Ann. Rev. Nucl. Part. Sci. 1984. 34 : 461–530

MAGNETIC MONOPOLES[1]

John Preskill[2]

California Institute of Technology, Pasadena, California 91125

CONTENTS

[1] Work supported in part by the US Department of Energy under contract DEAC-03-81-ER40050.

[2] Alfred P. Sloan Fellow.

0163–8998/84/1201–0461$02.00

1. INTRODUCTION

How is it possible to justify a lengthy review of the physics of the magnetic monopole when nobody has ever seen one? In spite of the unfortunate lack of favorable experimental evidence, there are sound theoretical reasons for believing that the magnetic monopole must exist. The case for its existence is surely as strong as the case for any other undiscovered particle. Moreover, as of this writing (early 1984), it is not certain that nobody has ever seen one. What seems certain is that nobody has ever seen two.

The idea that magnetic monopoles, stable particles carrying magnetic charges, ought to exist has proved to be remarkably durable. A persuasive argument was first put forward by Dirac in 1931 (1). He noted that, if monopoles exist, then electric charge must be quantized; that is, all electric charges must be integer multiples of a fundamental unit. Electric charge quantization is actually observed in Nature, and no other explanation for this deep phenomenon was known.

Many years later, another very good argument emerged. Polyakov (2) and 't Hooft (3) discovered that the existence of monopoles follows from quite general ideas about the unification of the fundamental interactions. A deeply held belief of many particle theorists is that the observed strong and electroweak gauge interactions, which have three apparently independent gauge coupling constants, actually become unified at extremely short distances into a single gauge interaction with just one gauge coupling constant (4, 5). Polyakov and 't Hooft showed that any such "grand unified" theory of particle physics necessarily contains magnetic monopoles. The implications of this discovery are rich and surprising and are still being explored.

While Dirac had demonstrated the consistency of magnetic monopoles with quantum electrodynamics, 't Hooft and Polyakov demonstrated the necessity of monopoles in grand unified gauge theories. Furthermore, the properties of the monopole are calculable, unambiguous predictions in a given unified model.

All grand unified theories possess a large group of exact gauge symmetries that mix the strong and electroweak interactions, but these symmetries become spontaneously broken at an exceedingly short distance scale M_X^{-1} (or, equivalently, an exceedingly large mass scale M_X). The properties of the magnetic monopole, such as its size and mass, are determined by the distance scale of the spontaneous symmetry breakdown (the "unification scale"). The prediction that magnetic monopoles must exist does not depend on the *mechanism* of the symmetry breakdown; for example, it does not matter whether the Goldstone bosons associated with the symmetry breakdown are elementary or composite. Nor does it matter

whether gravitation becomes unified with the other particle interactions at the unification scale.

The magnetic charge g of the monopole is typically the "Dirac charge" $g_D = 1/2e$. (Magnetic charge will be defined so that the total magnetic flux emanating from a charge g is $4\pi g$. Electric charge is defined so that the electric flux emanating from a charge e is e.) This magnetic charge is distributed over a core with a radius of order M_X^{-1}, the unification distance scale, and the mass of the monopole is comparable to the magnetostatic potential energy of the core.

The unification mass scale M_X varies from one grand unified model to another. But M_X can be calculated if we make a very strong assumption—the "desert hypothesis"—that is, if we assume that no unexpected new interactions or particles appear between present-day energies (of order 100 GeV) and the unification scale M_X. [This assumption is also the basis of the highly successful calculation (6) of the electroweak mixing angle $\sin^2 \theta_w$.] From the desert hypothesis follows the prediction $M_X \approx 10^{14}$ GeV (6); the properties of the monopole may then be summarized by

$$\begin{aligned} &\text{Charge:} && g = g_D = 1/2e, \\ &\text{Core size:} && R \approx M_X^{-1} \approx 10^{-28} \text{ cm}, \\ &\text{Mass:} && m \approx (4\pi/e^2) M_X \approx 10^{16} \text{ GeV}. \end{aligned} \qquad 1.$$

Here $e^2/4\pi$ is the running coupling constant renormalized at the mass scale M_X, making it somewhat larger than $\alpha \approx 1/137$.

Of course, the desert hypothesis could easily be wrong, even if the general idea of grand unification is correct. So the size and mass of the monopole could be much different from the estimates in Equation 1. It is nonetheless interesting to note that one can reasonably expect the monopole to be an *extremely* heavy stable elementary particle; 10^{16} GeV $\approx 10^{-8}$ g $\approx 10^6$ J is comparable to the mass of a bacterium, or the kinetic energy of a charging rhinoceros. It is hardly surprising that magnetic monopoles have not been produced by existing particle accelerators.

We also see from Equation 1 that the size R of the core of the monopole is expected to be larger than its Compton wavelength by a factor of order $4\pi/e^2$. In this sense, the monopole is a nearly classical object; quantum mechanics plays an insignificant role in determining the structure of its core, if e^2 is small. In fact, magnetic monopoles appear in spontaneously broken unified gauge theories even in the classical limit, as stable time-independent solutions to the classical field equations.

The stability of the classical monopole solution is ensured by a topological principle to be explained in detail below. Loosely speaking, the monopole is a "defect" in the scalar field that acts as an order parameter for

the spontaneous breakdown of the grand unified gauge symmetry. Trapped inside its core is a region in which the scalar field respects symmetries different from those respected by the vacuum state. This scalar field configuration is energetically unfavorable, so the core cannot expand. But the magnetostatic energy of the core prevents it from shrinking. So the core is stable.

While most of the mass of the monopole is concentrated in its tiny core of radius M_X^{-1}, the monopole has interesting structure on many different size scales (Figure 1). At distances less than $M_Z^{-1} \approx 10^{-16}$ cm from the center of the monopole, virtual W and Z bosons have important effects on its interactions with other particles. The monopole is also a hadron; it has a color magnetic field that extends out to distances of order 10^{-13} cm, and then becomes screened by nonperturbative strong-interaction effects. And, because of its large magnetic charge, the monopole is strongly coupled to a surrounding cloud of virtual electron-positron pairs, which extends out to distances of order $m_e^{-1} \approx 10^{-11}$ cm. In a grand unified theory in which new physics appears at energies below the unification scale M_X (so that the desert hypothesis does not apply), the structure of the monopole might be even more complicated.

