

The Aharonov-Bohm Effect and Weyl's Gauge Theory

William O. Straub, PhD
Pasadena, California
April 4, 2010

The presence of an electric or magnetic field is easy to detect—you just introduce a charged particle (fixed or moving at some initial velocity) and watch its behavior. If it moves or deflects, then there's an electromagnetic field nearby; the greater the deflection, the stronger the field. The electromagnetic field is, course, composed of the electric field \vec{E} and the magnetic field \vec{B} , either singly or in combination. But Maxwell's equations reveal a deeper phenomenon associated with these fields, one that was not experimentally demonstrated until relatively recently. It is the *potential*, and it comes in two forms—the 3-component *vector potential* A_i and the scalar potential φ , both of which are generally time-dependent functions of space. These functions appear seemingly out of nowhere from a consideration of the differential forms of two of Maxwell's equations: $\vec{\nabla} \cdot \vec{B} = 0$, which can be equivalently expressed by $\vec{B} = \vec{\nabla} \times \vec{A}$, and $\vec{\nabla} \times \vec{E} = -\partial \vec{B} / \partial x^0$, which leads to $\vec{E} = -\vec{\nabla} \varphi - \partial \vec{A} / \partial x^0$. The four potential quantities are joined into what is called the *four-potential* $A_\mu (= \varphi, A_1, A_2, A_3)$ which, unlike either \vec{E} or \vec{B} , transforms like a covariant Lorentz 4-vector. It is in this sense that the 4-potential is more fundamental than the electric and magnetic fields.

Invisibility of the Vector Potential

The scalar potential φ is familiar from high school electromagnetism, where it is commonly identified with the static potential difference or voltage V . But the vector potential A_i normally does not make its appearance until much later; indeed, many undergraduate science courses ignore it completely. Perhaps much of the reason for this stems from a preoccupation with the electric and magnetic fields themselves, which are usually the only “solutions” sought in undergraduate physics and engineering classes. When A_i is finally introduced, it's normally because some problems are easier to solve by first calculating the potential, then using $\vec{B} = \vec{\nabla} \times \vec{A}$ (in the same way that the scalar potential is used calculate \vec{E}).

But there are other reasons why A_i doesn't get much respect. One has to do with the fact that, unlike the magnetic field, it's almost impossible to observe. Another involves the fact that, like the scalar potential, the vector potential does not have a unique representation in any given application. For example, the potential difference $\Delta V = V - V_0$ of a charged sphere depends on how the *ground potential* V_0 is defined.; like potential energy, the only thing that's ever measured is the difference between two energy levels, as there is no *absolute* reference. The same is true for A_i . To see this, we consider a simultaneous change in φ and A_i given by the *gauge transformation*

$$\begin{aligned}\varphi' &= \varphi - \frac{\partial \lambda}{\partial x^0} \\ \vec{A}' &= \vec{A} + \vec{\nabla} \lambda\end{aligned}$$

where $\lambda(x, t)$ is an *arbitrary* function of the space-time coordinates. Substituting these new potentials into the expressions $\vec{B}' = \vec{\nabla} \times \vec{A}'$ and $\vec{E}' = -\vec{\nabla} \varphi' - \partial \vec{A}' / \partial x^0$ shows that

$$\begin{aligned}E' &= E \\ B' &= B\end{aligned}$$

That is, the electric and magnetic fields remain unchanged. By this, we say that Maxwell's equations are *gauge invariant*. Because of the arbitrariness of the gauge parameter $\lambda(x)$, the 4-potential A_μ has no unique definition. Very often, a clever choice of the gauge parameter can be used to simplify the calculation of \vec{E} and \vec{B} . Several examples are the so-called *Coulomb gauge* and *Lorentz gauge*, but we will not be needing them in this simple discussion.

For many years after Maxwell first set down his famous equations, the gauge property of the potentials was looked upon strictly as a useful computational device, and little real meaning was ascribed to the potentials themselves, the vector potential in particular. Indeed, because it could not actually be seen or detected, it took a back seat to

the magnetic field, which was considered the only real field. To give a concrete example of this, consider an ideal solenoid of very long (essentially infinite) length. When a current is sent through the coiled wire, a strong magnetic field is set up *within* the solenoid, while the magnetic field external to the coil is zero. If an observer now directs a moving particle of charge q outside the coil, the particle will not experience any deflection (recall the Lorentz force law $\vec{F} = q/c\vec{v} \times \vec{B}$), because the magnetic field is zero there. But the vector potential exterior to the solenoid is *not* zero; in fact, there is one non-zero component about the cylindrical axis of the solenoid given by

$$A_\phi = \frac{Br_s^2}{2r}$$

where r_s is the radius of the solenoid and r is the radial distance from the axis. The observer sees the particle proceeding on its merry way, and concludes that no field is present. The vector potential thus escapes detection.

