

Gauge transformations and symmetry violations

H. R. Reiss

*Max Born Institute, Berlin, Germany and
American University, Washington, DC, USA**

(6 December 2017)

Abstract

The choice of electromagnetic potentials employed to represent electric and magnetic fields is shown to carry with it the possibility of violating basic symmetries in both classical and quantum physical problems. This can lead to incorrect results in dynamical calculations with an improper set of potentials. Preservation of symmetries constitutes a robust reason to regard electromagnetic potentials as more fundamental than electromagnetic fields, representing a major extension of the quantum-only example of the Aharonov-Bohm effect for accepting the primacy of potentials over fields. The demonstration of symmetry violation caused by inappropriate gauge choice affects such fundamental matters as the fact that gauge transformations are not necessarily unitary, since they do not ensure preservation of the values of physical observables.

*Electronic address: reiss@american.edu

I. INTRODUCTION

For most of the history of exploring electromagnetic phenomena, it had been believed that knowledge of the electric and magnetic fields in a physical problem is sufficient to define the problem. The scalar and vector potentials, whose spatial and temporal derivatives yield the fields, had been regarded as auxiliary quantities that are convenient but not essential. This conclusion was apparently reinforced by the fact that the set of potentials to represent the fields is not unique. Subject to modest restrictions, there exist transformations to other sets of potentials that define the same fields. Such transformations are known as gauge transformations.

This seemingly straightforward situation was upset by the Aharonov-Bohm effect [1, 2]. The simplest realization of this phenomenon is that an electron beam passing outside a solenoid containing a magnetic field will be deflected, even though there is no field outside the solenoid. There is, however, a potential outside the solenoid that suffices to explain the deflection. The effect remained controversial until it was verified experimentally [3]. This has the basic consequence that potentials are more fundamental than fields. The Aharonov-Bohm effect is explicitly quantum-mechanical, and further commentary about its significance has been in terms of quantum mechanics [4, 5].

The concept explored here is different from that of the Aharonov-Bohm effect. It is shown that the choice of gauge can be such as to violate basic symmetries, even when the usual constraints on allowable gauge transformations have been satisfied. Furthermore, these symmetry violations can occur classically as well as quantum-mechanically. The consequences of these results are profound. It is shown that there exist contrasting sets of potentials that yield exactly the same fields, but where one set is consonant with physical requirements but another is not. This proves directly that the selection of the proper set of potentials is the critical matter, since the predicted fields are the same in both cases. It is the symmetry, and hence the physical applicability, that is preserved by one set and not by the other. A corollary is that gauge transformations are not unitary transformations, despite the contrary assertion to be found in textbooks. The assumption of unitarity (often implicit) underlies some of the influential articles that have been published on the subject of gauge choice over the span of many years.

The following Section in this paper sets the electromagnetic and relativistic conventions

that are employed. Section III discusses two basic examples that exhibit pairs of potential choices that describe exactly the same fields, but where one set of potentials is physically acceptable and the other is not. One example is the simplest possible case: the classical interaction of a charged particle with a constant electric field. There are two possibilities for gauge choice, one of which contradicts Noether's Theorem [6] in the sense that, despite the plain need for conservation of energy in this problem, the Lagrangian function acquires explicit time dependence. There is no such problem with the alternative gauge. However, even the flawed gauge describes some aspects of the problem correctly. The next example is the much more substantive case of the interaction of a charged particle with a plane-wave field, such as the field of a laser. In this case, the key factor is that the symmetry principle in question – preservation of the propagation property of a plane-wave field – is not often mentioned, even in the context of very strong fields where this symmetry is crucial [7]. The demand for the preservation of the propagation property is shown to impose a strong limitation on possible gauge transformations. In the presence of a simultaneous scalar interaction, like a Coulomb binding potential, it is shown that then only the radiation gauge is a possibility [8]. This restriction exists in both classical and quantum domains. An important practical aspect of this problem is that the widely-used dipole approximation in the description of laser-caused effects suppresses this symmetry.

