The Aharonov-Bohm effect

Joseph Samuel

Quantum mechanics teaches us that matter consists of waves. Interference of matter waves gives rise to delicate effects best illustrated by the double slit experiment. Aharonov and Bohm showed that the interference pattern of electrons in a multiply connected region can be influenced by magnetic fields outside that region. This surprising effect (now called the Aharonov-Bohm effect) has been measured in the laboratory. The process of understanding and coming to terms with this effect has deepened our understanding of both quantum mechanics and electromagnetism. This paper gives an elementary account of the Aharonov-Bohm effect.

In classical physics, electromagnetic effects are completely described[1] by the electric (**E**) and magnetic (**B**) fields. These fields can be directly measured by their effect on a test charge. Further, the behaviour of the test charge can be completely determined from a knowledge of these fields. However, it is usual and convenient[1] to introduce potentials Φ and **A**. The fields then emerge as derivatives of the potentials. For example, $\mathbf{B} = \nabla \times \mathbf{A}$. The advantage of introducing potentials is that some Maxwell equations (the homogeneous ones, like $\nabla \cdot \mathbf{B} = 0$) are automatically solved. Potentials are not directly measurable quantities and only their derivatives **E** and **B** have physical significance. Potentials which give rise to the same electric and magnetic fields are classically indistinguishable. In classical physics the introduction of potentials is a mathematical convenience and not a physical or logical necessity.

In quantum mechanics the situation is different. In order to couple the wave function $\psi(\mathbf{x}, t)$ of a test charge to the electromagnetic field, it is necessary to introduce vector potentials. However, the coupling is such that if one performs a 'gauge transformation', i.e. changes the potentials and the wave function ψ as follows:

$$\mathbf{A}'(\mathbf{x}, t) = \mathbf{A}(\mathbf{x}, t) + \nabla\chi(\mathbf{x}, t),$$

$$\Phi'(\mathbf{x}, t) = \Phi(\mathbf{x}, t) + \frac{1}{c}\frac{\partial \chi}{\partial t}(\mathbf{x}, t),$$

$$\psi'(\mathbf{x}, t) = \psi(\mathbf{x}, t)\exp\left[\frac{-ie}{\hbar c}\chi(\mathbf{x}, t)\right],$$

all measurable quantities are unchanged. Such a transformation does not alter the fields (which are, of course, measurable). And conversely, given the fields in a *simply connected* region, one can determine the potentials up to a gauge transformation. Since measurements cannot distinguish between 'gauge-related' potentials, one might expect that even in quantum mechanics, a knowledge of the fields in a region should enable one to completely determine the behaviour of a test charge. The truth is more subtle – and more interesting. In a multiply connected region, potentials which are not related by a gauge transformation can give rise to the *same* **E** and **B** fields. The work of Aharonov and Bohm[2] shows that one can experimentally distinguish between such potentials. The potentials contain physically measurable information which is not contained in the **E** and **B** fields. Thus, electromagnetic potentials play an indispensable role in quantum physics.

Consider the following experiment[2]: A beam of electrons is split into two and allowed to interfere on a screen (see Figure 1). The setup is exactly the same as in the double slit experiment[3], save for the solenoid S between the two interfering beams. Passing a current through the solenoid produces a magnetic field (normal to the plane of the figure) which we suppose, is contained entirely within the solenoid. Electrons are prevented from entering the solenoid. Since the electrons never directly experience the magnetic field, one might conclude that the interference pattern on the screen is unaffected by the presence of the field. Surprisingly, this conclusion is false[2].

To see that the field does affect the fringe pattern on the screen, let us compute the probability that an electron starting at $\mathbf{x}_1$ will arrive at a point $\mathbf{x}_2$ on the screen. According to the Feynman path integral approach[4], the probability amplitude for the particle to go from $\mathbf{x}_1$ to $\mathbf{x}_2$ is given by a sum over all possible paths connecting these points, weighted by $[\exp(i/\hbar)S_{\text{cl}}]$, where S_{cl} is the classical action along the path.

