The Role of the Potentials in Electromagnetism

It is well known that there are generally two ways of calculating the energy of an electromagnetic system. The first uses the field quantities and the second uses the products of the sources of charge and current with the potentials. The total energy is the same whichever method is chosen, but the energy distributions are quite different. If the potentials are used, the energy is confined to the material in which the charges and currents are located, whereas the use of the field quantities locates energy also in space devoid of matter.

From the point of view of computation the use of the potentials and sources often has considerable advantages because it confines the energy to the conductors. However, this leaves open the question of the actual energy distribution, which may affect the design and construction of a device. In this article we hope to answer the question by considering the physical and mathematical role of the potentials. Our enquiry should have the additional benefit of throwing light on the structure and meaning of the various electromagnetic quantities.

Historical Background

With the exception of the work of William Gilbert who discussed the idea of a magnetic field in his book 'De Magnete' in AD 1600, all early work on electricity and magnetism stems from Isaac Newton's mechanics. Newton's mechanics consists of two parts. First there is geometrical space which Newton regarded as an empty stationary container of infinite extent. The geometry of this space is Euclidean, so that location in space can be referred to Cartesian axes. All such axes are interchangeable because they can be related to each other by rotation and parallel displacement. Time progresses uniformly in this space. Secondly there are particles of mass moving in the empty space under the influence of forces. Amongst these forces are the gravitational attractions between the particles. Geometry is a property of space and physical action a property of matter.

The enormous success of Newton's ideas meant that it was natural to attribute electrical and magnetic effects to electric and magnetic particles. This idea was reinforced by the discovery of an inverse-square law of interaction for both types of particles just as in the law of gravitation between particles of mass. The existence of repulsive as well as attractive forces in electricity and magnetism was regarded as a minor difference. The discovery of electrical conductors and insulators lent further support for the particle view of electricity. In magnetism one simply had to assume that the particles were unable to pass from one molecule to the next and that the particles occurred in pairs.

The first serious difficulty arose when in 1820 Oersted discovered that there is a torque on a small magnet in the vicinity of an electric current. Ampère followed this with a thorough investigation of the reactions between currents, which led him to the conclusion that magnetic dipoles are equivalent to small current loops and that current circuits are equivalent to magnetic double layers or 'magnetic shells'. In the absence of such shells in nature Ampère concluded that magnetic effects are due to electricity in motion. He then faced the problem of how to incorporate electric currents into the Newtonian scheme of particle interaction. His solution was the 'current element' and he devised a law of force between such elements regarded as particles.

The law had the inverse-square relation and also the Newtonian requirement of equality between action and reaction, because the force acted along the line joining the two elements. However, there was no means of testing it, because there were no such things as isolated current elements. All that could be observed was the force on a short piece of circuit due to a complete circuit. The assumption of action along the line joining the elements was without experimental basis and Ampère's cumbersome formula is never used. In the absence of current particles the Newtonian scheme could not be applied to magnetic effects.

An even sharper attack on the Newtonian view came from the combined work of Faraday and Maxwell. Faraday had a strong aversion to the notion of isolated particles. He thought that an electric charge by itself could have no existence, because it would have no physical effect. He also distrusted the idea of action in a straight line, because he observed that the magnetic force curved round the current. This led him to the concept of linkage between electric current and magnetic field. Linkage demands a closed curve, which is a topological feature and not a local geometrical one. A closed curve cannot be replaced by an open one. This meant that the phenomenon could not be due to the interaction between discrete particles, because Newton's empty space has no topological features. Another equally important discovery made by Faraday was that time and 'rate of change' affect the processes, so that electromagnetic systems are dynamic rather than static.

These ideas were taken up by Maxwell, who clothed them in mathematical form. Although Maxwell used the terminology of mechanics in terms of force and momentum, he regarded their transference to electromagnetism as an analogy rather than a physical equivalence. Also he used the system dynamics of Lagrange, which are based on energy, rather than the particle dynamics of Newton. In Lagranian mechanics attention is transferred from the parts of a system to the energy distribution connecting the parts. There are two kinds of energy, kinetic and potential. Maxwell's crowning achievement lay in the discovery that in electromagnetism these two types of energy interact to produce a wave of energy travelling through space with a constant velocity. The experimental fact that energy is transmitted in this manner shows that there are no such things as empty space or universal time. Space and time are linked entities.

Einstein developed these ideas in his special theory of relativity. He took the velocity of the electromagnetic waves as a universal constant and concluded that every observer has his own set of coordinates of space and time. Since the velocity of light determines the relation between these coordinates, geometry and physics cannot be separated in the Newtonian manner. There is no geometry without physics, nor physics without geometry. Moreover the geometry of electromagnetic effects is non-Euclidean because the velocity of light acts as a barrier between the time-like regions of past and future and the space-like regions, which are 'elsewhere' and beyond the reach of the observer located at the origin of coordinates defined as 'here and now'. Electromagnetic space-time is essentially a system exhibiting curvature.