The existence of magnetic monopoles is a very general consequence of the unification of the fundamental interactions. But it is one thing to say that monopoles must exist, and quite another to say that we have a reasonable chance of observing one. If monopoles are as heavy as we expect, there is no hope of producing monopoles in any foreseen accelerator. Our best hope is to observe a monopole in cosmic rays. But since no process occurring in the present universe is sufficiently energetic to produce monopoles, any

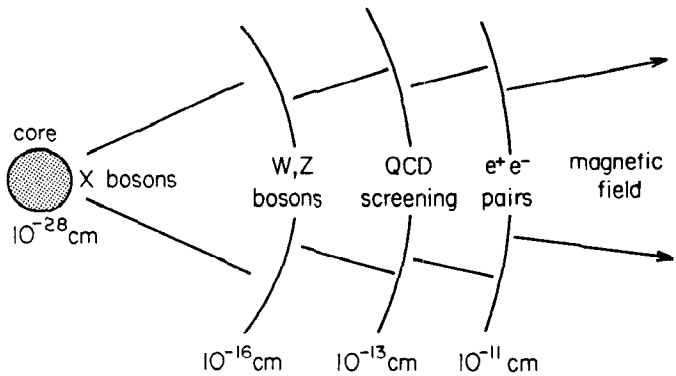

Figure 1 Structure of a grand unified monopole.

monopoles around today must have been produced in the very early universe, when higher energies were available. Thus, the abundance of magnetic monopoles is a cosmological issue (7–9).

In fact, estimates based on the standard cosmological scenario indicate that the monopole abundance should exceed by many orders of magnitude the current observational limits. Thus, our failure to observe a monopole is itself a significant piece of information, casting doubt on either the standard view of the evolution of the universe, or on cherished beliefs about particle physics at extremely short distances. This dilemma has led to revolutionary new developments in theoretical cosmology (10–12).

Significant as it may be not to see a monopole, it would be even more interesting to see one. But astrophysical arguments indicate that the flux of monopoles in cosmic rays is probably quite small (13, 14). Furthermore, if monopoles are very heavy, those bombarding the earth are likely to be moving relatively slowly, with velocities of order $10^{-3}\,c$. Detection of these slow, rare monopoles is a challenging problem for experimenters.

If a magnetic monopole is ever discovered, it will be a momentous occasion, with many fascinating implications. For one thing, that there are any monopoles at all would be evidence that the universe was once extremely hot. And severe constraints would be placed on our attempts at cosmological model building, for the observed monopole abundance would have to be explained by any realistic cosmological scenario.

Detection of a monopole would also confirm a very fundamental prediction of grand unification. The mass of the monopole, if it could be measured, would reveal the basic symmetry-breaking scale at which electrodynamics becomes truly united with the other particle interactions. More could be learned about very short-distance physics by studying the interactions of monopoles with fermions. Remarkably, a charged fermion (e.g. a quark or lepton) incident on a monopole at low energy can penetrate to the core of the monopole, and probe its structure (15, 16). Thus monopoles could provide us with a unique window on new physics at incredibly short distances.

But even if nobody ever sees a magnetic monopole, there is surely much to be gained by studying the theory of monopoles. Already, marvelous insights into gauge theory and quantum field theory have been derived from this study. There is little reason to doubt that further surprising discoveries await the dedicated student of the magnetic monopole.

The main purpose of this article is to present the basic results of monopole theory. For the most part, the presentation is intended to be accessible to a reader with a minimal background in theoretical particle physics. In Section 2, the connection between magnetic monopoles and the quantization of electric charge is explained, and in Section 3 the classical

monopole solution of 't Hooft and Polyakov is introduced. The theory of magnetic monopoles carrying nonabelian magnetic charge is developed in Section 4, and the general connection between the topology of a classical monopole solution and its magnetic charge is established there. Various examples illustrating and elucidating the formalism of Section 4 are discussed in Section 5. Section 6 is concerned with the properties of dyons, which carry both magnetic and electric charge. Aspects of the interactions of fermions and monopoles are considered in Section 7. In Section 8, the cosmological production of monopoles and astrophysical bounds on the monopole abundance are described. Some remarks about the detection of monopoles are contained in Section 9.

The reader who finds gaps in the present treatment may wish to consult some of the other excellent reviews of these topics. For a general review of grand unified theories, see (17, 18). For more about some of the topics in Section 2–4, see (19–21); for Section 6, see (21); for Section 8, see (22–26); and for Section 9, see (27, 28).

2. THE DIRAC MONOPOLE

2.1 *Monopoles and Charge Quantization*

Measured electric charges are always found to be integer multiples of the electron charge. This quantization of electric charge is a deep property of Nature crying out for an explanation. More than fifty years ago, Dirac (1) discovered that the existence of magnetic monopoles could "explain" electric charge quantization.

Dirac envisaged a magnetic monopole as a semi-infinitely long, infinitesimally thin solenoid (Figure 2). The end of such a solenoid looks like a

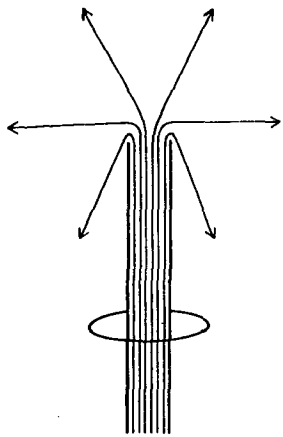

Figure 2 The end of a semi-infinite solenoid.

magnetic charge, but it makes sense to identify this object as a magnetic monopole only if no conceivable experiment can detect the infinitesimally thin solenoid.

We can imagine trying to detect the solenoid by doing an electron interference experiment (29); such an experiment gives a null result only if the phase picked up by the electron wave function, when the electron is transported along a closed path enclosing the solenoid, is trivial. Suppose a point monopole with magnetic charge g sits at the origin, so that the magnetic field is

$$\mathbf{B} = g\frac{\hat{r}}{r^2}, \qquad 2.$$

and that the solenoid lies on the negative z-axis. Then, in spherical coordinates, and in an appropriate gauge, the only nonvanishing component of the vector potential is

$$A_\phi = g(1-\cos\vartheta), \qquad 3.$$

where A_ϕ is defined by $\mathbf{A}\cdot\mathbf{dr} \equiv A_\phi\, d\phi$. The electron interference experiment fails to detect the solenoid if

$$\exp[-ie\oint\mathbf{A}\cdot\mathbf{dr}] = \exp[-i4\pi eg] = 1, \qquad 4.$$

where $-e$ is the electron charge. Hence, we require the magnetic charge g to satisfy Dirac's quantization condition (1)

$$eg = \frac{n}{2}. \qquad 5.$$

The minimum allowed magnetic charge $g_D = 1/2e$ is called the Dirac magnetic charge.