Weyl's Gauge Theory

In 1929, the German mathematical physicist Hermann Weyl, utilizing an earlier idea of his from 1918, proposed that the vector potential could in principle affect the wave function $\psi(x, t)$ of a quantum-mechanical particle or system. This proposal led Weyl to what could arguably be called the most profound and beautiful mathematical symmetry known to exist in nature, that of *gauge invariance*. Weyl noticed that a gauge transformation of the electromagnetic field would induce a similar change in the wave function ψ and its complex conjugate ψ^* as expressed by

$$\begin{aligned}\psi' &= e^{i\lambda}\psi \\ \psi^{*'} &= e^{-i\lambda}\psi^*\end{aligned}$$

In quantum mechanics, real and observable quantities are expressed as products of functions and their conjugates, so that

$$\psi^{*'}\psi' = \psi^{*'}(e^{-i\lambda}e^{i\lambda})\psi' = \psi^*\psi$$

Thus, the information that a gauge transformation (perhaps more properly called a *phase* transformation in this case) has acted on the wave function is hidden. Weyl discovered that this gauge symmetry is due to the presence of an electromagnetic field; he also discovered that the invariance of electric charge is a consequence of this symmetry.

The great German mathematician Emmy Noether famously discovered (also in 1918) that every mathematical symmetry in nature (typically an invariance resulting from the arbitrary change of some quantity in an action lagrangian) is intimately associated with a *conservation law*. For example, the symmetry of *translation*—the invariance of physical laws that doesn't change when you go from Pasadena, California to Bristol, England—necessarily requires that linear momentum be a conserved quantity. But what is often not associated with Noether's theorem is the fact that such symmetries are *hidden* from plain sight in some way. An arbitrary change of phase in the wave function, for example, cannot be seen or measured, but it manifests itself in the very observable fact that electric charge is conserved—an observation that, until Weyl and Noether came along—had to be taken as an *a priori* fact.

In the seventy or so years since Maxwell's equations appeared, the existence of the vector potential was viewed strictly as a mathematical oddity that had no real physical meaning. Why? Because it could not be seen or measured experimentally! Can we therefore associate some conservation law with the potential? The answer is yes and no—its true significance seems to lie only in the fact that the gauge transformation property of the potentials enables the principle of gauge invariance in modern quantum theory. But that principle leads to conservation laws, so perhaps it is only a matter of semantics.

The Aharonov-Bohm Idea

It was not until 1959 that a method was devised for demonstrating the reality of the scalar and vector potentials. Actually, it was little more than a thought experiment, because it could only be demonstrated mathematically, not experimentally. David Bohm and his graduate student Yakir Aharonov, both at the University of Bristol at

the time, showed theoretically that the vector potential could be detected when applied to the famous double-slit experiment. The proposed setup, shown in Figure 1, consists of a charged particle source, double slit, and detector screen. By firing the particles one at a time at the slit, each particle's wave function interferes with itself, resulting in the usual interference pattern at the detector. Aharonov's and Bohm's *ansatz* was to imagine placing a tiny solenoid immediately behind the slit, where presumably the solenoid's external vector potential would induce a measurable phase shift in the pattern.

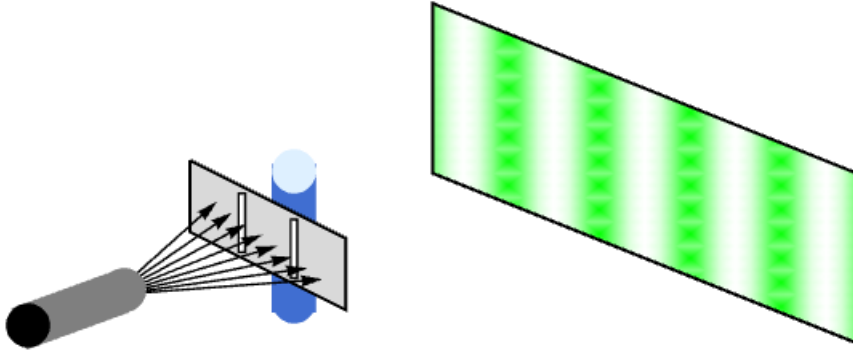


Figure 1. Experimental setup with solenoid turned off.

