to spatial shifts of wave packets disappears when we go beyond the physics of moving charges. We can replace the charged cylinders by a line of polarized neutrons producing magnetic flux due to quantum spins. In this case there is no spatial shift of wave packets. I am not aware of any pictorial explanation of the change of the phase of the spin state of the neutron, but in contrast to the phase of the electron in the standard approach to the AB effect, the phases of neutrons are changed locally due to the magnetic field of the electron. This is also an explanation of the Aharonov-Casher (AC) effect [17]: the local electric field acting on the moving neutron is responsible

for the appearance of the AC phase. Note, however, that it does not lead to a classical lag of the center of mass of the neutron [18,19].

I have not presented a general proof that in order to have an observable effect, the particles must pass through regions of nonzero fields. Rather, what I have shown is that the setups of the electric and magnetic AB effects do not contradict this assertion. Note, however, that the last example, in which there is an electric field almost everywhere except at the locations of the particles and this field causes no effect, strongly supports my claim.

Since the electromagnetic potential at any point along the trajectory of the electron can be gauged away, the standard approach to the AB effect leads to a paradoxical, in my view, nonlocal feature of quantum mechanics: the AB phase which has observable manifestation is acquired inside the interferometer in spite of the fact that there is no particular place or time where this happens. I have shown that this peculiarity disappears when all relevant parts of the system are considered: the phase is gradually acquired by the source of the electromagnetic potential.

This result does not question the validity of the AB effect and does not diminish the importance of its numerous applications. It removes, however, conceptual claims associated with the AB effect regarding nonlocality and the meaning of potentials. The AB effect does not prove that the evolution of a composite system of charged particles cannot be described completely by fields at locations of all particles. The potentials might be just a useful auxiliary mathematical tool after all.

I thank Noam Erez, Yaron Kedem, Shmuel Nussinov, and Philip Pearle for useful discussions. This work has been supported in part by the Binational Science Foundation Grant No. 32/08 and the Israel Science Foundation Grant No. 1125/10.

[1] Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485 (1959).
[2] W. H. Furry and N. F. Ramsey, Phys. Rev. **118**, 623 (1960).
[3] M. Peshkin, I. Talmi, and L. J. Tassie, Ann. Phys. (NY) **12**, 426 (1961).
[4] Y. Aharonov and D. Bohm, Phys. Rev. **123**, 1511 (1961).
[5] B. Liebowitz, Nuovo Cimento **38**, 932 (1965).
[6] T. H. Boyer, Phys. Rev. D **8**, 1679 (1973).
[7] T. T. Wu and C. N. Yang, Phys. Rev. D **12**, 3845 (1975).
[8] P. Bocchieri and A. Loinger, Nuovo Cimento **47**, 475 (1978).
[9] S. M. Roy, Phys. Rev. Lett. **44**, 111 (1980).
[10] M. Peshkin, Phys. Rep. **80**, 375 (1981).
[11] D. M. Greenberger, Phys. Rev. D **23**, 1460 (1981).
[12] S. Olariu and I. I. Popescu, Rev. Mod. Phys. **57**, 339 (1985).
[13] M. Peshkin and A. Tonomura, *The Aharonov-Bohm Effect* (Springer-Verlag, Berlin, 1989).
[14] Y. Aharonov, T. Kaufherr, and S. Nussiov, J. Phys.: Conf. Ser. **173**, 012020 (2009).
[15] A. Walstad, Int. J. Theor. Phys. **49**, 2929 (2010).
[16] T. Kaufherr, Y. Aharonov, S. Nussinov, S. Popescu, and J. Tollaksen, Phys. Rev. A **83**, 052127 (2011).
[17] Y. Aharonov and A. Casher, Phys. Rev. Lett. **53**, 319 (1984).
[18] T. H. Boyer, Phys. Rev. A **36**, 5083 (1987).
[19] Y. Aharonov, P. Pearle, and L. Vaidman, Phys. Rev. A **37**, 4052 (1988).

Concept of nonintegrable phase factors and global formulation of gauge fields

Tai Tsun Wu*
Gordon McKay Laboratory, Harvard University, Cambridge, Massachusetts 02138

Chen Ning Yang†
Institute for Theoretical Physics, State University of New York, Stony Brook, New York 11794
(Received 8 September 1975)

Through an examination of the Bohm-Aharonov experiment an intrinsic and complete description of electromagnetism in a space-time region is formulated in terms of a nonintegrable phase factor. This concept, in its global ramifications, is studied through an examination of Dirac's magnetic monopole field. Generalizations to non-Abelian groups are carried out, and result in identification with the mathematical concept of connections on principal fiber bundles.