Section IV is concerned with the corollaries of the results about symmetry violations associated with some gauges. Foremost among them is the conclusion that gauge transformations are not unitary transformations, in contradiction to explicit contrary statements to be found in the literature. The lack of unitarity of gauge transformations has its own set of corollaries. One of them is the fallacy of the much-discussed notion of the primacy of the length-gauge interaction Hamiltonian [9–13], which has long presented the conundrum of a scalar potential being favored for vector fields such as laser fields.

II. UNITS, DEFINITIONS, CONVENTIONS

Gaussian units are employed for all electromagnetic quantities. The expressions for the electric field \mathbf{E} and magnetic field \mathbf{B} in terms of the scalar potential ϕ and the 3-vector potential \mathbf{A} are

$$\mathbf{E} = -\nabla\phi - \frac{1}{c}\partial_t\mathbf{A}, \quad \mathbf{B} = \nabla \times \mathbf{A}. \quad (1)$$

A gauge transformation generated by the scalar function Λ is

$$\tilde{\phi} = \phi + \frac{1}{c}\partial_t\Lambda, \quad \tilde{\mathbf{A}} = \mathbf{A} - \nabla\Lambda, \quad (2)$$

where Λ must satisfy the homogeneous wave equation

$$\left(\frac{1}{c^2}\partial_t^2 - \nabla^2\right)\Lambda = \partial^\mu\partial_\mu\Lambda = 0. \quad (3)$$

The relativistic convention employed herein uses the time-favoring real metric, where $\mu = 0$ labels the “time” component and $\mu = 1, 2, 3$ are the “3-space” components. These names arise from the basic 4-vector, which is the spacetime 4-vector

$$x^\mu : (ct, \mathbf{r}). \quad (4)$$

The scalar product of two 4-vectors a^μ and b^μ is

$$a \cdot b = g_{\mu\nu}a^\mu b^\nu = a^\mu b_\mu = a_\mu b^\mu = a^0 b^0 - \mathbf{a} \cdot \mathbf{b}, \quad (5)$$

where $g_{\mu\nu}$ is the metric tensor whose matrix representation is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \quad (6)$$

The metric tensor also serves to convert upper (contravariant) indices into lower (covariant) indices, as shown in Eq. (5), as well as the other way around:

$$a_\mu = g_{\mu\nu}a^\nu, \quad a^\mu = g^{\mu\nu}a_\nu, \quad (7)$$

where $g^{\mu\nu}$ has the same matrix representation as $g_{\mu\nu}$. The metric tensor shown in Eq. (6) identifies the relativistic 4-space as non-Euclidean.

The 4-vector potential A^μ incorporates the scalar and 3-vector potentials as

$$A^\mu : (\phi, \mathbf{A}). \quad (8)$$

In 4-vector notation, the two gauge transformation expressions in Eq. (2) become the single expression

$$\tilde{A}^\mu = A^\mu + \partial^\mu\Lambda, \quad (9)$$

where the 4-vector differential operator has the components

$$\partial^\mu : \left(\frac{1}{c} \partial_t, -\nabla \right). \quad (10)$$

Both the initial and gauge-transformed 4-vector potentials must satisfy the Lorenz condition

$$\partial^\mu A_\mu = 0, \quad \partial^\mu \tilde{A}_\mu = 0. \quad (11)$$

The propagation 4-vector k^μ consists of the propagation 3-vector \mathbf{k} as the space part, and the amplitude $|\mathbf{k}| = \omega/c$ as the time component:

$$k^\mu : (\omega/c, \mathbf{k}). \quad (12)$$

This means that the 4-vector k^μ lies on the light cone and, according to the rule (5), it is “self-orthogonal”:

$$k^\mu k_\mu = (\omega/c)^2 - \mathbf{k}^2 = 0, \quad (13)$$

which is an important possibility in this non-Euclidean space.

The concept of transversality refers to that property of plane-wave fields expressed in Quantum Electrodynamics (QED) as *covariant transversality*

$$k^\mu A_\mu = 0, \quad (14)$$

expressed in terms of the 4-potential A^μ . In most classical textbooks on electromagnetic phenomena, transversality is defined as *geometrical transversality*

$$\mathbf{k} \cdot \mathbf{E} = 0 \text{ and } \mathbf{k} \cdot \mathbf{B} = 0, \quad (15)$$

expressed in terms of the electric and magnetic fields. It can be shown that covariant transversality infers geometrical transversality, but the converse is not true.