Dirac's reasoning shows that it is consistent in quantum mechanics to describe a magnetic monopole with the vector potential Equation 3, even though it has a "string" singularity for $\vartheta = -\pi$. The string is undetectable. In fact, we formulate in Section 4.1 a different mathematical description of the monopole, in which the string is avoided altogether.

The quantization condition Equation 5 requires all magnetic charges to be integer multiples of the Dirac charge $g_D = 1/2e$. We can also turn this argument around, as follows: Suppose there exists a magnetic monopole with magnetic charge g_D. Then it is consistent for a particle with electric charge Qe (and vanishing magnetic charge) to exist only if $\exp[i4\pi Qeg_D] = 1$, or

$$Q = (1/2eg_D)n = n, \qquad 6.$$

where n is an integer. Therefore, the existence of a magnetic monopole implies quantization of electric charge.

2.2 *Generalizations of the Quantization Condition*

To derive the Dirac quantization condition (Equation 5), we used the electron charge $-e$. But we believe that quarks exist, and the electric charge of a down quark, for example, is $-e/3$. Will not the same argument as before, applied to a down quark instead of an electron, lead to the conclusion that the minimal allowed magnetic charge is $3g_D$ instead of g_D?

No, not if quarks are confined (30). For if quarks are permanently confined in hadrons, it makes sense to speak of performing a quark interference experiment only over distances less than 10^{-13} cm, the size of a hadron. It is true that, when the down quark is transported around Dirac's string, its wave function acquires the nontrivial phase

$$\exp\left[-i(e/3)\oint \mathbf{A}_{em}\cdot d\mathbf{r}\right] = \exp(-i2\pi/3) \neq 1 \qquad 7.$$

due to the coupling of the down quark to the electromagnetic vector potential, if the monopole carries the Dirac magnetic charge g_D. But we must recall that the down quark carries another degree of freedom, color. The string is not detectable if the monopole also has a *color-magnetic field*, such that the phase acquired by the down quark wave function due to the color vector potential compensates for the phase due to the electromagnetic vector potential, or

$$\exp\left[ie_c\oint \mathbf{A}_{color}\cdot d\mathbf{r}\right] = \exp(i2\pi/3), \qquad 8.$$

where e_c is the color gauge coupling.

The correct conclusion, then, if quarks are confined, is not that the minimal magnetic charge is g_D, but rather that the monopole carrying magnetic charge g_D must also carry a color-magnetic charge. The color-magnetic field of the monopole becomes screened by nonperturbative strong-interaction effects at distances greater than 10^{-13} cm (21, 31). We also conclude that there cannot exist both *isolated* fractional electric charges and monopoles with the Dirac magnetic charge, unless there is some other (as yet unknown) long-range field that couples to both the monopoles and the fractional electric charges (32).

To state the Dirac quantization condition in its most general form, we note that the vector potential of a magnetic monopole carrying more than one type of magnetic charge can in general be written (33)

$$\sum_a e^a T^a A^a_\phi = \tfrac{1}{2}M(1-\cos\vartheta), \qquad 9.$$

where M is a constant matrix. The sum over a runs over all the generators of the gauge group, and the gauge couplings e^a have been absorbed into M. By

an argument similar to that invoked above (see also Section 4.2), we can derive the generalized Dirac quantization condition

$$\exp(i2\pi M) = 1. \qquad 10.$$

That is, M must have integer eigenvalues.

For example, in the SU(5) grand unified model, the electric charge generator may be written as a 5×5 matrix

$$Q_{\text{em}} = \text{diag}(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}, 0, -1), \qquad 11.$$

where the diag$(-, -, -, -, -)$ notation denotes a diagonal matrix with the indicated eigenvalues. The eigenvalues of Q_{em} are the electric charges, in units of e, of the elements of the 5 representation of SU(5)—antidown quarks, in three colors, the neutrino and the electron. The color SU(3) generators are traceless 3×3 matrices acting on the quarks only; one of these is

$$Q_{\text{color}} = \text{diag}(-\tfrac{1}{3}, -\tfrac{1}{3}, \tfrac{2}{3}, 0, 0). \qquad 12.$$

A matrix M that satisfies Equation 10 is

$$M = Q_{\text{em}} + Q_{\text{color}} = \text{diag}(0, 0, 1, 0, -1), \qquad 13.$$

and the magnetic charge of the monopole described by Equation 13 is the coefficient of eQ_{em} in $\frac{1}{2}M$, or $1/2e = g_{\text{D}}$, the Dirac charge.

In the SU(5) model, a restatement of the criteria in Equations 10 and 13 for the existence of a magnetic monopole with the Dirac charge is

$$\exp[i2\pi Q_{\text{em}}] = \text{diag}[\exp(i2\pi/3), \exp(i2\pi/3), \exp(i2\pi/3), 1, 1] \equiv Z \qquad 14.$$

where Z is a nontrivial element of Z_3, the center of color SU(3). Equation 14 is just a fancy way of saying that objects that carry trivial color SU(3) triality have integer electric charge (in units of e), even though objects with nontrivial triality have fractional charge. That the U(1) group generated by Q_{em} contains the center of color SU(3) also has a topological significance, which is elucidated in Sections 4 and 5.

Another interesting generalization of Equation 5 applies to *dyons*, objects that carry both electric and magnetic charge. Consider the two dyons with electric and magnetic charges $(Q_1 e, M_1 g_{\text{D}})$ and $(Q_2 e, M_2 g_{\text{D}})$. Each dyon is unable to detect the string of the other if and only if (34)

$$Q_1 M_2 - Q_2 M_1 = n, \qquad 15.$$

where n is an integer. The minus sign in Equation 15 arises because transporting the first dyon counterclockwise around the string of the second is equivalent to transporting the second dyon clockwise around the string of the first.

The condition represented by Equation 15 requires all magnetic charges to be integer multiples of g_D, if there exists a particle with $Q = 1$ and $M = 0$. But magnetically charged objects are allowed to carry anomalous electric charges. Equation 15 is satisfied if Q and M for all dyons are related by

$$Q = n - \frac{\vartheta}{2\pi} M, \qquad 16.$$

where n is an integer, and ϑ is an arbitrary parameter defined modulo 2π (see Figure 3). For a dyon carrying more than one type of magnetic charge, there is a distinct ϑ for each type.

The significance of the parameter ϑ is discussed further in Sections 6 and 7. Here we merely note that the dyon charge spectrum (Equation 16) violates CP unless ϑ is 0 or π, because Q is CP odd, and M is CP even.