Path Integral Approach to the Aharonov-Bohm Effect

The mathematical reasoning we will use for the predicted shift is equivalent but different than that employed by Aharonov and Bohm, but it will provide additional insight into some other ideas that Weyl had proposed. It will also give us the chance to use a particularly powerful quantum tool, which is the *path integral* approach originally pioneered by physicist Richard Feynman in his 1942 doctoral dissertation. The path integral provides a particularly clear and elegant solution to the thought experiment of Aharonov and Bohm.

I will not enter into a detailed discussion of the path integral (it's available elsewhere on my website), so I will just touch on it here. The path integral is a mathematical way of describing the probability amplitude that a particle will go from Point A to Point B in some finite period of time. It says that the particle can traverse *any* intermediate path on its way to Point B, over *any* time period. This means that, after leaving Point A, the particle can travel out to the Andromeda Galaxy, at which point it can then travel back in time a billion years to visit Mars. It can literally execute any of an *infinite* number of paths, crazy or otherwise, until it arrives at Point B. More amazingly, each path that the particle can take is just as important and logical as any other, *including* the classical, straight path from A to B. Does the particle really take all these paths? Mathematically, the answer is a definite yes. In reality, we can never really know just what the particle does. All we know for sure is that it *can*, if it wants to. But for each possible path there is a corresponding probability amplitude, and these amplitudes can interfere with one another constructively or destructively. The crazy paths tend to be the ones that get canceled out by destructive interference, while the logical, "classical" paths tend to reinforce one another. While the path integral is necessarily infinite-dimensional, it can in principle be calculated in closed form; for simple problems, such as the free particle and the harmonic oscillator, the calculation is straightforward and agrees perfectly with the results of classical and quantum physics.

Path integrals can also be applied to *fields*, in which case they give the probability amplitude that a field will propagate or transition from one field to another over a specified period of time. But the Aharonov-Bohm thought experiment described here will only be concerned with charged particles (like electrons) going from Point A (their source) through the double slit to Point B (the detector).

Here is the path integral for a particle going from Point A to Point B when the slits are not present:

$$I = \int \mathcal{D}x(t) \exp[iS(x, t)/\hbar]$$

Here, I represents the probability amplitude for the overall process; it is a complex number, and its conjugate square represents the real probability that the particle will leave A and arrive at B. The quantity $\mathcal{D}x$ is shorthand

for the infinite sequence $dx_1 dx_2 dx_3 \dots$, where each $x = x(t)$ represents a given path, and the single integral sign represents an infinite number of them. The quantity S in the exponential term is the *action*, which is defined as the integral of the lagrangian \mathcal{L} over time. Since we'll be considering the motion of classical charged particles moving with small velocities, we'll employ the non-relativistic form of the lagrangian. You may recall from your physics classes that the lagrangian for a free particle of mass m is just $\mathcal{L}_0 = 1/2 m(dx/dt)^2$. For a free particle having a charge q in the presence of a vector potential, it becomes

$$\mathcal{L} = \mathcal{L}_0 + \frac{q}{c} A_i \frac{dx^i}{dt}$$

(Note that I have set the scalar potential term $A_0 = \varphi$ to zero, because this field is absent in the Aharonov-Bohm setup.) The path now integral looks like

$$I = \int \mathcal{D}x(t) \exp\left(\frac{iS_0}{\hbar}\right) \exp\left(\frac{iq}{\hbar c} \int_0^t A_i \frac{dx^i}{dt} dt\right)$$

where $S_0 = \int_0^t \mathcal{L}_0 dt$. Lastly, note that the integral over time of the vector potential term becomes a line integral over space:

$$\int_0^t A_i \frac{dx^i}{dt} dt = \int_A^B A_i dx^i$$

so we have, finally,

$$I = \int \mathcal{D}x(t) \exp\left(\frac{iS_0}{\hbar}\right) \exp\left(\frac{iq}{\hbar c} \int_A^B A_i dx^i\right)$$

Armed now with the path integral (or at least an inkling of what it represents), let us proceed to the Aharonov-Bohm experiment itself.