I. MOTIVATION AND INTRODUCTION

The concept of the electromagnetic field was conceived by Faraday and Maxwell to describe electromagnetic effects in a space-time region. According to this concept, the field strenght $f_{\mu\nu}$ describes electromagnetism. It was later realized,[1] however, that $f_{\mu\nu}$ by itself does not, in quantum theory, completely describe all electromagnetic effects on the wave function of the electron. The famous Bohm-Aharonov experiment, first beautifully performed by Chambers,[2] showed that in a multiply connected region where $f_{\mu\nu}=0$ everywhere there are physical experiments for which the outcome depends on the loop integral

$$\frac{e}{\hbar c}\oint A_\mu dx^\mu \tag{1}$$

around an unshrinkable loop. This raises the question of what constitutes *an intrinsic and complete description* of electromagnetism. In the present paper we wish to discuss this question and also its generalization to non-Abelian gauge fields.

An examination of the Bohm-Aharonov experiment indicates that in fact only *the phase factor*

$$\exp\left(\frac{ie}{\hbar c}\oint A_\mu dx^\mu\right), \tag{2}$$

and *not the phase* (1), is physically meaningful. In other words, the phase (1) contains more information than the phase factor (2). But the additional information is not measurable. This simple point, probably implicitly recognized by many authors, is discussed in Sec. II. It leads to the concept of nonintegrable (i.e., path-dependent) phase factor as the basis of a description of electromagnetism.

This concept has been taken[3] as the basis of the definition of a gauge field. The discussions in Ref. 3, however, centered only on the local properties of gauge fields. To extend the concept to global problems we analyze in Sec. III the field produced by a magnetic monopole. We demonstrate how the quantization of the pole strength, a striking result due to Dirac,[4] is understood in this concept of electromagnetism. The demonstration is closely related to that in the original Dirac paper. Dirac discussed the phase factor of the wave function of an electron (which, among other things, depends on the electron energy). Our emphasis is on the nonintegrable electromagnetic phase factor (which does not depend on such quantities as the energy of the electron).

The monopole discussion leads to the recognition that in general the phase factor (and indeed the vector potential A_μ) can only be properly defined in each of many overlapping regions of space-time. In the overlap of any two regions there exists a gauge transformation relating the phase factors defined for the two regions. This discussion is made more precise in Sec. IV. It leads to the definition of global gauges and global gauge transformations.

In Sec. V generalizations to non-Abelian gauge groups are made. The special cases of SU_2 and SO_3 gauge fields are discussed in Secs. VI and VII. A surprising result is that the monopole types are quite different for SU_2 and SO_3 gauge fields and for electromagnetism.

The mathematics of these results is in fact well known to the mathematicians in *fiber bundle theory*. An identification table of terminologies is given in Sec. V. We should emphasize that our interest in this paper does not lie in the beautiful, deep, and general mathematical development in fiber bundle theory. Rather we are concerned with the necessary *concepts to describe the physics of gauge theories*. It is remarkable that these concepts have already been intensively studied as mathematical constructs.

Section VII discusses a *"gedanken"* generalized Bohm-Aharonov experiment for SU_2 gauge fields.

Unfortunately, the experiment is not feasible unless the mass of the gauge particle vanishes. In the last section we make several remarks.

II. DESCRIPTION OF ELECTROMAGNETISM

The Bohm-Aharonov experiment explores the electromagnetic effect on an electron beam (Fig. 1) in a doubly connected region where the electromagnetic field is zero. As predicted[1] by Aharonov and Bohm, the fringe shift is dependent on the phase factor (2), which is equal to

$$\exp\left(\frac{-ie}{\hbar c}\Omega\right),$$

where Ω is the magnetic flux in the cylinder. Thus two cases a and b for which

$$\Omega_a - \Omega_b = \text{integer} \times (hc/e) \tag{3}$$

give the same interference fringes in the experiment. This we shall state and prove as follows.

Theorem 1: If (3) is satisfied, no experiment outside of the cylinder can differentiate between cases a and b.

