A modification of electrodynamics is proposed, motivated by previously unremarked paradoxes that can occur in the standard formulation. It is shown by specific examples that gauge transformations exist that radically alter the nature of a problem, even while maintaining the values of many measurable quantities. In one example, a system with energy conservation is transformed to a system where energy is not conserved. The second example possesses a ponderomotive potential in one gauge, but this important measurable quantity does not appear in the gauge-transformed system. A resolution of the paradoxes comes from noting that the change in total action arising from the interaction term in the Lagrangian density cannot always be neglected, contrary to the usual assumption. The problem arises from the information lost by employing an adiabatic cutoff of the field. This is not necessary. Its replacement by a requirement that the total action should not change with a gauge transformation amounts to a supplementary condition for gauge invariance that can be employed to preserve the physical character of the problem. It is shown that the adiabatic cutoff procedure can also be eliminated in the construction of quantum transition amplitudes, thus retaining consistency between the way in which asymptotic conditions are applied in electrodynamics and in quantum mechanics. The 'gauge-invariant electrodynamics' of Schwinger is shown to depend on an ansatz equivalent to the condition found here for maintenance of the ponderomotive potential in a gauge transformation. Among the altered viewpoints required by the modified electrodynamics, in addition to the rejection of the adiabatic cutoff, is the recognition that the electric and magnetic fields do not completely determine a physical problem, and that the electromagnetic potentials supply additional information that is required for completeness of electrodynamics.

**Keywords:** potentials more fundamental than fields; gauge invariant ponderomotive potential; unphysical gauges; removal of adiabatic cutoff; Schwinger ansatz

### 1. Introduction

A basic intent of this paper is to show that the currently accepted rules for gauge transformations are incomplete: they leave room for gauge changes that are ‘pathological’ in the sense that clearly unphysical results can arise from them. Exploration of the implications of this problem show the need for additional constraints on allowable gauge transformations, and cast doubt on the need for asymptotic cutoff of the fields in the formalism. The most sweeping conclusion is that electromagnetic potentials contain more information than the fields that are derived from them. That is, potentials are more fundamental than fields.

Gauge invariance might appear to be an automatic property of problems in electrodynamics. If the four-vector potential of the field is $A^\mu$, then the field tensor

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$$

remains invariant under any gauge transformation of the form

$$A^\mu \rightarrow \tilde{A}^\mu = A^\mu + \partial^\mu \Lambda.$$  \hfill (2)

When $A^\mu$ satisfies the Lorenz condition

$$\partial_\mu A^\mu = 0,$$  \hfill (3)

and $A$ is a scalar function that satisfies the homogeneous wave equation

$$\partial^\mu \partial_\mu A = 0,$$  \hfill (4)

then there is a separation of the wave equation for the scalar potential $\phi$ from that for the three-vector potential $A$, and the Lorenz condition is preserved in any gauge transformation. The scalar and three-vector potentials constitute the time component and the spatial components of the four-vector potential $A^\mu$,

$$A^\mu : (\phi, A).$$  \hfill (5)
A Newtonian treatment of mechanics is formulated in terms of forces, and forces can be expressed entirely in terms of the electric and magnetic fields $E$ and $B$, related to the potentials by

$$E = -\nabla \phi - \frac{I}{c} \frac{\partial A}{\partial t},$$

$$B = \nabla \times A.$$  \hfill (6)

(Gaussian units are used here for maximum simplicity of notation.) This Newtonian result underlies the widely-accepted conclusion that only the fields are of fundamental importance in electrodynamics, and the potentials are merely convenient auxiliary quantities.

This seemingly secure point of view can be fundamentally upset, as will be shown here by two explicit, basic examples. In one, that can be viewed as a gedanken experiment (a thought experiment), a closed system is envisioned that contains a parallel-plate capacitor with the two plates holding equal and opposite electrical charges. The obvious way to describe the motion of a test charge between the plates is in terms of a simple scalar potential. As expected of an isolated physical system, it is found that energy is conserved. However, a gauge transformation is shown to exist that results in the violation of energy conservation, even though conditions (3) and (4) are obeyed. This is unacceptable, since the gedanken experiment is designed to ensure that the system is isolated, and energy conservation is a fundamental physical property of the system.

The second example is that of a charged particle immersed in a plane-wave field. As described in the radiation (or Coulomb) gauge, the particle possesses a ponderomotive potential in consequence of its interaction with the field. Were the charge to exit from the field, the ponderomotive potential would be converted to an equivalent kinetic energy. In actual laser experiments, environments exist where the ponderomotive potential is measured [1] (and has been measured [2]) in the laboratory. Nevertheless, a gauge transformation is exhibited that results in the complete loss of the ponderomotive potential.

Each of these examples – one with a pure longitudinal field and the other with a pure transverse field – is a paradox because, in both cases, a gauge transformation alters the essential character of the physical problem described. A gauge transformation should be devoid of physical consequences. These paradoxes are explicated in Sections 2 and 3. Section 4 introduces a concept whose application will be important for the latter part of the paper where the resolution of the paradoxical gauge situations will be set forth. Section 4 examines the action that follows from the Lagrangian density integrated over 4-space. Two terms occur that are related to the electromagnetic field. One is for the field alone. This is always gauge invariant since it depends only on the field tensor of Equation (1). Another term expresses the interaction between a charged particle and the field. This is conventionally discarded based on an asymptotic cutoff concept. It is this interaction term, however, that contains the key to the resolutions of the paradoxes.

Paradox resolution comes from alteration of the usual statement about how the field decouples from the system with which it interacts. Rather than have the field asymptotically decay in amplitude according to some algebraically strong model (for example, an exponential decay), it is sufficient to note that there will be some stage in the laboratory that precedes when the field and the target particles come into contact. After the interaction occurs, there will again be some interlude where the field and the target particles lose contact before the products of the interaction are detected by laboratory instruments. That is, asymptotic times can be treated as times before and after interaction has taken place, but without any specification that the field must be turned off in accordance with any algebraic rule. This is a concept that plays a role in the examination of the paradoxes, but it is also encountered in quantum dynamics in the general formulation of a transition amplitude. Section 5 contains a derivation of a general quantum transition amplitude, leading to a result that is well known. However, here also it is found to be unnecessary to specify a cutoff of the field amplitude, replacing it with the simple condition that the field and the targets have a time domain of interaction that is limited.