Derivation of the Aharonov-Bohm Effect

Let us consider the classic double-slit experiment, consisting of a source of identical, quantum-sized particles each having a charge q (electrons or protons will do nicely), an impenetrable screen with two closely-spaced, narrow slits, and detector. In practice, the slit widths will be on the order of several microns, separated by a section of screen of comparable dimension. Such small distances are necessary in view of the quantum nature of the double slit experiment itself. In addition, we place a solenoid behind the slit separation. The solenoid itself must be extremely small, and it must be shielded in such a way as to prevent physical interaction with the charged particles.

We start with the solenoid in place but with the current turned off, so $A_i = 0$. We now start firing charged particles at the screen, at a the rate of fire that prevents the particles from interacting with one another. Some of the particles will pass through Slit 1 and some through Slit 2, but in general each particle will seem to go through *both* slits. This is, after all, the nature of the quantum weirdness of the double slit experiment! We can associate a path integral for each slit:

$$I_1 = \int_1 \mathcal{D}x \exp\left(\frac{iS_0}{\hbar}\right)$$

$$I_2 = \int_2 \mathcal{D}x \exp\left(\frac{iS_0}{\hbar}\right)$$

where the subscripts refer to the paths through the two slits. The paths are constrained to go through their respective slits, but they're free-particle path integrals, and we know the solution to such integrals. For an unconstrained particle going from Point A to Point B, it's

$$I = \left(\frac{m}{2\pi i \hbar t}\right)^{1/2} \exp\left[\frac{im(B-A)^2}{2t}\right]$$

(The exponential term is a constant phase factor that will disappear when we take the conjugate square of I , so we'll ignore it.) In view of this, I_1 and I_2 can differ only in phase, and so we write

$$I_1 = \left(\frac{m}{2\pi i\hbar t}\right)^{1/2} \exp\left(\frac{i\theta_1}{\hbar}\right)$$

$$I_2 = \left(\frac{m}{2\pi i\hbar t}\right)^{1/2} \exp\left(\frac{i\theta_2}{\hbar}\right)$$

where the quantities θ_1 and θ_2 are phase constants reflecting the path constraints. The total path integral is then $I = I_1 + I_2$, which we can write as

$$I = \left(\frac{m}{2\pi i\hbar t}\right)^{1/2} \exp\left(\frac{i\theta_1}{\hbar}\right) \left[1 + \exp\left(\frac{i(\theta_2 - \theta_1)}{\hbar}\right)\right]$$

The quantity $(\theta_2 - \theta_1)/\hbar \equiv \Delta$ represents the *phase difference* between the combined paths; the detector at any given point will see constructive interference when $\Delta = 2n\pi$ and destructive interference when $\Delta = (2n + 1)\pi$, where n is an integer. The path integral I is a probability amplitude, and to get the probability we have to take the conjugate square of this quantity. It is not difficult to show that

$$I^*I = |I|^2 = \frac{2m}{\pi\hbar t} \cos^2 \frac{1}{2}\Delta \quad (1)$$

Thus, the path integral approach to the double-slit experiment explains the sinusoidal interference pattern seen at the detector, and this result has been verified by experiment.

We now start the current running into the solenoid. What effect, if any, can we expect? The presence of the solenoid induces a longitudinal magnetic field inside the solenoid, but this magnetic field is restricted to its *interior*; thus, there is no external magnetic field to affect the charged particles as they pass through the slits. As far as they're concerned, they see the same vacuum as they did when the solenoid was turned off. But now there's an additional term in the path integral, and we can't use the free-particle solution. Or can we? From elementary electrodynamics, we know that a line integral like $\int A_i dx^i$ is independent of the path taken from one point to another, so we can take this term out of the path integral and treat it as a phase coefficient:

$$\int \mathcal{D}x(t) \exp\left(\frac{iS}{\hbar}\right) \exp\left(\frac{iq}{\hbar c} \int A_i dx^i\right) = \exp\left(\frac{iq}{\hbar c} \int A_i dx^i\right) \int \mathcal{D}x(t) \exp\left(\frac{iS}{\hbar}\right)$$

$$= \left(\frac{m}{2\pi i\hbar t}\right)^{1/2} \exp\left(\frac{i\theta}{\hbar}\right) \exp\left(\frac{iq}{\hbar c} \int A_i dx^i\right)$$