So far, we have taken the magnetic monopole to be pointlike; the magnetic field (Equation 2) is singular at the origin. But it is obvious that our derivation of the quantization condition will also apply to a nonsingular field that approaches the form of Equation 2 at large distances. Such a nonsingular monopole is constructed in Section 3.

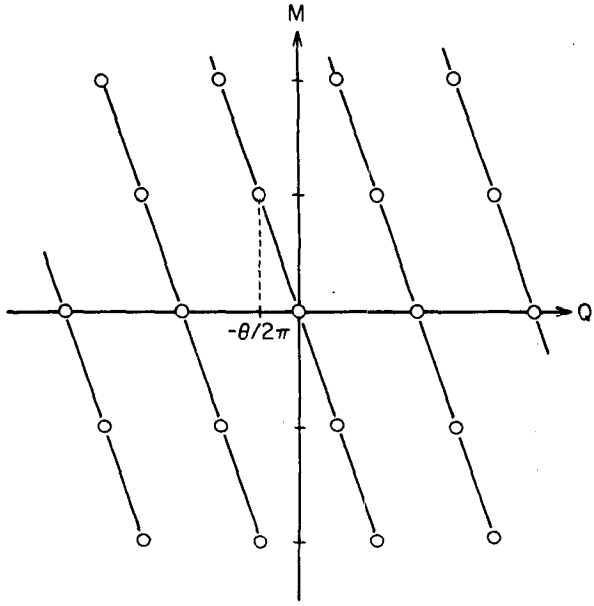

Figure 3 Electric charges (Q) and magnetic charges (M) allowed by the Dirac quantization condition.

3. MONOPOLES AND UNIFICATION

3.1 *Unification, Charge Quantization, and Monopoles*

Dirac showed that quantum mechanics does not preclude the existence of magnetic monopoles. Moreover, the existence of monopoles implies quantization of electric charge, a phenomenon observed in Nature. The monopole thus seems to be such an appealing theoretical construct that, to quote Dirac, "one would be surprised if Nature had made no use of it" (1).

Nowadays, we have another way of understanding why electric charge is quantized. Charge is quantized if the electromagnetic $U(1)_{em}$ gauge group is compact. But $U(1)_{em}$ is automatically compact in a *unified* gauge theory in which $U(1)_{em}$ is embedded in a nonabelian semisimple group. [Note that the standard Weinberg-Salam-Glashow (35) model is not "unified" according to this criterion.]

In other words, in a unified gauge theory, the electric charge operator obeys nontrivial commutation relations with other operators in the theory. Just as the angular momentum algebra requires the eigenvalues of J_z to be integer multiples of $\frac{1}{2}\hbar$, the commutation relations satisfied by the electric charge operator require its eigenvalues to be integer multiples of a fundamental unit. This conclusion holds even if the symmetries generated by the charges that fail to commute with electric charge are spontaneously broken.

These two apparently independent explanations of charge quantization are not really independent at all. Dirac found the existence of monopoles to imply charge quantization, but the converse, in a sense, is also true. Any unified gauge theory in which $U(1)_{em}$ is embedded in a spontaneously broken semisimple gauge group, and electric charge is thus automatically quantized, necessarily contains magnetic monopoles. The discovery of this remarkable result, by 't Hooft (3) and Polyakov (2), ushered in the modern era of monopole theory.

In contrast to Dirac's demonstration of the *consistency* of magnetic monopoles with quantum electrodynamics, t' Hooft and Polyakov demonstrated the *necessity* of monopoles in unified gauge theories. Furthermore, the properties of the monopole are calculable in a given unified model. In particular, its mass can be related to the masses of certain heavy vector bosons, while in Dirac's formulation of electrodynamics, the monopole mass must be regarded as an arbitrary free parameter.

There has been much speculation in recent years about "grand unified" models of elementary particle interactions, in which the standard low-energy gauge group $SU(3)_{color} \times [SU(2) \times U(1)]_{electroweak}$ is embedded in a simple gauge group that is spontaneously broken at a large mass scale. The simplest model of this type is the SU(5) model (4). But the prediction that

magnetic monopoles must exist applies to any grand unified model, and also to the even more ambitious models purporting to unify gravitation with the other particle interactions.

3.2 *Monopoles as Solitons*

In this section we show how magnetic monopoles arise in unified gauge theories as solutions to the classical field equations. A semiclassical expansion about the classical monopole solution can be carried out to arbitrary order in $\hbar$, but for now we confine our attention to the classical approximation. Some properties of the semiclassical expansion in higher order are discussed in Section 6.

Here we consider the simplest unified gauge theory containing a monopole solution (36). The generalization to more complicated models is described in Sections 4 and 5.

The model has the gauge group SU(2) and a Higgs field Φ in the triplet representation of the group; its Lagrangian is

$$\mathscr{L} = -\tfrac{1}{4}F^a_{\mu\nu}F^{\mu\nu a} + \tfrac{1}{2}D_\mu\Phi^a D^\mu\Phi^a - U(\Phi), \tag{17.}$$

where

$$U(\Phi) = \tfrac{1}{8}\lambda(\Phi^a\Phi^a - v^2)^2, \tag{18.}$$

$$D_\mu\Phi^a = \partial_\mu\Phi^a - e\varepsilon^{abc}W^b_\mu\Phi^c, \tag{19.}$$

$$F^a_{\mu\nu} = \partial_\mu W^a_\nu - \partial_\nu W^a_\mu - e\varepsilon^{abc}W^b_\mu W^c_\nu, \tag{20.}$$

and $a = 1, 2, 3$. The energy density can be written as

$$\mathscr{E} = \tfrac{1}{2}[E^a_i E^a_i + B^a_i B^a_i + D_i\Phi^a D_i\Phi^a] + U(\Phi), \tag{21.}$$

where

$$E^a_i = F^a_{0i}, \qquad B^a_i = \tfrac{1}{2}\varepsilon_{ijk}F^a_{jk}. \tag{22.}$$