III. SYMMETRY VIOLATION

The two examples presented have an important qualitative difference. The first example – a constant electric field – is so elementary that the proper choice of potentials is obvious, and there is no motivation to explore the properties of the symmetry-violating alternative potentials. The next example is quite different in that the improper choice of potentials is very attractive to a laser-physics community that is accustomed to the dipole approximation. The requisite propagation property never appears within the dipole approximation, and its violation is thereby invisible.

A. Constant electric field

The problem of a particle of mass m and charge q immersed in a constant electric field of magnitude E_0 is inherently one-dimensional. For present purposes, nothing is gained by going to three spatial dimensions. The most commonly employed potentials for representing a constant electric field are

$$\phi = -xE_0, \quad A = 0. \quad (16)$$

The Lagrangian function is the difference of the kinetic energy T and the potential energy U :

$$L = T - U \quad (17)$$

$$= \frac{1}{2}m\dot{x}^2 + qxE_0. \quad (18)$$

The Lagrangian has no explicit time dependence, so that Noether's Theorem [6] identifies this problem as one in which total energy is conserved. The Lagrangian equation of motion is

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} = m\ddot{x} - qE_0 = 0, \quad (19)$$

which is just the elementary Newtonian equation

$$m\ddot{x} = qE_0. \quad (20)$$

The simplest initial conditions for this problem – initial position and velocity set to zero – lead to the solution

$$x = \frac{qE_0}{2m}t^2. \quad (21)$$

From Eqs. (17) and (18), it follows that

$$T = \frac{1}{2m}(qE_0t)^2, \quad U = -\frac{1}{2m}(qE_0t)^2, \quad T + U = 0. \quad (22)$$

The anticipated conservation of energy holds true.

The gauge transformation generated by the function

$$\Lambda = ctxE_0 \quad (23)$$

is now introduced. This leads to the new set of potentials

$$\tilde{\phi} = 0, \quad \tilde{A} = -ctE_0, \quad (24)$$

and the new Lagrangian function [14]

$$\tilde{L} = \frac{1}{2}m\dot{x}^2 - qtE_0\dot{x}. \quad (25)$$

Although the kinetic energy is unaltered ($\tilde{T} = T$), the new potential energy is

$$\tilde{U} = qtE_0\dot{x}, \quad (26)$$

which is explicitly time-dependent. The new equation of motion is

$$\frac{d}{dt} \left(\frac{\partial \tilde{L}}{\partial \dot{x}} \right) - \frac{\partial \tilde{L}}{\partial x} = m\ddot{x} - qE_0 = 0. \quad (27)$$

This equation of motion is identical to that found in the original gauge, so that the solution is the same as Eq. (21). However, the altered gauge has introduced a fundamental change. The gauge-transformed potential energy is evaluated as

$$\tilde{U} = \frac{1}{m} (qE_0t)^2, \quad (28)$$

so that

$$\tilde{T} + \tilde{U} = \frac{3}{2m} (qE_0t)^2. \quad (29)$$

The total energy is not conserved, as was presaged by the explicit time dependence of the gauge-transformed Lagrangian (25).

How did this happen? One constraint placed on gauge transformations (see, for example, the classic text by Jackson [15]) is that the generating function must be a scalar function that satisfies the homogeneous wave equation, as in Eq. (3). This is satisfied by the function (23). The only other condition is the Lorenz condition (11), which can be verified as satisfied by the potentials before and after transformation. However, there is no condition that guarantees preservation of symmetries inherent in the physical problem. This means that it is not enough to employ appropriate fields; it is necessary to employ the appropriate potentials to ensure that all aspects of the physical problem are rendered properly.

This writer is unaware of any instance where inappropriate potentials have been accepted and employed in this exceedingly simple problem. The same cannot be said for the next example.

B. Plane-wave field

The plane-wave field is of central importance in all laser-caused processes, since laser fields are plane-wave fields. Of direct significance is the fact that a plane-wave field is the only manifestation of electromagnetic phenomena that has the ability to propagate indefinitely in vacuum without the continued presence of sources. In the typical laboratory experiments with lasers, the practical consequence of this ability to propagate without need for sources means that all fields that arrive at a target can only be a superposition of plane-wave fields. Any contamination introduced by optical elements like mirrors or gratings can persist for only a few wavelengths away from such elements. On the scale of a typical laboratory optical table, this is negligible.