In Section 6, it is shown that the interaction term in the Lagrangian density can be evaluated directly in the gedanken-experiment example in consequence of its simplicity. It is found that the gauge transformation introduces an important change in the total action. The change is not zero, as is customarily assumed. If the preservation of the action is imposed as a requirement for the validity of a gauge transformation, then the paradox of the loss of energy conservation in the gauge transformation is resolved. The gauge transformation that is proposed is NOT valid because it introduces a fundamental change in the physical action. The manner in which this demonstration is carried out circumvents altogether the usual reliance on an asymptotic cutoff of the field. This presages the general conclusion of this paper that such a cutoff requirement is unnecessary and counter-productive.

In general, it is difficult to achieve an explicit evaluation of the total action. A practical alternative
procedure is to impose an additional condition for the validity of a gauge transformation. In the gedanken experiment, it can be required that no gauge transformation can be allowed that will result in changing the energy conservation condition. In the simple case of this gedanken experiment, it is shown that the requirement for sustaining energy conservation removes all gauge freedom: there is only one gauge that can be employed.

The insight gained from the gedanken-experiment example is then applied to the plane-wave problem. It is shown that the gauge transformation that results in a loss of the ponderomotive energy also results in a substantive change in the action that follows from the interaction term in the Lagrangian density. As in the first case, a requirement that a gauge transformation should not change the physical action illustrates that the proposed gauge transformation does not always satisfy this expanded set of gauge conditions. An alternative is to require that a gauge transformation should not alter the periodicity that is present in the plane-wave example. Application of that condition is shown to guarantee the preservation of the ponderomotive potential under a gauge transformation. That is, only one of the two sets of potentials explored in the plane-wave problem satisfies the physical requirement for the existence of the ponderomotive potential. This constitutes a resolution of the paradox.

A corollary to the above demonstration is the important result that it is possible to exhibit a ponderomotive potential expression that is both Lorentz-invariant and gauge-invariant. This was heretofore an elusive goal.

The key to resolving both paradoxical cases presented here is the recognition that the total action should be preserved in a gauge transformation, and that it is not sufficient to have this requirement applied at some remote location in space–time where the electromagnetic field no longer exists. Rather, it is necessary to impose the action-conservation condition in regions of space–time before the interaction between particle and field has begun, and after the interaction has been completed, but where the electromagnetic field may be present at both limits. The condition applied in this fashion contains more physical information than an application of the usual adiabatic decoupling mechanism.

An obvious question arises in consequence of the work reported here: How is it possible to reconcile the conclusions about physical ambiguity existing with the electric and magnetic fields completely specified, in view of the formally gauge invariant theory of Schwinger [3]? It is shown in Section 7 that the formal invariance asserted by Schwinger is actually qualified by an initial ansatz about how plane-wave fields must behave. The Schwinger ansatz is identical in its effects to the condition found in Section 6 that is necessary for resolution of the plane-wave paradox. Schwinger’s formal gauge invariance exists only when the ansatz is satisfied. By not employing the Schwinger ansatz from the beginning, but rather by concluding eventually that it must be present, it is possible to reach important conclusions about electrodynamics that have heretofore eluded scrutiny.

Section 8 summarizes the essential conclusions reached here, reviews the implications for how gauge invariance is satisfied, and remarks on the connection to the Aharonov–Bohm effect [4,5] and to the long-established finding [6–11] that the Schrödinger equation cannot be written directly in terms of fields without making it nonlocal.

2. A gedanken experiment

Consider an isolated system consisting of a pair of flat, parallel, conducting plates, each having a uniform distribution of electric charge of opposite sign and equal magnitude to that of the other plate. The plates are large enough that edge effects need not be considered. There will be a constant electric field $E_0$ between the plates. A hypothetical test charge of mass $m$ and charge $q$ is to be released between the plates, and its equation of motion is to be determined.

The reason for regarding the system as isolated is that conservation of energy is thereby guaranteed.

2.1. Original gauge

It is sufficient to treat this problem in one spatial dimension $x$, parallel to the direction of $E_0$. A three-dimensional treatment adds nothing essential. A Lagrangian treatment of the problem begins with statements of the kinetic energy $T$ and potential energy $V$ given by

$$T = \frac{1}{2} m \dot{x}^2; \quad V = q \phi,$$

where $\phi$ is the scalar electromagnetic potential describing the constant electric field. This static-electric-field problem is completely described by a scalar potential; no vector potential is required. The electromagnetic potentials are

$$\phi = -x E_0; \quad A = 0. \quad (8)$$

The potentials of Equation (8) constitute a Lorentz gauge because they satisfy the relation

$$\frac{\partial \phi}{\partial t} + \frac{\partial A}{\partial x} = 0, \quad (9)$$
which is a statement in nonrelativistic notation of Equation (3). The Lagrangian function

\[ L(x, \dot{x}) = T - V = \frac{1}{2} m \dot{x}^2 + qxE_0, \]

when inserted into the equation of motion, gives

\[ \frac{d}{dt} \frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} = m \ddot{x} - qE_0 = 0. \]

This is identical to the Newtonian equation of motion

\[ m \ddot{x} = F = qE_0, \]

where \( F \) is the force exerted on the charge \( q \) by the electric field \( E_0 \).

A Hamiltonian approach is equally elementary. The Hamiltonian can be derived from the Lagrangian:

\[ H(x, p) = \dot{x}p - L(x, \dot{x}) = \frac{1}{2m} p^2 - qxE_0, \]

and the equations of motion are:

\[ \dot{x} = \frac{\partial H}{\partial p} = \frac{p}{m}, \quad \dot{p} = -\frac{\partial H}{\partial x} = qE_0. \]

These equations combine to give exactly the same equation of motion as (11) and (12).