Since $\mathscr{E} \geq 0$, the classical "vacuum" of this theory is a field configuration such that $\mathscr{E} = 0$. In the "unitary" gauge, the scalar field Φ may be written

$$\Phi = (0, 0, v + \varphi), \tag{23.}$$

and the vacuum configuration is

$$\varphi = 0, \qquad W^a_\mu = 0. \tag{24.}$$

To determine the perturbative spectrum in this gauge, we substitute Equation 23 into the Lagrangian. Since

$$\tfrac{1}{2}D_\mu\Phi^a D^\mu\Phi^a = \tfrac{1}{2}(\partial^\mu\varphi)^2 + \tfrac{1}{2}e^2(v^2 + \ldots)[(W^1_\mu)^2 + (W^2_\mu)^2], \tag{25.}$$

and

$$U(\Phi) = \tfrac{1}{2}\lambda v^2\varphi^2 + \ldots, \qquad 26.$$

we find that the theory has undergone the Higgs mechanism; there is a massless "photon" W_μ^3 that couples to the unbroken $U(1)_{em}$ current, as well as charged vector bosons $W_\mu^\pm$ with mass

$$M_W = ev \qquad 27.$$

and a neutral scalar with mass

$$M_H = \sqrt{\lambda}v. \qquad 28.$$

To investigate the spectrum of this theory beyond perturbation theory, let us determine whether there is a stable time-independent solution to the classical field equations other than the vacuum solution. Equivalently, we seek a field configuration at a fixed time that is a local minimum of the energy functional $\int d^3r\mathscr{E}$. Such a "soliton" configuration behaves like a particle in the classical theory, and can be expected to survive in the spectrum of the quantum theory.

Our search for a nontrivial local minimum of the energy functional will surely succeed if there are field configurations that cannot be continuously deformed to the vacuum configuration while the total energy remains finite. For if we start with such a configuration and deform it until a local minimum is obtained, the final configuration is guaranteed to be different from the vacuum solution.

Furthermore, it is easy to demonstrate the existence of such a configuration. For a field configuration of finite energy, the scalar field Φ is required to approach a minimum of the potential $U(\Phi)$ at large distances, but Φ is free to select different minima of U in different spatial directions. The asymptotic behavior of Φ defines a mapping $\Phi^a(\hat{r}) = \lim_{\alpha\to\infty} \Phi^a(\alpha\mathbf{r})$ such that

$$\Phi^a(\hat{r})\Phi^a(\hat{r}) = v^2; \qquad 29.$$

that is, a mapping from the sphere at spatial infinity to the sphere of minima of $U(\Phi)$.

Consider a "hedgehog" configuration such that the mapping $\Phi^a(\hat{r})$ is the identity mapping

$$\Phi^a(\hat{r}) = v\hat{r}^a. \qquad 30.$$

It is evident that there is no way of continuously deforming the mapping of Equation 30 to the trivial mapping $\Phi^a(\hat{r})$ = constant, while preserving the finite-energy condition, Equation 29. The number of times the mapping

$\Phi^a(\hat{r})$ "wraps around" the manifold of minima of $U(\Phi)$ is an integer. Since an integer cannot change continuously, this "winding number" is preserved by continuous deformations; it is said to be a "topological invariant." But the hedgehog configuration has winding number 1, and the vacuum configuration has winding number 0. Therefore, the vacuum configuration cannot be obtained by any continuous deformation of the hedgehog configuration that is consistent with Equation 29.

It only remains to verify that there really is a hedgehog configuration that asymptotically approaches Equation 30 and has finite energy. The contribution $\frac{1}{2}\int d^3x(D_i\Phi^a)^2$ to the energy is finite only if $D_i\Phi^a$ approaches zero at large r sufficiently rapidly. We therefore require

$$D_i\Phi^a \sim 0 \qquad 31.$$

for large r, or

$$W_i^a \sim \frac{\varepsilon_{iak}\hat{r}_k}{er}, \quad \text{and} \quad B_i^a \sim \frac{\hat{r}_i\hat{r}^a}{er^2}. \qquad 32.$$

The long-range gauge field (Equation 32) is a $U(1)_{em}$ gauge field that carries magnetic charge $g = 1/e$ [where $U(1)_{em}$ is the subgroup of SU(2) left unbroken by the scalar field, Equation 30]. The charge $1/e$ is really the Dirac magnetic charge in this model, since it is possible to introduce matter fields in the doublet representation of SU(2) that carry electric charge $e/2$.

We thus conclude that there must be a stable, finite-energy, time-independent solution to the classical equations of motion such that the asymptotic scalar field configuration $\Phi^a(\hat{r})$ has winding number 1. Finiteness of the energy requires the long-range gauge field of this soliton to be the field of a Dirac magnetic monopole.

In general, we may consider field configurations such that the winding number is an arbitrary integer. Since time evolution is continuous, and the winding number is discrete, it must be a constant of the motion in the classical field theory. We have seen that this "topological conservation law" is equivalent to conservation of magnetic charge. The conservation law survives in quantum theory because the probability of quantum mechanical tunneling between configurations with different winding numbers vanishes in the infinite volume limit.

The above discussion of the topological charge and its connection with magnetic charge is reformulated in much more general language in Section 4.

3.3 *The Monopole Solution*

We have demonstrated the existence of a time-independent monopole solution to the classical field equations. Let us now consider how the solution can be explicitly constructed.

The task of finding an explicit monopole solution is greatly simplified if we make the plausible assumption that the solution is spherically symmetric. In a gauge theory, it is not sensible to demand more than spherical symmetry up to a gauge transformation; we say that the scalar field configuration $\Phi^a(\mathbf{r})$, for example, is spherically symmetric if the effect of a spatial rotation of $\Phi^a(\mathbf{r})$ can be compensated by a gauge transformation. The asymptotic behavior of Φ^a given by Equation 30 and of W_i^a given by Equation 32 is invariant under a simultaneous rotation and global SU(2) gauge transformation. Let us assume that this invariance, and also invariance under the "parity" transformation

$$\mathbf{r} \to -\mathbf{r}, \qquad \Phi^a \to -\Phi^a, \qquad W_i^a \leftrightarrow -W_i^a, \tag{33.}$$

hold for all $\mathbf{r}$. We thus obtain the ansatz (2, 3)

$$\Phi^a(\mathbf{r}) = v\hat{r}_a H(M_W r)$$

$$W_i^a(\mathbf{r}) = \frac{\varepsilon_{iak}\hat{r}_k}{er}[1 - K(M_W r)]. \tag{34.}$$

Finite-energy solutions will obey the boundary conditions

$$\begin{aligned} H &= 0, \qquad K = 1 \qquad (r = 0); \\ H &= 1, \qquad K = 0 \qquad (r = \infty). \end{aligned} \tag{35.}$$

H and K satisfying the classical field equations can now be obtained by numerical methods (20, 37). [In fact, an analytical solution is possible in the limit $\lambda = 0$ (38, 39).] Here we merely note a few general features of the solution.