Consider first an electron outside of the cylinder. We look for a gauge transformation on the electron wave function ψ_a and the vector potential $(A_\mu)_a$ for case a, which changes them into the corresponding quantities for case b, i.e. we try to find $S = e^{-i\alpha}$ such that

$$S = S_{ab} = (S_{ba})^{-1},$$

$$\psi_b = S^{-1}\psi_a, \quad \text{or } \psi_b = e^{i\alpha}\psi_a, \tag{4}$$

$$(A_\mu)_b = (A_\mu)_a - \frac{i\hbar c}{e} S\frac{\partial S^{-1}}{\partial x^\mu}, \quad \text{or } (A_\mu)_b = (A_\mu)_a + \frac{\hbar c}{e}\frac{\partial \alpha}{\partial x^\mu}. \tag{5}$$

For this gauge transformation to be definable, S must be *single-valued*, but α itself need not be. Now $(A_\mu)_b - (A_\mu)_a$ is curlless; hence (5) can always be solved for α. But it is multiple-valued with an increment of

$$\Delta\alpha = \frac{e}{\hbar c}\oint [(A_\mu)_b - (A_\mu)_a]\, dx^\mu = \frac{e}{\hbar c}(\Omega_b - \Omega_a) \tag{6}$$

every time one goes around the cylinder. If (3) is satisfied, $\Delta\alpha = 2\pi \times$ integer and S is single-valued. Case a and case b outside of the cylinder are then gauge-transformable into each other, and no physically observable effects would differentiate them. The same argument obviously holds if one studies the wave function of an interacting system of particles provided the charges of the particles are all integral multiples of e. Thus we have shown the validity of Theorem 1.

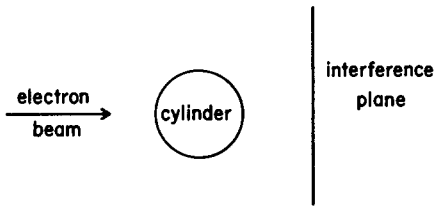


FIG. 1. Bohm-Aharonov experiment (Refs. 1, 2). A magnetic flux is in the cylinder. Outside of the cylinder the field strength $f_{\mu\nu} = 0$.

We conclude: (a) The field strength $f_{\mu\nu}$ underdescribes electromagnetism, i.e., different physical situations in a region may have the same $f_{\mu\nu}$. (b) The phase (1) overdescribes electromagnetism, i.e., different phases in a region may describe the same physical situation. What provides a complete description that is neither too much nor too little is the phase factor (2).

Expression (2) is less easy to use (especially when one makes generalizations to non-Abelian groups) as a fundamental concept than the concept of a phase factor for any path from P to Q

$$\Phi_{QP} = \exp\left(\frac{ie}{\hbar c}\int_P^Q A_\mu dx^\mu\right) \tag{7}$$

provided that an arbitrary gauge transformation

$$\exp\left(\frac{ie}{\hbar c}\int_P^Q A_\mu dx^\mu\right) \rightarrow \exp\left(\frac{ie}{\hbar c}a(Q)\right)\exp\left(\frac{ie}{\hbar c}\int_P^Q A_\mu dx^\mu\right)\exp\left(\frac{-ie}{\hbar c}a(P)\right) \tag{8}$$

does not change the prediction of the outcome of any physical measurements. Following Ref. 3, we shall call the phase factor (7) a nonintegrable (i.e., path-dependent) phase factor.

Electromagnetism is thus the gauge-invariant manifestation of a nonintegrable phase factor. We shall develop this theme further in the next section.

III. FIELD DUE TO A MAGNETIC MONOPOLE

The definition of a nonintegrable phase factor (7) in a general case may present problems. To illustrate the problem, let us study the magnetic monopole field of Dirac.[4] Consider a static magnetic monopole of strength $g \neq 0$ at the origin $\vec{r} = 0$ and take the region R of space-time under consideration to be all space-time minus the origin $\vec{r} = 0$. We shall now show the following:

dered by the loop. Notice that because of Dirac's quantization condition, the phase factor is the same whichever way one chooses the cap provided it does not pass through the point $\vec{r}=0$ (any t).

We have satisfactorily resolved the difficulty mentioned at the beginning of this section, provided Dirac's quantization condition (13) is satisfied. We shall now prove the following.

Theorem 3: If (13) is not satisfied (the above method of resolving the difficulty would not work since) there exists no division of R into overlapping regions $R_a, R_b, R_c, \ldots$ so that condition (i) and (ii) stated above, properly generalized to the case of more than two regions, would hold.