However, if time is taken to vary harmonically with a particular frequency, time and space can be separated from each other by using the time-axis as having imaginary numbers. Such space-time is associated with the name of Minkowski and is described as pseudo-Euclidean. It does not distinguish between past and future. Although it is then possible to retain the idea of parallelism in space, the Minkowski space-time retains the topological features which makes it different from Newtonian space and time.

One further general observation needs to be made about the use of vectors in electromagnetism. Vectors act at a point and are particularly appropriate to the investigation of systems of point-particles. When they are applied to electromagnetism they are less successful and exhibit various contradictory features such as the distinction between polar and axial vectors. The cause of these difficulties is that point vectors seek to separate physical action from geometry. The difficulties disappear when vectors are replaced by differential forms which combine space and time with observable physical features. However, in order to avoid unfamiliar notation we shall use vector algebra in this article.

The Electrostatic Potential

In a discussion of gravitational attraction Lagrange discovered a function which shortened the labour of calculating the force on a particle due to other particles. This function is now known as the Newtonian Potential. It consists of the sum of the masses of the particles each of which is divided by the distance from the point at which the function is to be calculated. The name 'potential' was given by Green, who showed that the function could be used in electrostatics, if the masses are replaced by charges, and in magnetostatics, if the masses are replaced by magnetic poles or dipoles. Time does not enter into these calculations and space is the flat infinite container of Newtonian mechanics. The energy is energy of position and is therefore potential energy. Hence the name 'potential' is very appropriate.

We can write the electrostatic potential as

$$\varphi_e = \frac{1}{4\pi\varepsilon}\sum_{i=1}^{n}\frac{Q_i}{r} = \frac{1}{4\pi\varepsilon}\int_{v'}\frac{\rho_e}{r}\,dv' \quad , \qquad (1)$$

where Q_i is the charge of the i^{th} particle, ε is the permittivity, r is the distance between the particle and the point at which the potential is calculated, ρ_e is the volume density of charge and v' is the volume occupied by the charges. The electric field is given by

$$\vec{E} = -grad\ \varphi_e \qquad (2)$$

The electric flux-density is given by the constitutive equation

$$\vec{D} = \varepsilon\,\vec{E} \qquad (3)$$

The electric field is conservative and the potential has a unique value apart from an arbitrary constant.

The Magnetostatic Potentials

If magnetostatics is regarded as the effect of magnetic particles we can write

$$\varphi_m = \frac{1}{4\pi\varepsilon}\int_{v'}\frac{\rho_m}{r}\,dv' \qquad (4)$$

However, if the sources are steady electric currents, we must start with the field equations

$$curl\ \vec{H} = \vec{J} \qquad (5)$$

$$div\ \vec{B} = 0 \qquad (6)$$

$$\vec{B} = \mu\,\vec{H} \qquad (7)$$

There is no unique scalar potential because the energy is kinetic and not potential.

However, eqn. (6) allows us to put

$$\vec{B} = curl\ \vec{A} \qquad (8)$$

Then from eqns. (5) and (7) we obtain

$$curl\ (curl\ \vec{A}) = curl\ \vec{B} = \mu\ curl\ \vec{H} = \mu\,\vec{J} \qquad (9)$$

The left-hand side of this equation can be transformed by the vector identity

$$curl\ (curl\ \vec{A}) = grad\ div\ \vec{A} - \nabla^2\,\vec{A} \qquad (10)$$

Hence

$$grad\ div\ \vec{A} - \nabla^2\,\vec{A} = \mu\,\vec{J} \qquad (11)$$

Since a vector is defined by its divergence and curl and since $\vec{A}$ has so far been defined only by its curl, we can put

$$div\ \vec{A} = 0 \qquad (12)$$

Hence

$$\nabla^2\,\vec{A} = -\mu\,\vec{J} \qquad (13)$$

This has the integral solution

$$\vec{A} = \frac{\mu}{4\pi}\int_{v'}\frac{\vec{J}}{r}\,dv' \qquad (14)$$

Comparison of eqns (4) and (14) shows that $\vec{A}$ is the vector analogue of the scalar potential φ_m. This has led to it being called the 'vector potential', a somewhat unfortunate term since potential energy cannot be a vector.

For a current element we can write

$$\vec{A} = \frac{\mu}{4\pi}\frac{I\,\vec{dl'}}{r} \qquad (15)$$

Eqn (15) suggests that we have solved Ampère's problem of finding an expression for the effect of an isolated current element. We have certainly obtained a very useful tool for calculating the field of currents. A particularly useful feature of eqn (14) is the fact that $\vec{A}$ and $\vec{J}$ are parallel to each other.