All plane-wave fields propagate at the speed of light in vacuum; they are fundamentally relativistic. This means that the 1905 principle of Einstein with regard to such fields is basic: the speed of light is the same in all inertial frames of reference [16]. The mathematical statement of this principle is that any description of a plane-wave field can depend on the spacetime coordinate x^μ only as a scalar product with the propagation 4-vector k^μ . An instructive example of the way this principle is applied in practice, even when the field is modulated by an envelope function, is to be found in the article by Sarachik and Schappert [17]. The consequence of this projection of the spacetime 4-vector onto the light cone is that any change of gauge must be such as to be confined to the light cone. That is, with the definition

$$\varphi \equiv k^\mu x_\mu, \quad (30)$$

the field 4-vector must be such that

$$A_{plane-wave}^\mu = A_{plane-wave}^\mu(\varphi). \quad (31)$$

When the gauge transformation of Eq. (9) is applied, the gauge-altered 4-vector potential is confined by the condition (31) to the form [8]

$$\tilde{A}^\mu = A^\mu + k^\mu \Lambda', \quad (32)$$

where the gauge-change generating function can itself depend on x^μ only in the form of φ , and

$$\Lambda' = \frac{d}{d\varphi} \Lambda(\varphi). \quad (33)$$

As is evident from Eq. (13), transversality is maintained by the gauge transformation (32).

A further limitation arises if an electron is subjected to a scalar binding potential in addition to the vector potential associated with the laser field. A relativistic Hamiltonian function for a charged particle in a plane-wave field contains a term of the form

$$\left(i\hbar\partial^\mu - \frac{q}{c}A^\mu\right)\left(i\hbar\partial_\mu - \frac{q}{c}A_\mu\right). \quad (34)$$

This occurs in the classical case, in the Klein-Gordon equation of quantum mechanics, and in the second-order Dirac equation of quantum mechanics [18, 19]. The expansion of the expression in Eq. (34) contains the squared time part

$$\left(i\hbar\partial_t - \frac{q}{c}A^0\right)^2. \quad (35)$$

This poses a problem if A^0 contains contributions from both a scalar potential and the time part of the plane-wave 4-vector potential, since executing the square in Eq. (35) will give a term containing the product of these two scalar potentials that is not physical; it does not occur in the reduction of relativistic equations of motion to their nonrelativistic counterparts [8]. This applies specifically to applications in Atomic, Molecular, and Optical (AMO) physics. That is, it must be true that [8, 20]

$$A_{plane-wave}^0 = \phi_{plane-wave} = 0. \quad (36)$$

This means that gauge freedom vanishes. Only the gauge known as *radiation gauge* (also known as *Coulomb gauge*) is possible. This is the gauge in which scalar binding influences are described by scalar potentials ϕ and laser fields are described by 3-vector potentials \mathbf{A} .

Now consider the gauge transformation generated by the function [20]

$$\Lambda = -A^\mu(\varphi)x_\mu. \quad (37)$$

This leads to the transformed gauge

$$\tilde{A}^\mu = -k^\mu x^\nu \left(\frac{d}{d\varphi}A_\nu\right), \quad (38)$$

which was introduced in Ref. [20] in an attempt to base the Keldysh approximation [21] on plane-wave fields rather than on quasistatic electric fields. The transformed potential can also be written as [20]

$$\tilde{A}^\mu = -\frac{k^\mu}{\omega/c}\mathbf{r} \cdot \mathbf{E}, \quad (39)$$

thus suggesting a relativistic extension of the length gauge used by Keldysh and widely employed within the AMO community. The problem with the \tilde{A}^μ of Eq. (38) or (39) is that it violates the symmetry (31) required of a propagating field. Nevertheless, this \tilde{A}^μ satisfies the Lorenz condition (11) and the transversality condition (14), and the generating function of Eq. (37) satisfies the homogeneous wave equation of Eq. (3) [20]. That is, all the usual requirements for a gauge transformation are met even though the transformed 4-vector potential \tilde{A}^μ of Eq. (38) or (39) violates the symmetry required of a propagating field like a laser field.