The classical action for a charged particle in an external vector potential **A** is

$$S_{\text{cl}} = S_{\text{cl}}^{o} + \frac{e}{c}\int \mathbf{v} \cdot \mathbf{A}dt$$

$$= S_{\text{cl}}^{o} + \frac{e}{c}\int \mathbf{A} \cdot d\mathbf{x},$$

where $S_{\text{cl}}^{o} = \frac{1}{2}\int mv^2 dt$ is the classical action of the free particle. The principal contribution to the probability amplitude will come from classes of paths that pass to the left and right of the solenoid. (There are also paths that wind around the solenoid before arriving at $\mathbf{x}_2$. In the semiclassical approximation these would contribute negligibly.) The amplitude to go from $\mathbf{x}_1$ to $\mathbf{x}_2$ is therefore

$$K(\mathbf{x}_1 t_1, \mathbf{x}_2 t_2) = K_L \exp\left[\frac{ie}{\hbar c}\int_L \mathbf{A} \cdot d\mathbf{x}\right]$$

$$+ K_R \exp\left[\frac{ie}{\hbar c}\int_R \mathbf{A} \cdot d\mathbf{x}\right],$$

where $K_L(K_R)$ is the sum of $[\exp iS_{\text{cl}}^{o}]$ over all paths that pass to the left (right) of the solenoid. Squaring the amplitude to find the probability of arrival at $\mathbf{x}_2$, we find

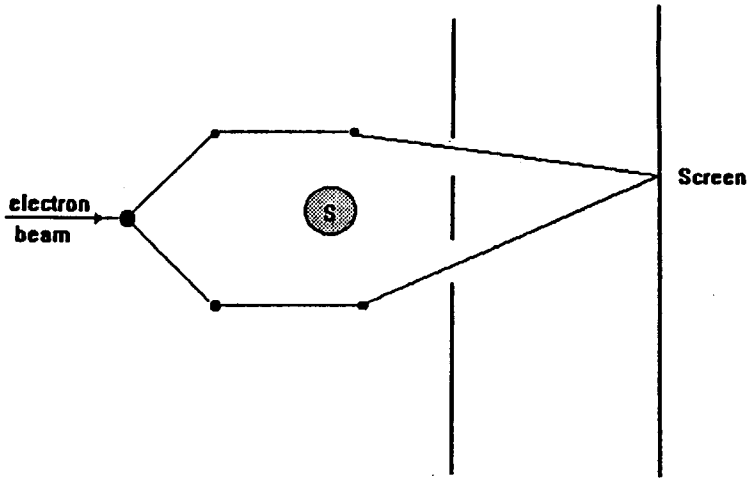

Figure 1. A schematic illustration of the experimental set-up for the Aharonov–Bohm effect. It differs from the usual double slit experiment only in the presence of the solenoid *S* between the two interfering beams.

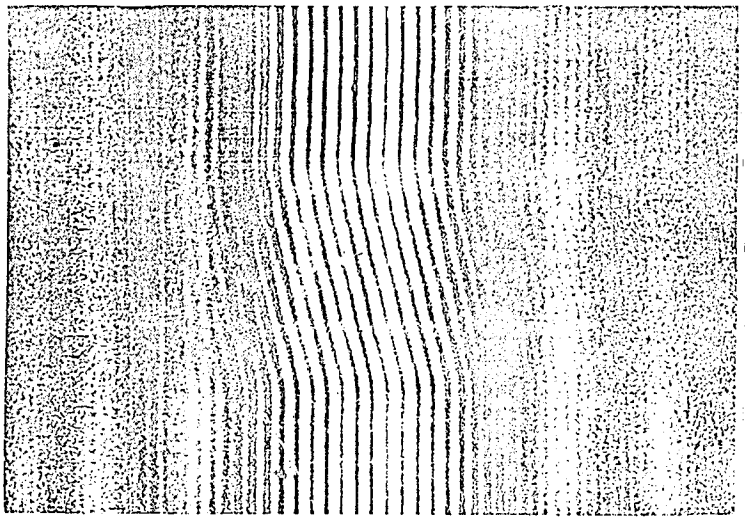

Figure 2. Dynamically recorded fringe system showing the displacement of the fringes caused by a continuously varying magnetic flux between the split beams. From Missiroli *et al.*[6]