However, this parallelism leads us back to the Euclidean property of Newtonian space. In fact we have separated space from physical action and have lost sight of Faraday's discovery of linkage. We have also lost sight of the finite velocity of light and its relativistic effects but this is less important because that velocity is so enormous. In many practical cases we can regard it as infinite and this is what we have done in using the magnetostatic equation (5), where we have omitted Maxwell's displacement current.

Clearly the idea of linkage must not be disregarded and it is included in eqn (5). We have got rid of it in the expressions for $\vec{A}$ by making the assumption of eqn (12). This is an arbitrary choice and therefore $\vec{A}$ unlike φ has no unique value. It is not an observable quantity, but is subject to a 'gauge transformation' given by

$$\vec{A}' = \vec{A} + grad\ \Lambda \qquad (16)$$

where Λ is a scalar function. This leaves $\vec{B}$ unchanged, since

$$curl\ \vec{A}' = curl\ \vec{A} + curl\ grad\ \Lambda = curl\ \vec{A} \qquad (17)$$

Eqn (12) assumes that Λ is a harmonic function because

$$div\ grad\ \Lambda = \nabla^2 \Lambda = 0 \qquad (18)$$

Apart from this Λ is arbitrary. We are now able to answer the question of the correct energy distribution. Since $\vec{A}$ is not unique and is not observable, the energy density $\vec{A}.\vec{J}$ is not unique either. On the other hand $\vec{B}$ and $\vec{H}$ are observable and the field energy is therefore uniquely correct.

We need now to investigate the physical reason for the gauge invariance of $\vec{A}$. This must be something to do with the idea of linkage and the fact that the absence of current particles makes it impossible to use the idea of Newtonian space. However, before we discuss these matters we need to include the effects of time on the phenomena.

Time - Varying Potentials

We start with the observable field relationships.

$$curl\ \vec{E} = -\frac{\partial \vec{B}}{\partial t} \qquad (19)$$

$$curl\ \vec{H} = \vec{J} + \frac{\partial \vec{D}}{\partial t} \qquad (20)$$

$$div\ \vec{B} = 0 \qquad (21)$$

$$div\ \vec{D} = \rho \qquad (22)$$

$$\vec{B} = \mu \vec{H} \qquad (23)$$

$$\vec{D} = \varepsilon \vec{E} \qquad (24)$$

From eqn (21) we put

$$\vec{B} = curl\ \vec{A} \qquad (25)$$

Then from eqn (19)

$$\vec{E} = -\frac{\partial \vec{A}}{\partial t} - grad\ \varphi \qquad (26)$$

and from eqns (20), (24) and (26)

$$curl\ (curl\ \vec{A}) = \mu \vec{J} + \mu\varepsilon\left(-\frac{\partial^2 \vec{A}}{\partial t^2} - grad\frac{\partial\varphi}{\partial t}\right) \qquad (27)$$

Hence

$$grad\ div\ \vec{A} - \nabla^2 \vec{A} = \mu \vec{J} + \mu\varepsilon\left(-\frac{\partial^2 \vec{A}}{\partial t^2} - grad\frac{\partial\varphi}{\partial t}\right) \qquad (28)$$

Rearranging the terms in this equation and using the experimental result

$$\mu\varepsilon = \frac{1}{c^2} \qquad (29)$$

where c is the velocity of light, we have

$$\nabla^2 \vec{A} - \frac{1}{c^2}\frac{\partial^2 \vec{A}}{\partial t^2} = -\mu \vec{J} + grad\left(div\ \vec{A} + \frac{1}{c^2}\frac{\partial\varphi}{\partial t}\right) \qquad (30)$$

Since $div\ \vec{A}$ has not been defined, we can use the Lorentz gauge condition

$$div\ \vec{A} + \frac{1}{c^2}\frac{\partial\varphi}{\partial t} = 0 \qquad (31)$$

Hence

$$\nabla^2 \vec{A} - \frac{1}{c^2}\frac{\partial^2 \vec{A}}{\partial t^2} = -\mu \vec{J} \qquad (32)$$

This is a wave-equation and has the integral solution

$$\vec{A} = \frac{\mu}{4\pi}\int_{v'}\frac{[\vec{J}]}{r}dv' \qquad (33)$$

The square bracket indicates that J must be taken at an earlier time given by $t' = t - r/c$, where t is the time at which $\vec{A}$ is to be calculated.