With the initial conditions \( x(0) = x_0, \quad \dot{x}(0) = v_0, \)
the particle trajectory that follows from solving (11) or (12) is

\[ x(t) = \frac{qE_0}{2m} \dot{x}^2 + v_0 t + x_0. \]

For later reference, it can be specified that the origin of the \( x \) coordinate is midway between the capacitor plates that are separated by a distance \( L \), so that

\[ -\frac{L}{2} \leq x \leq \frac{L}{2}. \]

The essential property emphasized here is that neither the Lagrangian nor the Hamiltonian has explicit dependence on the time \( t \). The canonically conjugate quantity to the time is the energy, so Noether’s theorem states that the energy is a conserved quantity. When total energy is conserved, then the potential energy is well defined, and has the value

\[ U = -qxE_0. \]

With trajectory information inserted, the kinetic and potential energies are

\[ T = \frac{1}{2m} (qE_0 t)^2 + qE_0 v_0 t + \frac{1}{2} m \dot{x}_0^2; \]
\[ U = -\frac{1}{2m} (qE_0 t)^2 - qE_0 v_0 t - qE_0 x_0. \]

The sum of the two is, indeed, constant:

\[ T + U = \frac{1}{2} m \dot{x}_0^2 - qE_0 x_0. \]

2.2. Transformed gauge

Now a gauge transformation is introduced that is generated by the function

\[ A = -E_0 c t x. \]

The gauge-transformed potentials are

\[ \phi \rightarrow \widetilde{\phi} = \phi - \frac{1}{c} \frac{\partial A}{\partial t} = 0, \]
\[ A \rightarrow \widetilde{A} = A + \frac{\partial A}{\partial x} = -E_0 c t. \]

Whereas the initial gauge had a non-vanishing scalar potential and a zero vector potential, the opposite is true in the gauge-transformed system.

The requirement for a gauge transformation that

\[ \left( \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} \right) A = 0 \]

is satisfied here. In addition, the Lorenz condition

\[ \frac{\partial \widetilde{\phi}}{\partial \tilde{t}} + \frac{\partial \widetilde{A}}{\partial \tilde{x}} = 0 \]

is satisfied in the new gauge as it was in the original gauge.

The new Hamiltonian is

\[ \tilde{H} = \frac{1}{2m} \left( \tilde{p} - \frac{q}{c} \tilde{x} \right)^2 = \frac{1}{2m} (\tilde{p} + qE_0 t)^2, \]

and the new equations of motion are

\[ \dot{x} = \frac{\partial \tilde{H}}{\partial \tilde{p}} = \frac{1}{m} (\tilde{p} + qE_0 t); \quad \dot{\tilde{p}} = -\frac{\partial \tilde{H}}{\partial \tilde{x}} = 0. \]

Combining the two equations in (24) gives exactly the same Newtonian equation (12) as before, so the trajectory has not changed. The kinetic energy remains as given in Equation (16). The trajectory and the kinetic energy are ‘observables’, so they are expected to be gauge-invariant.

When the Hamiltonian has explicit time dependence, there is no longer a well-defined potential energy, and the Hamiltonian can no longer be interpreted as a total energy. A gauge-transformed Lagrangian function is

\[ \tilde{L} = \frac{1}{2} m \dot{x}_0^2 - qE_0 \tilde{x}. \]
so that the Lagrangian in the form $\mathcal{L} = \mathcal{F} - \mathcal{K}$ yields

$$\mathcal{K} = qE_0 \dot{x}.$$  

There is no form of total energy that is independent of time. As expected from the explicit time dependence of $\mathcal{L}$ and $\mathcal{H}$, energy is not conserved.

### 2.3. The paradox

Despite the preservation of the trajectory and the kinetic energy in the gauge transformation, the gauge-transformed system has changed fundamentally from the system described by the potentials given in Equation (8). Energy conservation has disappeared.

There is now real cause for concern. The gedanken experiment is explicitly designed as an isolated system in order to ensure that no energy can be transported into or out of the system. Nevertheless, the gauge transformation generated by the $A$ of Equation (18) has altered the underlying physics of the problem.

It is easily verified that the gauge transformation function satisfies the homogeneous wave equation, as in Equation (4). Both the original and the transformed gauges satisfy the Lorenz condition (9). The gauge transformation is thus entirely ‘legitimate’ according to the accepted rules, but the physical nature of the system has changed. This is a paradox. The conceptual difficulty of this problem is emphasized by the fact that, in the original gauge, the potential energy is not only well defined, but it is measurable by a simple voltmeter. After the gauge transformation, a voltmeter reading has become meaningless.

Before this paradox is resolved, a further paradoxical result will be presented that is perhaps even more striking in the qualitative extent to which a gauge transformation alters the apparent physical meaning of a basic phenomenon in electrodynamics.

### 3. Charged particle in a plane-wave field

#### 3.1. Original gauge

A simple monochromatic plane wave in the radiation gauge (or Coulomb gauge) is described by the potentials

$$\phi_{\text{rad}} = 0, \quad A_{\text{rad}} = A_0 \epsilon \cos \varphi, \quad (25)$$

$$\varphi \equiv k \cdot x = k^\mu x_\mu = \omega t - k \cdot r, \quad (26)$$

where $\epsilon$ is a unit polarization vector and script-phi ($\varphi$) is the phase of the plane wave. The 4-vector $k^\mu$ is the propagation vector of the plane-wave field:

$$k^\mu : \left( \frac{\omega}{c}, \mathbf{k} \right),$$

and $x^\mu$ is the space–time 4-vector $x^\mu : (ct, r)$.

The radiation-gauge 4-vector potential can be written directly from the components defined in Equation (25):

$$A_{\text{rad}}^\mu : (\phi_{\text{rad}}, A_0 \epsilon \cos \varphi). \quad (27)$$

The 4-vector potential of (27) satisfies the Lorenz condition

$$\partial_\mu A_{\text{rad}}^\mu = \left( \partial_\mu \phi_{\text{rad}} \right) \frac{d}{d\varphi} A_{\text{rad}}^\mu = k_\mu \left( d_\varphi A_{\text{rad}}^\mu \right) = -k \cdot \epsilon = 0, \quad (28)$$

as a result of the transversality of the plane-wave field:

$$k \cdot \epsilon = 0. \quad (29)$$

Equations (27) and (29) lead to the 4-dimensional transversality condition

$$k_\mu A_{\text{rad}}^\mu = 0, \quad (30)$$

and to the Lorenz condition of (28).