To prove this statement, observe that if such a division is possible, one could generalize (15) and arrive at a satisfactory definition of the phase factor. The phase factor around a loop is then a continuous function of the loop. Take the loop to be a parallel on the sphere r fixed, $t=0$, θ fixed, $\phi=0\to 2\pi$. The phase factor defined by the generalization of (15) is equal to

$$\exp\left[\frac{ie}{\hbar c}\Omega(r,\theta)\right]=\exp\left[\frac{ie}{\hbar c}2\pi g(1-\cos\theta)\right]. \quad (17)$$

This is not equal to unity when $\theta=\pi$, since (13) is assumed to be invalid. Thus we have a contradiction.

Theorem 3 shows that if Dirac's quantization condition (13) is not satisfied, then the field of a magnetic monopole of strength g cannot be taken as a realizable physical situation in R. (Of course, if one excludes the half-line $x=y=0$, $z<0$, or any half-line starting from $\vec{r}=0$ leading to infinity, then it is possible to have any value for g.) This conclusion is the same as Dirac's, but viewed from a somewhat different point of emphasis.

IV. GENERAL DEFINITION OF GAUGE AND GLOBAL GAUGE TRANSFORMATION

Assuming that (13) holds, to round out our concept of a nonintegrable phase factor the question of the flexibility in the choice of the overlapping regions and the flexibility in the choice of A_μ in the regions must be faced. Both of these questions are related to gauge transformations.

Consider a gauge transformation ξ in R_b (ξ will be assumed to be many times differentiable, but not necessarily analytic), resulting in a new potential $(A_\mu)'_b$. We shall illustrate schematically the transformation by "elevating" the region b in Figure 3(a).

One could extend the region b. One could also contract it, provided the whole R remain covered.

One could create a new region by considering a subregion of b as an additional region R_c [Figure 3(b)], and define the gauge transformation connecting them as the identity transformation so that $(A_\mu)_c=(A_\mu)_b$. One can then "elevate" R_c and contract R_b, which results in Fig. 3(c).

Through operations of the kind mentioned in the last three paragraphs, which we shall call *distortions*, we arrive at a large number of possibilities, each with a particular choice of overlapping regions and with a particular choice of gauge transformation from the original $(A_\mu)_a$ or $(A_\mu)_b$ to the new A_μ in each region. Each of such possibilities will be called a *gauge* (or *global gauge*). This definition is a natural generalization of the usual concept, extended to deal with the intricacies of the field of a magnetic monopole.

For each choice of gauge there is a definition of a nonintegrable phase factor for every path. The group condition $\Phi_{C_cBA_a}=\Phi_{C_cB_b}\Phi_{B_bA_a}$ is always satisfied.

Notice that the original gauge we started with was characterized by (a) specifying [in (10)] the regions [R_a and R_b] and (b) specifying the gauge transformation factor (12′) in the overlap (between R_a and R_b). *It does not refer to any specific* A_μ. [A distortion may of course lead to no changes in characterizations (a) and (b). Thus two different gauges may share the same characterizations (a) and (b).] In the case of the monopole field, we had chosen the vector potential to be given by (11). But, in fact, we can attach to this gauge any $(A_\mu)_a$ and $(A_\mu)_b$ provided they are gauge-transformed into each other by (12′) in the region of overlap. (The resultant $f_{\mu\nu}$ is, of course, not a monopole field in general.) *Thus a gauge is a concept not tied to any specific vector potential.* We shall call the process of distortion leading from one gauge to another a *global gauge transformation*. It is also a concept not tied to any specific vector potential. It is a natural generalization of the usual gauge transformation.

The collection of gauges that can be globally gauge-transformed into each other will be said to

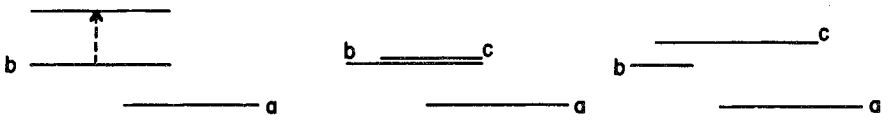


FIG. 3. Distortions allowed in gauge transformation.

TABLE I. Translation of terminology.