$$P(\mathbf{x}_1 t_1 \mathbf{x}_2 t_2) = |K|^2 = |K_L|^2 + |K_R|^2 + 2\,\mathrm{Re}\left[K_L^* K_R \exp\frac{2\pi i\phi}{\phi_o}\right], \qquad (1)$$

where $\varphi_o = hc/e$ and $\varphi = \oint \mathbf{A}\cdot d\mathbf{x}$ is the total flux through the solenoid. Thus, the probability of arrival at $\mathbf{x}_2$ does depend on ϕ. A change in ϕ would produce a shift in the interference fringes. The calculation sketched above was approximate. An exact calculation yields the same qualitative result: although the electron is excluded from the location of the magnetic field, its behaviour is changed by the field. This discussion suggests that in quantum mechanics, the potentials *do* have physical significance. The fields (**E** and **B**) in the region outside the solenoid incompletely describe the effects of electromagnetism in quantum physics.

The experiment described above was performed shortly afterwards by Chambers[5] and the theoretical prediction confirmed. In practice, it is hard to obtain large separations between the electron beams, so Chambers used a magnetic whisker rather than a solenoid to produce a magnetic field. Since then the experiment has been done with more and more sophistication. For a review see ref. 6. Figure 2 shows a photograph of the interference fringes changing as ϕ is increased (along the vertical axis). The envelope of the fine interference fringes is the diffraction pattern of the slits. Notice that the envelope does not move when the ϕ is changed. This shows that the individual beams do not experience any deflection when the magnetic field in the solenoid is varied.

It is appropriate to mention an earlier piece of work[7] that arrived at the same conclusion as ref. 2, but went completely unnoticed. Ehrenberg and Siday[7], in their studies of electron microscopy, had the main conclusion of ref. 2, quite explicitly stated in their paper. But, to use the delicately phrased words of Chambers[5] 'they did not sufficiently emphasize the remarkable nature of their result'. In less delicate words, it is not enough to make a remarkable discovery to get credit for it, one must also make a noise about it!

In some accounts of the Aharonov-Bohm effect (see for instance, ref. 8) one gets the impression that the wave function 'splits into two parts':

$$\psi = \psi_L + \psi_R,$$

each passing on either side of the solenoid and acquiring different phases. This is a physically incorrect description of the effect. There is only one wavefunction, which furthermore is a single-valued function[9] on the configuration space. One is, of course, at liberty to split the wave function into two parts, but there is no reason why each part should separately obey the Schrödinger equation. The Feynman path integral derivation given above avoids this pitfall. Another way to derive the Aharonov-Bohm effect is to solve the Schrödinger equation for the wave function in the external potential **A**, which is what Aharonov and Bohm did.

If one computes the probability exactly (and not approximately as we did above), one finds[10] that the probability amplitude in equation (1) is an infinite sum over all paths that wind around the solenoid. The probability then becomes an infinite sum

$$P(\phi) = \Sigma_n P_n \exp\left[i2\pi n \frac{\phi}{\phi_0}\right].$$

The dependence of the probability on ϕ is not sinusoidal any more, but acquires higher harmonics. Of course, $P(\phi)$ is still periodic in ϕ with period ϕ_0. One cannot measure the flux ϕ in the solenoid through interference effects outside it. One can only determine its fractional part modulo ϕ_0.