The ponderomotive potential $U_p$ is an important property of a plane-wave field. It represents the interaction energy of a charged particle with the plane-wave field. As mentioned in Section 1, $U_p$ can be very large in strong laser fields. It is also a quantity that can be measured directly. It is given by the 4-vector potential of Equation (27) as

$$U_p = \left\langle |A^\mu A_\mu| \right\rangle. \quad (31)$$

The absolute value must be indicated because, with a time-favoring real metric as employed in Equation (27), its value would otherwise be negative. The angle brackets signify an average over a cycle of the plane-wave field.

When the dipole approximation is applicable (NOT the case here), there is a widely-used gauge transformation due to Göppert-Mayer [12] that makes it appear that a plane wave can be represented as a quasistatic electric field. This is commonly called the length gauge. That is, when

$$\varphi \approx \omega t, \quad (32)$$

the Göppert-Mayer (GM) gauge (length-gauge) potentials are

$$\phi_{\text{GM}} = -r \cdot E(t), \quad A_{\text{GM}} = 0, \quad (33)$$

where $E(t)$ is the dipole-approximation electric field vector. This gauge finds much favor in the AMO (Atomic/Molecular/Optical) community because it appears to make it possible to describe a vector field (a propagating plane wave field) by a scalar potential. The potentials of Equation (33) are simply a
generalization of the potentials of Equation (8) to three dimensions and to a time-varying electric field. That is, Equation (33) represents a quasistatic electric field.

3.2. Transformed gauge

Many years ago, it was shown [13] that there is an extension of the GM gauge to fully relativistic conditions where the dipole approximation is not applicable. This Relativistic Göppert-Mayer (RGM) gauge is given by the potentials

\[ \tilde{\phi}_{\text{RGM}} = -r \cdot \mathbf{E}(\varphi), \quad \tilde{A}_{\text{RGM}} = -\frac{k}{\omega/c}[r \cdot \mathbf{E}(\varphi)], \quad (34) \]

that can be combined into the single expression

\[ \tilde{\mathbf{A}}_{\text{RGM}}^\mu = -\frac{k^\mu}{\omega/c}[r \cdot \mathbf{E}(\varphi)], \quad (35) \]

or that can be rendered in covariant form as

\[ \tilde{\mathbf{A}}_{\text{RGM}}^\mu = -k^\mu \left[ x \cdot A_{\text{rad}}'(\varphi) \right] ; \quad A_{\text{rad}}'(\varphi) = \frac{d}{d\varphi} A_{\text{rad}}(\varphi). \quad (36) \]

That is, \( A'(\varphi) \) is the total derivative of \( A(\varphi) \) with respect to \( \varphi \), where \( \varphi \) is the phase of the wave, as given in Equation (26).

The function \( A \) that generates the transformation from the radiation gauge to the RGM gauge is

\[ A = -x \cdot A_{\text{rad}}(\varphi) = -x_{\mu} A_{\text{rad}}^\mu(\varphi). \quad (37) \]

It is straightforward to show that

\[ \partial_\mu \tilde{A}^\mu = -\partial_\mu A^\mu - k \cdot A_{\text{rad}}' - (k \cdot k)(x \cdot A_{\text{rad}}') = 0, \quad (38) \]

where the null result follows from the Lorenz condition satisfied by the radiation-gauge 4-potential \( (\partial_\mu A^\mu = 0) \), transversality of the plane wave \( (k \cdot A_{\text{rad}}' = 0) \), and the self-orthogonality property of a 4-vector on the light cone \( (k \cdot k = 0) \). It is also true that

\[ \partial_\mu \tilde{A}_{\text{RGM}}^\mu = 0. \quad (39) \]

Therefore, the transformation to the RGM gauge is a proper gauge transformation that satisfies Equations (3) and (4).

However, a basic change in physical significance arises from the gauge transformation. It follows immediately from Equation (35) or Equation (36) that

\[ \left( \tilde{U}_{\text{p}} \right)_{\text{RGM}} = \left( |\tilde{A}_{\text{RGM}}^\mu| \right)_{\text{RGM}} \left( A_{\text{RGM}}^\mu \right)_{\text{RGM}} = 0. \quad (40) \]

In view of what has been said about the very large magnitude of \( U_{\text{p}} \) that is possible with strong laser fields, and in view of the fact that \( U_{\text{p}} \) can be used to assess the onset of magnetic field effects [14,15] and of relativistic effects [15,16], it follows that the absence of the ponderomotive potential in the RGM gauge signifies a major change in physical understanding of the plane-wave problem.

This radical alteration in physical meaning can be illustrated graphically in a comparison of Figure 1 with Figure 2. These figures cover a range of field frequencies and intensities that include almost all experimental conditions that can be achieved with modern high-field lasers. Figure 1 shows the physical interpretation that is connected with the RGM gauge as expressed in Equations (34) or (35). These equations contain the electromagnetic field solely in terms of the electric field vector. One would judge the intensity of the field entirely in terms of the magnitude of the electric field. This qualitative situation is exhibited in Figure 1, which shows an electric field of one atomic unit (a.u.) as the nominal separator of electromagnetic phenomena into a strong-field domain where the electric field strength exceeds a magnitude of 1 a.u., and a weak-field domain where the field has less than a unit value of the electric field.

There is an arrow in Figure 1 labeled ‘Path to \( \omega = 0 \)’ that begins at a value typical of experiments with atoms and molecules (a wavelength of 800 nm and an intensity of \( 10^{14} \text{ W cm}^{-2} \)) and shows a progression to longer wavelengths at a constant electric field. This refers to two separate (although related) matters that have received much recent attention in the literature. One is the search [17,18] for the so-called ‘tunneling limit’ nominally expected to be approached in laser experiments as \( \omega \to 0 \). Another matter is the widespread notion that a theory can be judged for correctness based on the criterion that the low-frequency limit should approach results known to be true for static electric fields [19,20].

Figure 2 demonstrates that both of the above goals are chimerical. The increase in \( U_{\text{p}} \) that occurs as \( \omega \to 0 \) first causes magnetic fields to become important; further decline in frequency ushers in a relativistic domain. In no case is it possible with a laser field to achieve the zero-frequency limit. It is emphatically impossible with the dipole-approximation, nonrelativistic theories that have been applied to these problems.