Gauge field terminology	Bundle terminology
gauge (or global gauge)	principal coordinate bundle
gauge type	principal fiber bundle
gauge potential b_μ^k	connection on a principal fiber bundle
S_{ba} (see Sec. V)	transition function
phase factor Φ_{QP}	parallel displacement
field strength $f_{\mu\nu}^k$	curvature
source [a] J_μ^K	?
electromagnetism	connection on a $U_1(1)$ bundle
isotopic spin gauge field	connection on a SU_2 bundle
Dirac's monopole quantization	classification of $U_1(1)$ bundle according to first Chern class
electromagnetism without monopole	connection on a trivial $U_1(1)$ bundle
electromagnetism with monopole	connection on a nontrivial $U_1(1)$ bundle

[a] I.e., electric source. This is the generalization (see Ref. 3) of the concept of electric charges and currents.

minimum absolute value is $\pm\frac{1}{2}$. Therefore the minimum "charge" of all physical states can be read off from (24) by taking the 2×2 irreducible representation of X_k:

$$X_k = -\frac{i\sigma_k}{2}\,, \tag{26}$$

where σ_k are the Pauli matrices. Thus

$$\text{minimum "charge"} = \frac{e}{2}\,. \tag{27}$$

The particle of the gauge field belongs to the adjoint representation. Its "charges" are e, 0, and $-e$. Thus

$$\frac{\text{"charge" of gauge particle}}{\text{minimum "charge"}} = 2 \quad \text{for } SU_2\,. \tag{28}$$

We shall now try to define a Dirac monopole field as a special SU_2 field along only one isospin direction $k=3$, i.e., we define

$$b_\mu^1 = b_\mu^2 = 0\,, \quad b_\mu^3 = A_\mu\,, \tag{29}$$

where A_μ is given in the two regions (10) by (11). In the overlapping region, transformation factor S of (12) and (14) now becomes

$$S_{ab} = \exp\left(-\frac{2ge}{\hbar c}\phi X_3\right) \tag{30}$$

by replacement (25). This is single-valued if and only if the quantization condition

$$\frac{eg}{\hbar c} = \text{integer} = D \tag{31}$$

is satisfied because for SU_2

$$\exp(4\pi X_3) = 1\,, \quad \exp(2\pi X_3) \neq 1\,,$$

which follows from the existence of half-integral representations such as (26).

The phase factor (30) describes a great circle, wound D times, on the manifold of SU_2 when ϕ varies from $0 \to 2\pi$. Such a circle can be continuously shrunk to the identity element, in contrast with the situation for electromagnetism. Thus, by a global gauge transformation S may be changed to $S'=1$, and the two regions a and b after the global gauge transformation can be *fused into one single region*. The gauge potential b_μ^k is then defined *everywhere* in R as a single region. Thus we have the following theorem.

Theorem 9: For the SU_2 gauge group, the gauges $\mathcal{G}_D$ for different D can be transformed into each other by global gauge transformations. The different monopole fields are therefore of the same type.

We shall only exhibit the global transformation for the case $\mathcal{G}_{-1}$ for which

$$S_{ba} = \exp(-2\phi X_3)\,, \tag{32}$$

$$\frac{e}{\hbar c} = \frac{-1}{g}\,. \tag{33}$$

The gauge transformations we shall seek are illustrated in Fig. 5. We shall choose

$$\xi = \exp[\theta(X_1\sin\phi - X_2\cos\phi)]\,, \tag{34}$$

$$\eta = \exp[(\pi-\theta)(X_1\sin\phi - X_2\cos\phi)]\exp(\pi X_2)\,. \tag{35}$$

It is easy to see that ξ is analytic in the coordinates x^μ at all points in R_a. (One only has to verify this statement at $\theta=0$, which is easily done.) Similarly η is analytic in R_b. ξ and η are therefore allowed gauge transformations in, respectively, R_a and R_b.

it does not satisfy the Bianchi identity[3] at the origin. Thus, although solution (12a) of Ref. 9 is (electrically) sourceless at all points, including the origin, it is not a proper gauge field at the origin, a fact we did not realize before. All three solutions, (12a), (12d), and (12e), are, of course, of the same gauge type.

(c) In Sec. II it was emphasized that $f_{\mu\nu}$ underdescribes electromagnetism because of the Bohm-Aharonov experiment which involves a doubly connected space region. For non-Abelian cases, the field strength $f^k_{\mu\nu}$ underdescribes the gauge field even in a singly connected region. An example of this underdescription was given in Ref. 13.

(d) For the region of space-time outside of the cylinder of Fig. 1 there is only one gauge type. All electromagnetic fields in the region can be continuously distorted into each other by the movement of electric charges and currents inside and outside the cylinder.