The gauge field W_i^a rapidly approaches its asymptotic value outside a core with radius of order R_c; the heavy gauge fields are excited only inside the core. The size R_c is chosen to minimize the sum of the energy stored in the magnetic field outside the core and the energy due to the scalar field gradient inside the core. In order of magnitude these are

$$\begin{aligned} E_{\text{mag}} &\sim 4\pi g^2 R_c^{-1} \sim (4\pi/e^2)R_c^{-1}, \\ E_{\text{core}} &\sim 4\pi R_c v^2 \sim (4\pi/e^2)M_W^2 R_c, \end{aligned} \tag{36.}$$

so the core size is determined to be

$$R_c \sim M_W^{-1}. \tag{37.}$$

The energy of the solution, the monopole mass m in the classical approximation, does not depend sensitively on the scalar self-coupling λ;

one finds

$$m = \frac{4\pi}{e^2} M_W f(\lambda/e^2), \tag{38.}$$

where f is a monotonically increasing function such that (37)

$$f(0) = 1$$
$$f(\infty) = 1.787. \tag{39.}$$

The mass m becomes independent of λ for large λ because the scalar field approaches its asymptotic form outside an inner core with radius $R_H \sim M_H^{-1}$, and the scalar field energy stored in the inner core is of order

$$E_{\text{scalar}} \sim 4\pi\lambda v^4 R_H^3 \sim m(M_W/M_H), \tag{40.}$$

which becomes negligible for large λ.

Comparing Equations 37 and 38, we see that the size of the monopole core is larger by the factor $\alpha^{-1} = (4\pi/e^2)$ than the monopole Compton wavelength. As a result, the quantum corrections to the structure of the monopole are under control, if α is small. Even though the coupling $g = 1/e$ is large, the effects of virtual monopole pairs are small, because the monopole is a complicated coherent excitation that cannot be easily produced as a quantum fluctuation. (See Section 6.)

This situation should be contrasted with the quantum mechanics of a point monopole. Virtual monopole pairs have a drastic effect on the structure of the point monopole, for which g is a genuine strong coupling. In fact, the vacuum-polarization cloud of a point monopole must extend out to distances of order $(\alpha m)^{-1}$, because the magnetic self-energy of a monopole of that size is of order m. Thus, both the nonsingular monopole and the point monopole have a complicated structure in a region with radius of order $(\alpha m)^{-1}$. But for the nonsingular monopole, we have an explicit classical description of this structure, and quantum corrections are small and calculable if α is small. The point monopole, on the other hand, is a genuine strong-coupling problem. We cannot calculate anything.

We have shown how the magnetic monopole arises as a solution to the classical field equations in a simple SU(2) gauge theory. The discussion is generalized in Section 4 and various more complicated examples are cited in Section 5.

It turns out (40) that in many, but not all (41), more complicated examples it is possible to construct a monopole solution that satisfies a suitable generalization of the spherically symmetric ansatz, Equation 34. But nothing further is said here about the construction of explicit solutions.

4. MONOPOLES AND TOPOLOGY

4.1 *Monopoles without Strings*

The Dirac string is a considerable embarrassment in monopole theory. It is disconcerting to find that the vector potential that describes a Dirac monopole has a string singularity along which the magnetic field is formally infinite, even though we can argue that the string is undetectable. One is therefore encouraged to discover that it is possible to eliminate the string (42).

Let us consider the vector potential on a sphere centered at the monopole. (The monopole may be either pointlike or nonsingular; in the latter case we choose the radius of the sphere to be much larger than the core radius.) The trick by which we avoid the string is to divide the sphere into upper and lower hemispheres, and define a vector potential on each. For example, we may choose the nonvanishing component of **A** on each hemisphere to be

$$A_\varphi^{\mathrm{U}} = g(1-\cos\vartheta), \qquad \text{upper}\left(0 \le \vartheta \le \frac{\pi}{2}\right),$$

$$A_\varphi^{\mathrm{L}} = -g(1+\cos\vartheta), \quad \text{lower}\left(\frac{\pi}{2} \le \vartheta \le \pi\right), \tag{41.}$$

where A_φ is defined by $\mathbf{A}\cdot\mathrm{d}\mathbf{r} = A_\varphi \mathrm{d}\varphi$. Both A^{U} and A^{L} are nonsingular on their respective hemispheres, and both have the curl

$$\mathbf{B} = g\frac{\hat{r}}{r^2}. \tag{42.}$$

On the region where the hemispheres intersect, the equator ($\vartheta = \pi/2$), we must require that A^{U} and A^{L} describe the same physics; therefore, they differ by a gauge transformation. And, indeed

$$A_\varphi^{\mathrm{U}}\left(\vartheta = \frac{\pi}{2}\right) - A_\varphi^{\mathrm{L}}\left(\vartheta = \frac{\pi}{2}\right) = 2g = \frac{1}{ie}(\partial_\varphi\Omega)\Omega^{-1}, \tag{43.}$$

where

$$\Omega(\varphi) = \exp[i2eg\varphi]. \tag{44.}$$

If Ω is not single-valued, then the change in the phase of the wave function of an electron, as the electron is transported around the equator, is ill defined. So we must demand

$$eg = \frac{n}{2}, \tag{45.}$$

where n is an integer. We have thus found an alternative derivation of the Dirac quantization condition, in which the string singularity makes no appearance.

It is easy to see that this quantization condition applies to any vector potential on the sphere, not just one with the special form of Equation 41. In general, if the nonsingular vector potentials A^U and A^L defined on the upper and lower hemispheres differ by a gauge transformation $\Omega(\varphi)$ at the equator, then we may interpret $\Omega(\varphi)$ as an object that detects the total magnetic flux Φ through the sphere. If $\Omega(\varphi = 0) = 1$, then $\Omega(\varphi = 2\pi)$ satisfies

$$\Omega(\varphi = 2\pi) = \exp[ie\oint d\mathbf{x}\cdot(\mathbf{A}^U - \mathbf{A}^L)] = \exp[ie(\Phi^U + \Phi^L)]$$
$$= \exp[ie(4\pi g)], \qquad 46.$$

where the line integral is taken along the equator, and g is the magnetic charge enclosed by the sphere. Single-valuedness of $\Omega(\varphi)$ again implies Equation 45.

The integer n, the magnetic charge of the monopole in Dirac units, is a winding number; it is the number of times $\Omega(\varphi)$ covers the $U(1)_{em}$ gauge group as φ varies from 0 to 2π. So we have discovered a topological basis for the Dirac quantization condition. Magnetic charge is quantized because the winding number must be an integer.

If we now allow the radius r of the sphere to vary, $\Omega(\varphi, r)$ and the winding number n are continuous functions of r as long as A^U and A^L are nonsingular. Since n is required to be an integer, it must be a constant, independent of r. If n is nonzero, we are forced to conclude that the magnetic charge g is contained in an arbitrarily small sphere; the monopole is a point singularity.