A simple way to emphasize the impossibility of reaching zero frequency with plane-wave fields by approaching that limit along any path direction is to write \( U_{\text{p}} \) as a function of the electric field and frequency. In atomic units, the expression is

\[ U_{\text{p}} = \frac{E^2}{4\omega^2}. \quad (41) \]
Figure 1. This figure shows a range of frequencies and intensities appropriate to strong-field laser experiments. From the standpoint of the RGM gauge, the electromagnetic properties are determined by the strength of the electric field. Starting from a point at a typical laser intensity of $10^{14}$ W cm$^{-2}$ and a wavelength of 800 nm, there appears to be no impediment to extrapolating to zero frequency at a fixed electric field strength. (The color version of this figure is included in the online version of the journal.)

Figure 2. When the same frequency-intensity range as in Figure 1 is examined from the point of view of the radiation (or Coulomb) gauge, the ponderomotive potential plays a major role in establishing where magnetic field effects and relativistic effects become important. Starting at an intensity of $10^{14}$ W cm$^{-2}$ and a wavelength of 800 nm, a path to zero frequency at a fixed electric field strength or in any other direction is not possible without encountering relativistic conditions. (The color version of this figure is included in the online version of the journal.)
It is obvious that the zero-frequency limit of a plane wave does not exist.

3.3. The paradox

The RGM gauge does not exhibit any ponderomotive potential. It thereby misses altogether the low frequency limitations on plane-wave phenomena that it nominally represents. All emphasis in the RGM gauge is on the magnitude of the electric field, so that it provides no cautions about the non-existence of a tunneling limit as the frequency goes to zero, nor does it show the error that is committed when correspondence with a static-electric-field result is posited as a test of the correctness of a theory. The danger sign provided by the divergent limit of the ponderomotive potential, shown in Equation (41), has become invisible.

Nevertheless, the RGM gauge is related to the seemingly completely different radiation gauge by a valid gauge transformation that satisfies the conditions (3) and (4), as shown in Equations (38) and (39). That means that the electric and magnetic fields are identical in the two gauges, as can be verified by direct calculations of the fields from both gauges. This is paradoxical; the two gauges possess quite different qualitative properties. Furthermore, the ponderomotive potential is a measurable quantity; it is present in the radiation gauge but absent in the RGM gauge. This deepens the paradox.

The following section provides the key to the resolution of both paradoxes that have been discussed, although that resolution does not become evident until Section 6.

4. Physical action

The action $S$ can be written as

$$S = \int \mathcal{L} \, d^4x,$$

where $\mathcal{L}$ is the Lagrangian density. The Lagrangian density for an electron in a plane-wave field consists of three terms representing the free electromagnetic field, a free charged particle, and a term for the interaction between the particle and the field. The aim here is to resolve paradoxes that follow from gauge changes in the potentials that represent the field, so the free charged particle term is not needed for this investigation. The Lagrangian density to be examined consists of the two field-dependent terms

$$\mathcal{L} = -\frac{1}{16\pi} F_{\mu\nu} F^{\mu\nu} + \frac{1}{c} j^\mu A^\mu.$$

The first term in (43) is for the free electromagnetic field. Since the field tensor $F^{\mu\nu}$ is automatically gauge-invariant, there is no need to examine this term further. The second term (the interaction term) is to be explored. The difference between a gauge-transformed Lagrangian density $\mathcal{L}$ and the original $\mathcal{L}$ is

$$\mathcal{L} - \mathcal{L} = \frac{1}{c} j^\mu \left( A^\mu - A^\mu_0 \right) = \frac{1}{c} j^\mu \partial_\mu A^\mu,$$

when Equation (2) is used. The change in the action is then

$$\tilde{S} - S = \frac{1}{c} \int d^4x \partial_\mu j^\mu A^\mu.$$

This change is universally disregarded. The justification for this omission will now be examined.

The usual procedure is to employ the equivalence $\partial_\mu j^\mu A^\mu = j^\mu \partial_\mu A^\mu + A \partial_\mu j^\mu$. The expression of charge conservation is

$$\partial_\mu j^\mu = 0,$$

so that the right-hand side of Equation (45) is just the integrand of Equation (44). This places the action difference of Equation (44) in the form

$$\tilde{S} - S = \frac{1}{c} \int d^4x \partial_\mu (j^\mu A^\mu).$$

The integrand in (47) is exactly a 4-divergence expression, so that the 4-dimensional Gauss integral theorem makes it possible to convert the 4-volume integral into an integral over the surface $\Sigma$ enclosing the original 4-volume:

$$\tilde{S} - S = \frac{1}{c} \oint_{\Sigma} A \, d\sigma_\mu j^\mu,$$

where $d\sigma_\mu$ is the outwardly-directed normal to an infinitesimal element of the bounding surface $\Sigma$.

The usual assumption is that the function $A$ falls off sufficiently rapidly at large distances from an interaction region that the surface integral in Equation (48) vanishes. When this is true, then the action is gauge-invariant and all is well. That requirement is not posited for now, and the change in the action due to a gauge transformation will be left temporarily in the form of Equation (48).

5. Quantum transition amplitude without adiabatic cutoff

The intent of this section is to show that quantum transition amplitudes can be expressed in a form that does not require any assumption about how the electromagnetic field turns off at large distances.
This task is undertaken to illustrate how a similar goal can be pursued in the application of a gauge transformation of the electromagnetic fields.

The derivation of a general S matrix, or quantum transition amplitude, as given here, is designed to mimic the way in which information is gathered in the laboratory. Another goal is to minimize the imposition of unnecessary conditions or restrictions.

The model of how quantum transitions occur is based upon how experiments are conducted with pulsed, focused laser beams. The model is widely appropriate, and is not limited to the focused-laser-beam model. Nevertheless, there are exceptional circumstances (for example, where measurements are somehow accomplished \textit{in situ} while the system scrutinized remains within the influence of the transition-causing interaction) where the underlying precepts of this method do not apply. Such special cases must be examined individually.