(e) The phase factor for the group U_1 is the phase factor of the algebra of complex numbers. It is perhaps not accidental that such a phase factor provides the basis for the description of a physically realized gauge field—electromagnetism. Now the only possible more complicated division algebra is the *algebra of quaternions*. The phase factors of the quaternions form the group SO_3. It is tempting to speculate that such a phase factor provides the basis for the description of a physically realized gauge field—the SU_2 gauge field. Speculation about the possible relationship between quaternions and isospin has been made before.[14] Such speculations were, however, not made with reference to gauge fields. If one believes that gauge fields give the underlying basis for strong and/or weak interactions, then the fact that gauge fields are fundamentally *phase factors* adds weight to the speculation that quaternion algebra is the real basis of isospin invariance.

(f) It is a widely held view among mathematicians that the fiber bundle is a natural geometrical concept.[15] Since gauge fields, including in particular the electromagnetic field, are fiber bundles, *all gauge fields are thus based on geometry*.[16] To us it is remarkable that a geometrical concept formulated without reference to physics should turn out to be exactly the basis of one, and indeed maybe all, of the fundamental interactions of the physical world.

ACKNOWLEDGMENTS

It is a pleasure to thank Professor Shiing-shen Chern for correspondence and discussions. We are especially indebted to Professor J. Simons, whose lectures and patient explanations have revealed to us glimpses of the beauty of the mathematics of fiber bundles.

While we were making corrections on the draft of this paper, a report on the experimental discovery of a magnetic monopole[17] reached us.

Additional references to fiber bundles, monopoles and quaternions are given in footnote 18.

---

*Work supported in part by the U. S. ERDA under Contract No. AT(11-1)-3227.

†Work supported in part by the National Science Foundation under Grant No. MPS74-13208 A01.

[1] Y. Aharonov and D. Bohm, Phys. Rev. 115, 485 (1959). See also W. Ehrenberg and R. E. Siday, Proc. Phys. Soc. London B62, 8 (1949).

[2] R. G. Chambers, Phys. Rev. Lett. 5, 3 (1960).

[3] Chen Ning Yang, Phys. Rev. Lett. 33, 445 (1974). This paper introduced the formulation of gauge fields in terms of the concept of nonintegrable phase factors. The differential formulation of gauge fields for Abelian groups was first discussed by H. Weyl, Z. Phys. 56, 330 (1929); for non-Abelian groups it was first discussed by Chen Ning Yang and Robert L. Mills, Phys. Rev. 96, 191 (1954). See also S. Mandelstam, Ann. Phys. (N.Y.) 19, 1 (1962); 19, 25 (1962); I. Białynicki-Birula, Bull. Acad. Pol. Sci., Ser. Sci. Math. Astron. Phys. 11, 135 (1963); N. Cabibbo and E. Ferrari, Nuovo Cimento 23, 1146 (1962); R. J. Finkelstein, Rev. Mod. Phys. 36, 632 (1964); N. Christ, Phys. Rev. Lett. 34, 355 (1975); and A. Trautman, in *The Physicist's Conception of Nature*, edited by J. Mehra (Reidel, Boston, 1973), p. 179.

[4] P. A. M. Dirac, Proc. R. Soc. London A133, 60 (1931). Since this brilliant work of Dirac, there have been several hundred papers on the magnetic monopole. For a listing of papers until 1970, see the bibliography by D. M. Stevens, Virginia Polytechnic Institute Report No. VPI-EPP-70-6, 1970 (unpublished).

[5] See J. Milnor and J. Stasheff, *Characteristic Classes* (Princeton Univ. Press, Princeton, N.J., 1974); C. B. Allendoerfer and A. Weil, Trans. Am. Math. Soc. 53, 101 (1943); Shiing-shen Chern, Ann. Math. 45, 747 (1944). See also H. Weyl, Amer. J. Math. 61, 461 (1939), and Ref. 6 below.

[6] There are many books on fiber bundles. See e.g., N. Steenrod, *The Topology of Fibre Bundles* (Princeton Univ. Press, Princeton, N. J., 1951). For connection, see, e.g., S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, Vol. I-1963, Vol. II-1969).

[7] The notation here is the same as that in Ref. 3, except for the normalization factor $e/\hbar c$ which was absorbed into b in Ref. 3. To avoid confusion with the azimuthal angle, we write Φ for the ϕ of Ref. 3. Notice that

$$\Phi_{(A+dx)A} = I - \frac{e}{\hbar c} b^k_\mu(x) X_k dx^\mu .$$