From eqn (22)

$$\nabla^2\varphi - \frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} = -\frac{\rho}{\varepsilon} \qquad (34)$$

and

$$\varphi = \frac{1}{4\pi\varepsilon} \int_{v'} \frac{[\rho]}{r} dv' \qquad (35)$$

We notice that $\vec{A}$ and φ are related to each other, so that the potential now has four components: three spatial components of $\vec{A}$ and a time component φ. The combined 4-potential is subject to a gauge invariance of

$$\vec{A}', \varphi' = \vec{A}_1 \varphi + grad\ \Lambda \qquad (36)$$

where the gradient has a time component as well as three space components. The gauge transformation leaves the electric and magnetic fields unchanged. The choice of the Lorentz gauge implies that

$$\nabla^2 \Lambda - \frac{1}{c^2}\frac{\partial^2 \Lambda}{\partial t^2} = 0 \qquad (37)$$

This arbitrary choice makes it possible to separate current and charge in eqns (32) and (34) and that makes possible the parallelism between $\vec{A}$ and J in eqn (33). We have also had to use Cartesian coordinates, because only in these coordinates it is possible to separate $\nabla^2 \vec{A}$ into three components.

Once again we have to conclude that the potentials are not unique, because they depend on a choice of gauge and a choice of coordinates. These choices are arbitrary and the potentials are therefore not physically observable. The energy distribution associated with the potential has no physical significance apart from the total energy of a system.

What is the Physical Significance of the Gauge Invariance?

We have seen the potential was originally defined for a system of interacting particles. Such a system can be described in terms of potential energy, which is a scalar quantity. Faraday's discovery of the mutual linkage between electric currents and magnetic fields drew attention to a topological property which conflicts with the idea of the empty flat space used in Newtonian mechanics. It is not possible to describe electromagnetic interaction in Newtonian terms, because space and physical action cannot be separated. There has to be a complete re-appraisal of the relationship between geometry and physics.

The vector potential was devised as a tool for describing the interaction of sources having direction as well as position. Such interaction requires a comparison of vectors and the comparison involves transportation of a vector in space. In a flat space this presents no difficulty, because the direction of the vector is not affected by its displacement, but in a curved space the direction depends on the curvature and on the path taken. This becomes clear when we consider, for example, the displacement of a vector in a curved two-dimensional space such as the surface of a sphere.

The direction of a vector is independent of the choice of coordinates. Its components depend on that choice, but the vector itself remains invariant. This explains its usefulness in describing physical phenomena. However, the derivative of a vector depends on the coordinates, because in a curved space it depends on the path for the displacement. It becomes helpful to define a 'covariant derivative' which is independent of the choice of coordinates. The technicalities are not important for our present purpose. The important results are that the covariant derivative is equal to the ordinary derivative plus a term consisting of the product of the vector with a geometrical object called the 'affine connection'. This connection depends on the curvature and the path. It is not a unique local quantity, but as its name implies it makes a connection. We also find that the covariant derivative of a vector around a small closed loop is equal to the local curvature.

These ideas can be applied to the electromagnetic interaction. Consider again the interaction between current-elements. For this we need the vector potential and we have found that it provides the connection. The vector potential is subject to a choice of gauge and is not a unique local quantity. Thus there is a close analogy between its behaviour and that of the affine connection in a curved space. That is not surprising, because Faraday's linkage exhibits curvature and the absence of independent current particles shows that there is curvature in the electromagnetic space containing current sources.

There is an even deeper physical significance in the gauge invariance of the electromagnetic field. So far we have treated electric charge as consisting of particles and current as the motion of such particles. However, we learn from quantum theory that these particles are inherently wave-functions which can be described in terms of a magnitude and phase-angle. The magnitude relates to probability and is observable. The phase-angle can be observed only in terms of phase-differences and is not a uniquely local quantity. The phase transformation of the wave-function is closely related to the gauge transformation of the potential, so that the phase-angle can be expressed in terms of the gauge of the potential. A charge of gauge, therefore, does not effect the magnitude of the wave-function, but the derivative of the wave-function is affected by a change of gauge.

Once again it is useful to define a covariant derivative and as expected, the potential behaves as an affine connection. Moreover the covariant derivative around a small closed loop is equal to the local electromagnetic field. We conclude that the field describes the curvature which characterizes the electromagnetic interaction. It is an observable local quantity. This remarkable result shows the underlying unity between macroscopic electromagnetism and the microscopic quantum behaviour.

End-Note

A full account of the subjects discussed in this article is given in the book: 'Geometry of Electromagnetic Systems' by D. Baldomir and P. Hammond (Oxford University Press 1996).

The book explains the advantages of using differential forms in the description of electromagnetic phenomena, because they combine geometry and physics.

This book also gives a much more complete account of the relationship between the gauge of the potential and the phase of the wave-function, which is mentioned briefly in the article.

Percy Hammond