The before-interaction time domain designated ‘\( t \to -\infty \)’ refers to times prior to the interaction of the laser beam with the target of the experiment. It is not meant to infer any ‘switching’ on or off of the field itself. For example, it could refer to a time prior to the intersection of a laser beam with a beam of atoms or ions, but with neither beam ‘turned off’. The Schrödinger equation describing the target system in full interaction with the field

\[ i\hbar \partial_t \Psi = H \Psi, \quad (49) \]

is no different in the ‘\( t \to -\infty \)’ domain from the Schrödinger equation

\[ i\hbar \partial_t \Phi = H_0 \Phi \]

that lacks altogether any interaction with the transition-causing mechanism. The difference in the two Hamiltonians

\[ H_1 = H - H_0, \quad (51) \]

called the interaction Hamiltonian, vanishes in the ‘\( t \to -\infty \)’ domain. It is then possible to place the complete set of solutions of one Schrödinger equation in one-to-one correspondence with the solutions of the other:

\[ \{ \Psi_n \} = \{ \Phi_n \}, \quad \forall n, \ t \to -\infty. \quad (52) \]

When this association is made, the \( \Phi_n \) states are called \textit{in-states}, and they are often identified with the superscript (+). This will be taken to be understood here, and the in-state superscript will be omitted.

It will be convenient to express Equation (52) as

\[ \lim_{t \to -\infty} (\Phi_n, \Psi_n) = \delta_{nn}. \quad (53) \]

After an interaction associated with \( H_1 \) has occurred, the products of that interaction are sensed by the measuring instruments designed for that purpose. The time at which this detection occurs will be designated the ‘\( t \to +\infty \)’ domain, although all that is required to warrant this designation is that the interaction products have been detected. There is no requirement that the field causing the interaction has been turned off. The laboratory instruments themselves are never exposed to the field, and any information they receive has to be interpreted in terms of non-interacting \( \Phi \) states. A fully interacting state that began as \( \Psi \) will, as \( t \to \infty \), have information about the nature of changes that have occurred due to \( H_1 \). These changes are evaluated by forming overlaps with the possible final non-interacting states \( \Phi_i \). This is the S matrix:

\[ S_{ii} = \lim_{t \to +\infty} (\Phi_i, \Psi_i). \quad (54) \]

The probability amplitude that nothing has changed can be subtracted from this in the form

\[ (S - 1)_{ii} = \lim_{t \to +\infty} (\Phi_i, \Psi_i) - \lim_{t \to -\infty} (\Phi_i, \Psi_i), \quad (55) \]

using Equation (53), and expressing the Kronecker delta as the matrix element of the unit operator:

\[ (1)_{ii} = \delta_{ii}. \]

Equation (55) is in the form of the integration of an exact differential

\[ (S - 1)_{ii} = \int_{-\infty}^{\infty} dt \frac{\partial}{\partial t} (\Phi_i, \Psi_i). \quad (56) \]

If the indicated differentiation is carried out in the integrand of Equation (56), and the Schrödinger equations (49) and (50) are employed, the end result is

\[ (S - 1)_{ii} = -\frac{i}{\hbar} \int_{-\infty}^{\infty} dt (\Phi_i, H_1 \Psi_i). \quad (57) \]

Equation (57) is the final form of the transition amplitude based on the use of in-states. An entirely analogous procedure can be employed if the interacting and non-interacting states are correlated in the ‘\( t \to +\infty \)’ domain

\[ \{ \Psi_n \} = \{ \Phi_n \}, \quad \forall n, \ t \to +\infty, \quad (58) \]

and the inquiry is made about how this could have evolved from non-interacting states \( \Phi \) in the ‘\( t \to -\infty \)’ domain. The end result of this alternative procedure is

\[ (S - 1)_{ii} = -\frac{i}{\hbar} \int_{-\infty}^{\infty} dt (\Psi_i, H_1 \Phi_i). \quad (59) \]
where the interacting states organized according to Equation (58) are called out-states, and are often identified with the superscript (−), omitted here.

Scattering theory terminology identifies the form (57) as the direct-time or post transition amplitude, and the form (59) as the time-reversed or prior transition amplitude. The circumstances of any particular problem will usually dictate which form is the more convenient one.

Even though the S-matrix terminology was originally understood to refer to scattering, and some of the conventions employed here come from the literature on scattering, the results (57) and (59) are quite general, and are not restricted to scattering problems.

An essential reminder is that the 't → +∞' and 't → −∞' domains do not require any statement about how the interaction-causing mechanism is turned on or turned off. There needs to be no statement about ‘adiabatic cutoffs’, or specification of exponential or Gaussian behavior of the field amplitude at large times. All that is required is that one limit refers to times before the field and target beams or domains have interacted, and the other limit refers to times after the interactions have occurred and the end products have been measured. These same properties will be ascribed to how the change in the total action is to be assessed upon a change of gauge.

6. Paradox resolution

Before proceeding any further it is necessary to be explicit about the sense in which the word ‘paradox’ is being used, and to be clear about the nature of the paradoxes that are being examined.

A definition of the word ‘paradox’ is: ‘A seemingly absurd or contradictory statement or proposition which when investigated may prove to be well-founded or true’ [21].

The paradox being investigated here is that, in two examples, a change of gauge has made a fundamental alteration in the physical character of a problem. This is paradoxical because the same electric and magnetic fields are present before and after a gauge transformation, and it is necessary to ascertain how it is possible to have two importantly different physical situations described by the same configuration of fields.

Resolution of the paradox will be accomplished in two steps:

1. In each of the two paradoxes it is shown that the change in the physical action expressed in Equation (48) does not vanish. This means that the usual conditions for establishing gauge invariance are not sufficient.

2. The required extra condition is then specified in the minimal form necessary to avoid the change in the physical attribute that is the underlying cause of the paradox.

6.1. Change in the action

The requirements for gauge invariance stated to this point are very simple. With a gauge transformation required to be of the form given in Equation (2), the only additional conditions imposed were that the generating function Λ satisfy the homogeneous wave Equation (4) and that the initial and transformed potentials should satisfy the Lorenz condition (3). A gauge-induced change of the action was hinted at in Equation (48), but then set aside for later consideration. As mentioned in the paragraph following Equation (48), it is customary to neglect this term on the grounds of the existence of an asymptotic cutoff of the field-dependent Λ function.

6.1.1. Constant electric field problem

The gedanken experiment specifies an explicit bounding surface. That will now be identified as the surface Σ in Equation (48).