It is possible to avoid the singularity only if Ω is allowed to wander through a larger gauge group containing $U(1)_{em}$. This is precisely the option exercised by the nonsingular monopole described in Section 3, which has nonabelian gauge fields excited in its core.

4.2 *Topological Classification of Monopoles*

It is easy to generalize the above discussion to apply to magnetic monopoles with nonabelian, long-range gauge fields, and thus obtain a topological definition of magnetic charge appropriate for the nonabelian case (21, 43).

Let us consider gauge fields, defined on a sphere, in the Lie algebra of an arbitrary Lie group H. As before, we describe the gauge field configuration by specifying nonsingular gauge potentials A^U and A^L on the upper and lower hemispheres, and a single-valued gauge transformation $\Omega(\varphi)$, which relates A^U and A^L on the equator. The gauge transformation $\Omega(\varphi)$ is a

"loop" in the gauge group H, a mapping from the circle into H. We define the magnetic charge enclosed by the sphere to be the winding number of $\Omega(\varphi)$. This is the natural nonabelian generalization of the abelian magnetic charge.

For example, suppose that the gauge group is $H = SO(3)$. It is well known that SO(3) is topologically equivalent to a three-dimensional sphere with antipodal points identified. Therefore, there are closed paths in SO(3), those beginning at one point of the three-sphere and ending at an antipodal point, which cannot be continuously deformed to a point. Such a path is said to have winding number 1. But a path that begins and ends at the same point of the three-sphere can be continuously deformed to a point; it has winding number 0. Thus, the winding number of a loop in SO(3) can have only two possible values, 0 and 1, and the magnetic charge in an SO(3) gauge theory can have only the values 0 and 1. In particular, a magnetic monopole is indistinguishable from an antimonopole.

In general, the closed paths in a Lie group H beginning and ending at the identity element of H fall into topological equivalence classes, called "homotopy" classes (44). Two paths are in the same class if they can be continuously deformed into one another. The classes are endowed with a natural group structure, since the composition of two paths may be defined to be a path that traces the two paths in succession. This group is called $\pi_1(H)$, the "first homotopy group" of H. According to the above remarks, $\pi_1[SO(3)]$ is Z_2, the additive group of the integers defined modulo 2.

The example $H = SO(3)$ exhibits all the essential features of the general case. Every Lie group H has a covering group $\bar{H}$, which is simply connected; that is, such that $\pi_1(\bar{H})$ is trivial. For $H = SO(3)$, the covering group is $\bar{H} = SU(2)$, which is topologically equivalent to the three-sphere itself, without antipodal points identified. The Lie group H is always isomorphic to the quotient group $\bar{H}/K$, where K is a subgroup of the center of $\bar{H}$. The center is a discrete subgroup of $\bar{H}$ that commutes with all elements of $\bar{H}$. For $\bar{H} = SU(2)$, the center is Z_2, consisting of the elements 1 and -1, and SO(3) is isomorphic to $SU(2)/Z_2$. In general, we may think of the group H as the group $\bar{H}$, but with elements differing by multiplication by an element of K identified as the same element.

All paths in H that begin and end at the identity element of H correspond to paths in $\bar{H}$ that begin at the identity and end at an element of K. And the topological class of a path in H can be labeled by the end point of the corresponding path in $\bar{H}$, just as the class of a path in SO(3) is determined by whether it ends at its starting point on the three-sphere or at the antipodal point. So we finally have

$$\pi_1(H) = \pi_1(\bar{H}/K) = K. \qquad 47.$$

For $H = U(1)$, which is covered by $\bar{H} = \mathbb{R}$, the additive group of the real numbers, K is Z, the integers. For $H = SO(3)$, K is Z_2, and for any simple Lie group H, K is Z_N, for some integer N.

Our topological definition of the nonabelian magnetic charge is sensible. As long as the gauge fields are nonsingular and Ω is an element of H, the winding number must be a constant, independent of the radius of the sphere. So the magnetic charge is not carried by the long-range field of a monopole; it resides on a point singularity (Dirac monopole) or a core in which gauge fields other than the H gauge fields are excited (nonsingular monopole). And this magnetic charge is obviously conserved. It is a discrete quantity. But time evolution is continuous, and a discrete quantity can be continuous only by being constant.

While other gauge-invariant definitions of magnetic charge are possible (33), only the topological definition, which requires a magnetic monopole to have a point singularity or a core, can guarantee the stability of a monopole. If we assign "magnetic charge" to an H gauge field configuration that is nonsingular everywhere in space, nothing can prevent this "magnetic charge" from propagating to spatial infinity as nonabelian radiation (21, 45).

So far we have only considered magnetic monopole configurations in classical gauge field theory. Eventually, we must worry about quantum mechanical effects on the magnetic field. There is really something to worry about, because we believe that nonabelian gauge field theories are confining and have no massless excitations. Therefore the magnetic field cannot survive at arbitrarily large distances, it must be screened at distances larger than the confinement distance scale (21, 31). The mechanism of magnetic screening is briefly discussed in Section 6.4.

Fortunately, since our definition of magnetic charge is topological, it can be applied in the quantum theory, and conservation of magnetic charge is still guaranteed. The gluon fluctuations about the classical long-range magnetic field, which cause the magnetic screening, cannot change the winding number of the classical field.

4.3 *Magnetic Charge of a Topological Soliton*

The object of this section is to generalize the discussion of topological solitons in Section 3 to an arbitrary gauge group, and to demonstrate the general connection between the topology of a soliton and its magnetic charge.

We consider an arbitrary gauge field theory, with gauge group G, which undergoes spontaneous symmetry breakdown to a subgroup H. Acting as an order parameter for this symmetry-breaking pattern is a multiplet of scalar fields Φ, transforming as some (in general reducible) representation of

G. The classical potential $U(\Phi)$ has many degenerate minima, and we identify one arbitrarily chosen minimum as Φ_0. H is the "stability group" of Φ_0, the subgroup of G that leaves Φ_0 invariant.

We find it convenient in this section to assume that G is simply connected, $\pi_1(G) = 0$. This assumption entails no loss of generality, because we may always consider G to be the covering group of a specified Lie group.