This violation of a basic requirement for a laser field has consequences. The most obvious is that the covariant statement of the all-important [7] ponderomotive energy U_p produces a null result since

$$U_p \sim \tilde{A}^\mu \tilde{A}_\mu = 0 \quad (40)$$

as a consequence of the self-orthogonality of the propagation 4-vector k^μ . Furthermore, the resemblance of Eq. (39) to the length-gauge representation of a quasistatic electric field suggests a tunneling model for the relativistic case [22, 23], which is inappropriate for strong laser fields. Tunneling can occur only through interference between scalar potentials, and a strong laser field is inherently vector, not scalar.

IV. GAUGE TRANSFORMATIONS AND UNITARITY

Unitary transformations in quantum physics preserve the values of physical observables. It was shown in the foregoing Section that not all gauge transformations produce physically acceptable results. Therefore, gauge transformations are not unitary transformations.

The Schrödinger equation

$$i\hbar\partial_t\Psi(t) = H\Psi(t), \quad (41)$$

when viewed as a statement in a Hilbert space (that is, without selecting a representation such as the configuration representation or the momentum representation) states that the effect of rotating a state vector $\Psi(t)$ by the operator H produces the same effect as differentiating the vector with respect to time (and multiplied by $i\hbar$). Time t is an external parameter upon which the state vectors depend, which accounts for why Eq. (41) specifies t as a label independent of the Hilbert space. That is, the Schrödinger equation specifies that differentiation with respect to the external time parameter produces exactly the same effect

as the operator H that acts within the Hilbert space. A gauge transformation preserves this equivalence. This can be stated as

$$i\hbar\partial_t - \tilde{H} = U (i\hbar\partial_t - H) U^{-1}, \quad (42)$$

where U is the operator that generates the gauge transformation. Equation (42) is equivalent to the “form invariance” of the Schrödinger equation that is demonstrated in quantum mechanics texts (see, for example Ref. [24]). Form invariance means that the Schrödinger equation has the same form when the potentials corresponding to any gauge are employed. Since Eq. (42) can be written as

$$\tilde{H} = U H U^{-1} + U (i\hbar\partial_t U^{-1}), \quad (43)$$

this shows explicitly that the Hamiltonian does not transform unitarily if there is any time dependence in U . An important example is the Göppert-Mayer transformation [25] that is so widely employed in the AMO community. This transformation is given by

$$U = \exp \left(\frac{ie}{\hbar c} \mathbf{r} \cdot \mathbf{A}(t) \right), \quad (44)$$

which depends explicitly on time when $\mathbf{A}(t)$ describes a laser field.

The fact that, in general,

$$\tilde{H} \neq U H U^{-1}, \quad (45)$$

is the explanation for the curious result to be found in many papers (for example, Refs. [9–13]) that the $\mathbf{r} \cdot \mathbf{E}$ potential is a preferred potential. If any other potential is employed in solving the Schrödinger equation, then the claim is made that a transformation factor must be employed even on a non-interacting state. There is a logical contradiction inherent in the requirement that a non-interacting state must incorporate a factor that depends on an interaction, but the list of published papers that accept this premise is much longer than the salient examples cited here. The underlying problem is the assumption that a gauge transformation is a unitary transformation. That problem exists in all of the references just cited, although it is usually submerged in complicated manipulations. It is especially clear in Ref. [13], where it is specified that all operators O transform under a gauge transformation according to the unitary-transformation rule

$$\tilde{O} = U O U^{-1}. \quad (46)$$

That specification is applied to H , in violation of the condition (43), and to the interaction Hamiltonian H_I , with no explanation for how it is possible to gauge-transform from the length-gauge interaction to any other gauge in view of the absence of operators in the scalar potential $\mathbf{r} \cdot \mathbf{E}$.