The gedanken experiment will be defined to consider only those initial velocities that are non-negative:

\[ v_0 \geq 0. \]

The origin of time (t = 0) will be the smallest time that could occur in the problem, which would be with the test charge at the left capacitor plate (x = −L/2) and \( v_0 = 0 \). With this initial condition, the largest time \( \tau \) corresponds to the elapsed time for the test charge to traverse the full distance \( L \) at the smallest velocity, which gives:

\[ \tau = (2mL/qE_0)^{1/2}. \]

The surface Σ consists of the lines at \( t=0 \) and \( t=\tau \), and the lines at \( x=−L/2 \) and \( x=L/2 \). The scalar product in the integrand in (48) is:

\[ d\sigma_μ^μ = (dtqc)|_{t=0,\tau} - (d\epsilon p_χ)|_{x=\pm L/2}. \]

For the sides where the capacitor plates are located, the current is always non-negative at \( x=−L/2 \) because of the restriction to positive initial velocities, and aligned anti-parallel to \( dx \); and at \( x=L/2 \), the current is always positive and aligned parallel to \( dx \). However, the factor \( A = −E_0\epsilon x \) has opposite signs at the two capacitor plates because \( x=±L/2 \) has opposite signs. Therefore, the contributions from the spacelike normals to the surface Σ are of the same sign at both ends. For the timelike normals to Σ, there will be no contribution.
at $t = 0$ because of the time factor in $\Lambda$, whereas $t = \tau$ will be quite different. The overall result for $\vec{S} - \overline{S}$ in Equation (48) is non-zero.

This means that the gauge specified in Equation (8) is NOT equivalent to the transformed gauge given in Equations (19) and (20), despite meeting the usual requirements for gauge equivalence of Equations (21) and (22).

6.1.2. Plane wave problem

To make the point that there can be an important difference of the action following from a gauge transformation, it is required only to show sufficiency, not necessity. Therefore, a very symmetrical situation can be examined to simplify the demonstration.

Consider an atomic ionization experiment performed with a long-pulse circularly polarized laser beam. Complete cylindrical symmetry can be assumed. The surface $\Sigma$ will be taken to encompass the interaction domain. There is no need or justification to specify a cutoff of the field. Were the integrand in Equation (48) to contain only $d\Omega j_{\mu}$, the sum of the atomic electrons entering $\Sigma$ (see Figure 3) through the surface with timelike normal at negative times would be balanced by the outflow of photoelectrons through the surfaces with spacelike normal, plus the remaining un-ionized electrons emerging from the surface with timelike normal at positive times.

However, this factor in the integrand of Equation (48) is multiplied by $\Lambda$, given in Equation (37). With the help of Equation (27), this can be written as

$$\Lambda = r \cdot A_{\text{rad}}(\phi).$$

With (60) in place in the integrand of (48), the exact cancellation of the integral due to balances of currents passing through the timelike and spacelike surfaces of $\Sigma$ is no longer possible. The change in action due to the gauge transformation cannot always cancel.

6.2. Preservation of physical significance

6.2.1. Constant electric field problem

The gedanken experiment was constructed such that energy conservation is a requirement, and the initial gauge selected was consistent with that demand. However, a simple gauge transformation negated that essential property. If the necessity of energy conservation is applied as an additional condition that a gauge transformation be acceptable, what would be necessary to enforce that condition?

This rhetorical question is easily answered. The original gauge was associated with Lagrangian (10) and Hamiltonian (13) functions that have no explicit time dependence. That is the hallmark of energy conservation. The gauge transformation generated by Equation (18) introduced explicit time dependence and

Figure 3. The left part of the figure is a simple view of the space–time surface enclosing the volume encountered in defining an $S$ matrix. The arrows labeled $T_i$ show outwardly directed spacelike normals to the surface, and those labeled $T_T$ and $T_f$ are outwardly directed timelike normals to the surface at initial and final times. The right part of the figure is a covariant generalization of the left part. This figure is taken from [14].
thereby destroyed energy conservation. A basic constraint on possible gauge transformations that do not have this outcome is to require that \( A \) must be independent of time. Equation (2) means that there can be no change in the scalar potential. The gauge generating function \( A \) must also satisfy the homogeneous wave equation lacking time differentiation, so that Equation (4) reduces to

\[
\frac{\partial^2}{\partial x^2} A = 0.
\]

A single integration leads to

\[
\frac{\partial}{\partial x} A = C = \text{constant}.
\]

The gauge transformation rule (2) then means that the vector potential can be nothing more than the constant \( C \). The conclusion is that neither the scalar nor the vector potential can undergo a meaningful gauge transformation if energy is to be conserved.

In other words, the requirement for energy conservation in this simple constant-electric-field problem is equivalent to the statement that no gauge transformation at all is possible. The potentials of Equation (8) are unique to within an additive constant in the vector potential.

6.2.2. Plane wave problem

It will be shown here that the demand for preservation of the periodicity of the simple plane wave problem is a sufficient condition for the preservation of the ponderomotive energy.

If periodicity is not to be altered by a gauge transformation, then the generator of any such transformation can depend on space–time coordinates \( x^\mu \) only in the form of \( k \cdot x \):

\[
A = A(k \cdot x) = A(\varphi).
\]  
(61)

This means that

\[
\partial^\mu A(\varphi) = (\partial^\mu \varphi) \frac{d}{d\varphi} A = k^\mu A',
\]  
(62)

where the prime on \( A \) signifies the total derivative with respect to \( \varphi \). The wave equation for \( A \) is thus

\[
\partial^\mu \partial_\mu A(\varphi) = k^\mu k_\mu A''.
\]

The condition that \( A \) can be a function of \( \varphi \) only means that \( A' \) and \( A'' \) can only be functions of \( \varphi \) (or else simply constant). Since any 4-vector on the light cone is self-orthogonal:

\[
k^\mu k_\mu = 0,
\]  
(63)

then the requirement that \( A \) should satisfy the homogeneous wave equation is automatically satisfied.

The central issue to be addressed here is the value of the ponderomotive potential. For this, the value of \( \tilde{A}^\mu A_\mu \) is required, where

\[
\tilde{A}^\mu = A^\mu + \partial^\mu A = A^\mu + k^\mu A',
\]

when Equation (62) is used. The square of this is

\[
\tilde{A}^\mu \tilde{A}_\mu = (A^\mu + k^\mu A')(A_\mu + k_\mu A') = A^\mu A_\mu,
\]

when the transversality property

\[
k_\mu A^\mu = 0
\]

is used along with Equation (63). Therefore, in view of the defining statement (31), the ponderomotive potential is preserved in any gauge transformation that maintains the periodicity of the plane wave.