(Note the use of a generic θ in this expression.) As before, the total path integral is $I = I_1 + I_2$, and with the solenoid turned on it can be written as

$$I = \left(\frac{m}{2\pi i\hbar t}\right)^{1/2} \exp\left(\frac{i\theta_1}{\hbar}\right) \exp\left(\frac{iq}{\hbar c} \int_1 A_i dx^i\right) \left[1 + \exp\left(\frac{i(\theta_2 - \theta_1)}{\hbar}\right) \exp\left[\frac{iq}{\hbar c} \left(\int_2 A_i dx^i - \int_1 A_i dx^i\right)\right]\right]$$

$$= \left(\frac{m}{2\pi i\hbar t}\right)^{1/2} \exp\left(\frac{i\theta_1}{\hbar}\right) \exp\left(\frac{iq}{\hbar c} \int_1 A_i dx^i\right) \left[1 + \exp(i\Delta) \exp\left[\frac{iq}{\hbar c} \left(\int_2 A_i dx^i - \int_1 A_i dx^i\right)\right]\right] \quad (2)$$

But the last quantity represents a closed path from the particle source to the detector and back, and we have

$$\int_2 A_i dx^i - \int_1 A_i dx^i = -\oint A_i dx^i$$

where the minus sign reminds us that the closed path under consideration is clockwise. Using Stoke's theorem, this integral can be written as the surface integral

$$\oint A_i dx^i = \iint \vec{\nabla} \times \vec{A} dS$$

where the surface in question is the cross section of the solenoid. But this is just the magnetic flux Φ through the axis of the solenoid, given by

$$\iint \vec{\nabla} \times \vec{A} dS = \iint \vec{B} \cdot \hat{n} dS = \Phi$$

Equation (2) now becomes

$$I = \left(\frac{m}{2\pi i \hbar t}\right)^{1/2} \exp\left(\frac{i\theta_1}{\hbar}\right) \exp\left(\frac{iq}{\hbar c} \int_1 A_i dx^i\right) \left[1 + \exp(i\Delta) \exp\left(\frac{-iq\Phi}{\hbar c}\right)\right]$$

The square of this path integral is easily shown to be

$$|I|^2 = \frac{2m}{\pi \hbar t} \cos^2 \frac{1}{2} \left(\Delta - \frac{q\Phi}{\hbar c}\right)$$

Comparing this result with (1), we see that the effect of the solenoid on the charged particles is to shift the pattern in accordance with

$$\Delta \rightarrow \Delta - \frac{q\Phi}{\hbar c}$$

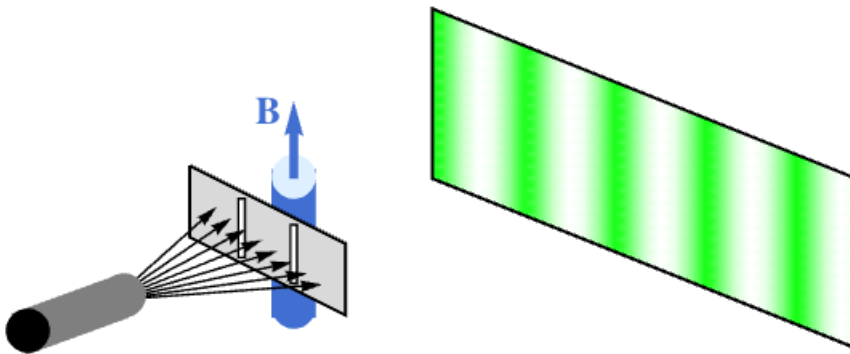
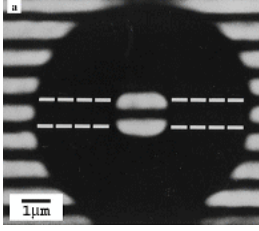


Figure 2. Experimental setup with solenoid turned on. Pattern has shifted to the right.

At the time Aharonov and Bohm derived this formula in 1959 there was no way to test it experimentally—the required solenoid had to be perfectly shielded and of such a small size (no more than several microns in diameter) that fabrication was impossible. And while there was hardly any doubt that they had made a potentially major discovery (no pun intended), many in the physics community continued to doubt that the vector potential could ever be detected. No less an authority than Niels Bohr expressed doubt on the Aharonov-Bohm prediction.