-
- [1] W. Ehrenberg and R. E. Siday, *The refractive index in electron optics and the principles of dynamics*, Proc. R. Soc. B **62**, 8-21 (1949).
 - [2] Y. Aharonov and D. Bohm, *Significance of potentials in the quantum theory*, Phys. Rev. **115**, 485-491 (1959).
 - [3] A. Tonomura, N. Osakabe, T. Matsuda, T. Kawasaki, and J. Endo, *Evidence for Aharonov-Bohm Effect with Magnetic Field Completely Shielded from Electron wave*, Phys. Rev. Lett. **56**, 792-795 (1986).
 - [4] W. H. Furry and N. F. Ramsey, *Significance of potentials in quantum theory*, Phys. Rev. **118**, 623-626 (1960).
 - [5] L. Vaidman, *Role of potentials in the Aharonov-Bohm effect*, Phys. Rev. A **86**, 040101 (2012).
 - [6] E. Noether, *Invariante Variationsprobleme*, Nachr. d. Königl. Gesellsch. D. Wiss. Zu Göttingen, Math-phys. Klasse **2**, 235-257 (1918).
 - [7] H. R. Reiss, *Mass shell of strong-field quantum electrodynamics*, Phys. Rev. A **89**, 022116 (2014).
 - [8] H. R. Reiss, *Physical restrictions on the choice of electromagnetic gauge and their practical consequences*, J. Phys. B **50**, 075003 (2017).
 - [9] K.-H. Yang, *Gauge transformations and quantum mechanics: I. Gauge invariant interpretation of quantum mechanics*, Ann. Phys. NY **101**, 62 (1976).
 - [10] D. H. Kobe and A. L. Smirl, *Gauge invariant formulation of the interaction of electromagnetic radiation and matter*, Am. J. Phys. **46**, 624-633 (1978).
 - [11] R. R. Schlicher, W. Becker, J. Bergou, and M. O. Scully, *Interaction Hamiltonian in Quantum Optics* in A. O. Barut, ed., *Quantum Electrodynamics and Quantum Optics* (New York, Plenum, 1984).
 - [12] W. E. Lamb, R. R. Schlicher, and M. O. Scully, *Matter-field interaction in atomic physics and quantum optics*, Phys. Rev. A **36**, 2763-2772 (1987).

- [13] J. H. Bauer, *Simple proof of gauge invariance for the S-matrix element of strong-field photoionization*, Phys. Scr. **77**, 015303
- [14] H. R. Reiss, *On a modified electrodynamics*, J. Mod. Opt. **59**, 1371-1383 (2012); **60** 687 (2013).
- [15] J. D. Jackson, *Classical Electrodynamics* (New York, Wiley, 1975).
- [16] A. Einstein, *Zur Elektrodynamik bewegter Körper*, Ann. Phys. **17**, 891-921 (1905).
- [17] E. S. Sarachik and G. T. Schappert, *Classical theory of the scattering of intense laser radiation by free electrons*, Phys. Rev. D **1**, 2738-2753 (1970).
- [18] R. P. Feynman and M. Gell-Mann, *Theory of the Fermi Interaction*, Phys. Rev. **109**, 193-198 (1958).
- [19] S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (New York, Harper & Row, 1962).
- [20] H. R. Reiss, *Field intensity and relativistic considerations in the choice of gauge in electrodynamics*, Phys. Rev. A **19**, 1140-1150 (1979).
- [21] L. V. Keldysh, *Ionization in the field of a strong electromagnetic wave*, Sov. Phys. JETP **20**, 1307-1314 (1965) [Zh. Eksp. Teor. Fiz. **47**, 1945-1957 (1964)].
- [22] V. S. Popov, B. M. Karnakov, V. D. Mur, and S. G. Pozdnyakov, *Relativistic theory of tunnel and multiphoton ionization of atoms in a strong laser field*, JETP **102**, 760-775 (2006) [Zh. Eksp. Teor. Fiz. **129**, 871-887 (2006)].
- [23] M. Klaiber, K. Z. Hatsagortsyan, and C. H. Keitel, *Gauge-invariant strong-field approximation*, Phys. Rev. A **73**, 053411 (2006).
- [24] C. Cohen-Tannoudji, B. Diu, and F. Laloe, *Quantum Mechanics* (Paris, Hermann, 1977)
- [25] M. Göppert-Mayer, *Über Elementarakte mit zwei Quantensprüngen*, Ann. Phys., Lpz. **9**, 273-294 (1931).