7. Schwinger gauge invariance

In a seminal 1951 paper [3], Schwinger describes what he refers to as a ‘formally gauge invariant theory’, and explores its consequences for such matters as vacuum polarization. The formal gauge invariance, however, comes with qualifications that are asserted but not examined.

Schwinger’s equations (4.1) through (4.3) imply transversality of a plane wave field (a basic geometrical property of a plane wave, also assumed here), but they furthermore require a dependence on 4-vector coordinates \( x^\mu \) only in the form of \( k^\mu, x^\mu \), a requirement that does not appear in the present treatment until Equation (61), in a discussion of the resolution of the plane-wave paradox contained in Section 6.2.2. The delayed employment in the present work of this ansatz of Schwinger places in evidence essential characteristics of electrodynamics that are otherwise obscured.

8. Comments

Perhaps the most significant result established above is that it is possible to have one set of electric and magnetic fields imply two quite different sets of physical consequences. The necessity of avoiding that dilemma led to the introduction of an additional constraint on possible gauge transformations (preservation of the total action), but also to the realization that a given configuration of fields in itself is not necessarily unique. A specification of potentials, on the other hand, does imply uniqueness. That is, potentials contain more information than fields.

This statement introduces other considerations.
About a half-century ago there was a concerted effort to answer the question of why the Schrödinger equation contains only potentials, whereas it had seemed that a direct expression in terms of fields should be possible. Each such attempt [6–11] ended with a theory containing fields, but where the field dependence was nonlocal. That is equivalent to the statement that potentials really are more fundamental than the fields. Fields are obtained from potentials by the operation of differentiation, which is a local procedure; potentials can be obtained from fields only by the operation of integration, which is a nonlocal procedure. The nonlocality of the Schrödinger equation when expressed directly in terms of the fields thus implies that potentials contain essential information beyond what is available from the fields.

A related matter is the statement in Section 1 that a Newtonian formulation is based on forces and thus requires nothing more than electric and magnetic fields. Other formulations of mechanics that are based on system functions such as the Lagrangian and Hamiltonian functions, are stated in terms of potentials. They thus contain more information than a Newtonian formulation. It is notable that it is a standard procedure in mechanics textbooks to show that the Lagrangian or Hamiltonian formulations imply the Newtonian version. The converse is not seen in textbooks, and the reason is clear. Formulations that require potentials are more complete than those expressed only in terms of fields.

In light of these results, it is not surprising that a phenomenon such as the Aharanov–Bohm effect [4,5] can occur. This effect refers to the finding that it is possible to have physical consequences arise from a situation in which a potential is present, but not a field. This theoretical prediction has been confirmed experimentally [22–24].

Another basic matter considered here is the mischief that can be caused by an asymptotic cutoff of the field. The cutoff procedure was shown to obscure the change in total action that can occur in realizable laboratory circumstances. The use of an asymptotic cutoff is generally regarded as necessary. That has been shown not to be true.

Finally, the very important ponderomotive potential associated with a plane-wave field is transparently Lorentz-invariant, but its gauge invariance had here-tofore been problematical. The ponderomotive potential of a plane-wave field has here been shown to be gauge invariant subject only to the simple condition that any gauge transformation must preserve periodicity of the potentials.

References

[1] Kibble, T.W.B. Phys. Rev. 1966, 150, 1060–1069.
[2] Bucksbaum, P.H.; Bashkansky, M.; McIlrath, T.J. Phys. Rev. Lett. 1987, 58, 349–352.
[3] Schwinger, J. Phys. Rev. 1951, 82, 664–679.
[4] Ehrenberg, W.; Siday, R.E. Proc. Roy. Soc. B 1949, 62, 8–21.
[5] Aharonov, Y.; Bohm, D. Phys. Rev. 1959, 115, 485–491.
[6] Mandelstam, S. Ann. Phys. (N. Y.) 1962, 19, 1–24.
[7] DeWitt, B.S. Phys. Rev. 1962, 125, 2189–2191.
[8] Belinfante, F.J. Phys. Rev. 1962, 128, 2832–2837.
[9] Lévy, M. Nucl. Phys. 1964, 57, 152–190.
[10] Rohrlich, F.; Strocchi, F. Phys. Rev. 1965, 139, B476–B482.
[11] Priou, M. Nucl. Phys. 1966, 81, 641–664.
[12] Göppert-Mayer, M. Ann. Phys. (Leipzig) 1931, 9, 273–294.
[13] Reiss, H.R. Phys. Rev. A 1979, 19, 1140–1150.
[14] Reiss, H.R. Prog. Quantum Electron. 1962, 16, 1–71.
[15] Reiss, H.R. Phys. Rev. Lett. 2008, 101, 043002, 159901(E).
[16] Reiss, H.R. J. Opt. Soc. Am. B 1990, 7, 574–586.
[17] Tisch, J.W.B. Nat. Phys. 2008, 4, 350–351.
[18] Colosimo, P.; Doumy, G.; Blaga, C.I.; Wheeler, J.; Hauri, C.; Catoire, F.; Tate, J.; Chirla, R.; March, A.M.; Paulus, G.G.; Muller, H.G.; Agostini, P.; DiMauro, L.F. Nat. Phys. 2008, 4, 386–389.
[19] Popov, V.S. Phys. Usp. 2004, 47, 855–885.
[20] Gribakin, G.F.; Kuchiev, M.Yu. Phys. Rev. A 1997, 55, 3760–3771.
[21] Oxford Dictionary. http://oxforddictionaries.com/definition/paradox.
[22] Olariu, S.; Popescu, I.I. Rev. Mod. Phys. 1985, 57, 339–436.
[23] Webb, R.A.; Washburn, S.; Umbach, C.P.; Laibowitz, R.B. Phys. Rev. Lett. 1985, 54, 2696–2699.
[24] Osakabe, N.; Matsuda, T.; Kawasaki, T.; Endo, J.; Tonomura, A.; Yano, S.; Yamada, H. Phys. Rev. A 1986, 34, 815–822.