The impact of ref. 2 has been felt in many diverse areas of physics. We mention two here for illustration. The work leads us to ask the question: What is a complete description of the electromagnetic field? It appears from the Aharonov–Bohm effect that the (electric and magnetic) fields do not have enough information and the potentials have too much (remember they can be altered by gauge transformations, which do not change physics). What about line integrals like $\oint \mathbf{A} \cdot d\mathbf{x}$? These still contain too much information. All that appears in the Aharonov–Bohm effect is the phase $[\exp 2\pi i(\phi/\phi_0)]$. All effects are periodic in ϕ with period ϕ_0. Indeed, one can make a gauge transformation in the region outside the solenoid to change ϕ by this amount. Thus, a complete description[11] of the electromagnetic field is via 'path-dependent phase factors' or Wilson loops $-[\exp \frac{ie}{\hbar c} \oint \mathbf{A} \cdot d\mathbf{x}]$. Wilson loops have played a role in our understanding of other gauge theories.

There are two ways of looking at the problem of a particle moving in the region around a solenoid. One can regard the configuration space of the system to be a plane, with an infinite potential used to keep the particle away from the solenoid. With a slight shift of emphasis, one can regard the configuration space of the system as the plane minus a disc, which is topologically distinct from the plane. In the second point of view, the Aharonov–Bohm phase appears as a quantization ambiguity[12]. If one studies quantization on spaces which are multiply connected (like the plane with a hole) one finds that the quantum theory is ambiguous. For each 'hole' in the configuration space, one finds a possible Aharonov–Bohm phase. In the above elementary example, the second point of view is a lofty way of looking at an easy problem. However, in the quantum theory of fields, this is the only point of view available. The configuration space has 'holes' in it, which cannot be 'filled in'. This leads to alternative quantizations of the same classical field theory. Examples are the emergence of θ vacuua in quantum chromodynamics and the fractional spin of topological geons in quantum gravity[13].

The Aharonov–Bohm paper was a conceptually important step forward in our understanding of gauge fields and quantum mechanics. As is evident from the two examples quoted above, it has left its mark on our thinking. What was 'new' about this paper was not a new technique or a new calculation. It was a new perspective.

1. Jackson, J. D., *Classical Electrodynamics*, John Wiley, New York, 1962.
2. Aharonov, Y. and Bohm, D., *Phys. Rev.*, 1959, **115**, 485.
3. Feynman, R. P., Leighton, B. B. and Sands, M., *The Feynman Lectures on Physics*. Addison-Wesley, Reading, Mass. Vol. 3, 1965.
4. Feynman, R. P. and Hibbs, A. R., *Path Integrals and Quantum Mechanics*, McGraw-Hill, New York, 1965.
5. Chambers, R. G., *Phys. Rev. Lett.*, 1960, **5**, 3.
6. Missiroli, G. F., Pozzi, G. and Valdrè, U., *J. Phys.*, 1981, **E14**, 649.
7. Ehrenberg, W. and Siday, R. E., *Proc. Phys. Soc.*, 1949, **B62**.
8. Sakurai, J. J., *Advanced Quantum Mechanics*, Addison-Wesley, Reading Mass. 1967.
9. It is possible to correctly formalise the intuitive argument that is intended in ref. 8 by using two patches to cover the configuration space. The wave function is defined locally in each patch and is mathematically, a *section*, not an ordinary function[11]. This is a more advanced point of view that we do not pursue here.
10. Morandi, G. and Menossi, E., *Eur. J. Phys.*, 1985, **5**.
11. Wu, T. T. and Yang, C. N., *Phys. Rev.*, 1975, **D12**, 3845.
12. Schulman, L. S., *Phys. Rev.*, 1968, **176**, 1558; Schulman, L. S., *Techniques and applications of Path Integrals*, John Wiley, New York; 1981, Laidlaw, M. G. G. and Dewitt, C. M., *Phys. Rev.*, 1971, **D3**, 1375.
13. Friedman, J. L. and Sorkin, R. D., *Phys. Rev. Lett.*, 1980, **44**, 1100.

Joseph Samuel is in the Raman Research Institute, Bangalore 560 080, India