Experimental Verification of the Aharonov-Bohm Effect

In the years following publication of the Aharonov-Bohm paper, several attempts were made by various researchers to construct a suitable solenoid. For a while, it appeared that microscopic magnetized iron fibers might be workable, and results were obtained that appeared to verify the Aharonov-Bohm prediction. But indisputable results were not achieved until 1986, when Akira Tonomura and his colleagues (following up on earlier efforts they conducted in 1982) succeeded in producing a 6-micron diameter, micro fabricated toroidal solenoid utilizing the Meissner effect. The results matched the predicted phase shift perfectly, and the Aharonov-Bohm thought experiment was finally verified. For his work, Tonomura received the Nishina Memorial Prize, the Asahi Prize, the Japan Academy and Imperial Prize, and the Benjamin Franklin Medal in Physics.



Pattern shift in Tonomura experiment.

The Aharonov-Bohm Effect and Weyl's Gauge Theory

In the path integral approach that we used to derive the Aharonov-Bohm prediction, a factor appears that is associated with the action term. It is

$$\exp\left(\frac{iq}{\hbar c} \oint A_\mu dx^\mu\right)$$

(I have restored the scalar potential term.) It is interesting to note that this same term appears as a consequence of a theory Hermann Weyl developed in 1918, in which he attempted to unify the gravitational and electromagnetic fields using a generalization of Einstein's 1915 gravity theory. In Weyl's theory, the term appears as a *scaling factor* for the metric tensor $g_{\mu\nu}$ of Riemannian geometry, which determines the length or magnitude of vector quantities. Weyl's theory produced a variant of Einstein's gravitational field equations that initially appeared to be of great importance with regard to the description of electromagnetism as a purely geometric construct. Unfortunately, the theory required that the Riemannian line element, $ds^2 = g_{\mu\nu} dx^\mu dx^\nu$, be itself rescaled from point to point in space-time. As Einstein himself pointed out, this would mean that certain properties of matter would change from point to point and from one time to the next, and thus have differing *histories* as they moved through an electromagnetic field. In particular, the line spacings of atomic spectra, which of course are fixed, would change with time. Furthermore, since the quantity \hbar/mc^2 has units of time, even the mass of particles would have to change as they propagate through an electromagnetic field.

In attempting to avert disaster, Weyl sought to somehow fix the scaling of the metric tensor $g_{\mu\nu}$ so that appropriate physical quantities would remain invariant. In consideration of the transformation

$$g_{\mu\nu} \rightarrow \exp\left(\frac{iq}{\hbar c} \oint A_\mu dx^\mu\right) g_{\mu\nu}$$

Weyl noted that $g_{\mu\nu}$ would remain unchanged if

$$\frac{q}{\hbar c} \oint A_\mu dx^\mu = 2n\pi$$

This idea appeared workable, and London used it to derive the electron orbits in the Bohr atomic model. But, having conceded to Einstein's argument, Weyl eventually abandoned his theory.

But Weyl never gave up on his basic premise, which was that rescaling (which he called a *gauge transformation*) was such a beautiful idea that nature must surely take advantage of it. In a seminal paper published in 1929, Weyl applied his gauge idea to quantum mechanics, where it forever changed physics. Today, the concept of gauge invariance (the invariance of physical laws with respect to a rescaling of the wave function) is one of the foundations of all modern theories of the strong, weak and electromagnetic interactions. And it is not beyond the realm of possibility that Weyl's ideas may yet be instrumental in bringing the sole remaining force, gravity (no doubt kicking and screaming), into what we all hope will be the true Unified Theory of Everything.

References

1. Yakir Aharonov and David Bohm, *Significance of electromagnetic potentials in quantum theory*. Physical Review 115, 1959, 485–491.

2. Akira Tonomura et al., *Evidence for Aharonov-Bohm effect with magnetic field completely shielded from electron wave*. Physical Review Letters (ISSN 0031-9007), vol. 56, Feb. 24, 1986, p. 792-795.
3. Fritz London, *Quantum-mechanical interpretation of Weyl's theory*. Zeitschrift f. Physik 42, 1927.
4. Wolfram Research, *Aharonov-Bohm animation*. <http://demonstrations.wolfram.com/AharonovBohmEffect/>