We wish to construct finite energy solutions to the classical field equations of this gauge field theory. Therefore, we restrict our attention to field configurations such that Φ approaches a minimum of $U(\Phi)$ at spatial infinity. Barring "accidental" degeneracy, degenerate minima of U not required by G symmetry, the manifold of minima of U is equivalent to the coset space G/H,

$$G/H = \{\Phi : \Phi = \Omega\Phi_0, \Omega \in G\}. \qquad 48.$$

Associated with each finite-energy field configuration is a mapping from the two-dimensional sphere S^2 at spatial infinity into the vacuum manifold G/H. As noted in Section 3.2, a field configuration is a topological soliton if this mapping cannot be continuously deformed to the trivial constant mapping that takes all points on S^2 to Φ_0.

By multiplying by an appropriate constant element of G, we may turn any mapping from S^2 into G/H into a mapping that takes an arbitrarily chosen reference point, the north pole, say, to Φ_0. [This procedure suffers from an ambiguity if H is not connected (19). A consequence of this ambiguity is explained in Section 5.4.] Mappings from S^2 into G/H that take the north pole to Φ_0 fall into topological equivalence classes, homotopy classes, such that mappings in the same class can be continuously deformed into one another. These classes are endowed with a natural group structure, since there is a natural way of composing two mappings that both take the north pole to Φ_0 (see Figure 4). This group is $\pi_2(G/H)$, the "second homotopy group" of G/H (44).

The group $\pi_2(G/H)$ is discrete; its elements are the possible "topological charges" of finite-energy field configurations. Since time evolution is continuous, the discrete topological charge must be a constant of the motion. The classical field theory has a "topological conservation law."

We found that the topological charge of the soliton constructed in Section 3 could be identified with its magnetic charge. We can now show that this identification applies in general (19).

Mappings from the sphere S^2 into the coset space G/H are not very easy to visualize. Fortunately, we can, by a trick, reduce the topological classification of these mappings to the topological classification of closed paths in H. That is, we can reduce the calculation of $\pi_2(G/H)$ to the calculation of $\pi_1(H)$, and we already know how to calculate $\pi_1(H)$.

The trick is to cut the sphere into two hemispheres, along the equator. Each point (ϑ, φ) on the sphere is mapped to some $\Phi(\vartheta, \varphi) \in G/H$. On the upper and lower hemispheres we can find smooth gauge transformations Ω_U and Ω_L that take Φ to Φ_0:

$$\Omega_U(\vartheta, \varphi)\Phi(\vartheta, \varphi) = \Phi_0, \qquad \text{upper}\left(0 \leq \vartheta \leq \frac{\pi}{2}\right),$$

$$\Omega_L(\vartheta, \varphi)\Phi(\vartheta, \varphi) = \Phi_0, \qquad \text{lower}\left(\frac{\pi}{2} \leq \vartheta \leq \pi\right). \tag{49.}$$

On the region where the hemispheres intersect, the equator ($\vartheta = \pi/2$), the gauge transformation $\Omega_U\Omega_L^{-1}$ is defined. It leaves Φ_0 invariant, and is therefore an element of H. So

$$\Omega_U\left(\vartheta = \frac{\pi}{2}, \varphi\right)\Omega_L^{-1}\left(\vartheta = \frac{\pi}{2}, \varphi\right) \equiv \Omega(\varphi) \in H \tag{50.}$$

defines a closed path in H. We have thus found a natural way of associating with each mapping from S^2 into G/H a closed path in H.

This association actually defines a group homomorphism from $\pi_2(G/H)$ to $\pi_1(H)$. If we choose the arbitrary reference point that is mapped to Φ_0 to be a point on the equator, instead of the north pole, then it is obvious that the composition of mappings from S^2 into G/H corresponds to the composition of loops in H, and the group structure is preserved.

The kernel of this homomorphism is trivial because, if $\Omega(\varphi)$ has winding number zero, then it can be continuously deformed to the trivial loop $\Omega(\varphi) = 1$. Therefore, there is a smooth gauge transformation in G, defined on the whole sphere, which takes $\Phi(\vartheta, \varphi)$ to Φ_0. Furthermore, it is known that $\pi_2(G) = 0$ for any compact Lie group G (46). Thus, this gauge transformation can be continuously deformed to a trivial gauge transformation, and the mapping $\Phi(\vartheta, \varphi)$ can be continuously deformed to Φ_0. Therefore, the

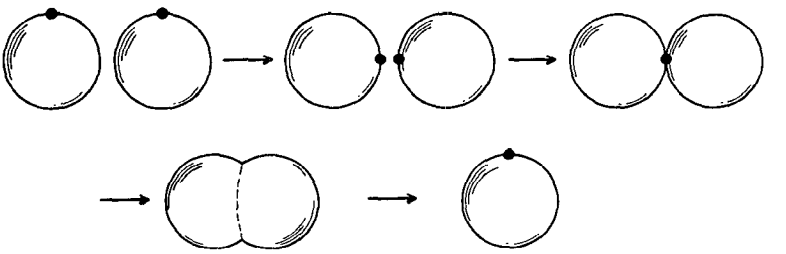

Figure 4 Composition of two mappings from S^2 into G/H that take the north pole (*black dot*) to Φ_0.

homomorphism takes only the identity element of $\pi_2(G/H)$ to the identity element of $\pi_1(H)$.

Moreover, if G is simply connected, we can show that this homomorphism is onto; every element of $\pi_1(H)$ is the image of some element of $\pi_2(G/H)$. Given any loop $\Omega(\varphi)$ in H we can find smooth gauge transformations Ω_U and Ω_L in G, defined on the upper and lower hemispheres, such that

$$\Omega_U\left(\vartheta = \frac{\pi}{2}, \varphi\right) = \Omega(\varphi),$$

$$\Omega_L\left(\vartheta = \frac{\pi}{2}, \varphi\right) = 1, \qquad 51.$$

because we may choose $\Omega_U(\vartheta, \varphi)$ to be the continuous deformation of the loop $\Omega_U(\vartheta = \pi/2, \varphi)$ to the point $\Omega_U(\vartheta = 0)$, which is guaranteed to exist if G is simply connected. Then

$$\Phi(\vartheta, \varphi) = \begin{cases} \Omega_U(\vartheta, \varphi)\Phi_0, & 0 \le \vartheta \le \dfrac{\pi}{2}, \\ \Phi_0 & \dfrac{\pi}{2} \le \vartheta \le \pi, \end{cases} \qquad 52.$$

is a smooth mapping from S^2 into G/H associated with the loop $\Omega(\varphi)$.

We have now established the group isomorphism

$$\pi_2(G/H) = \pi_1(H), \qquad 53.$$

which holds if G is simply connected. [It is easy to see, by slightly modifying the above argument, that the general result is $\pi_2(G/H) = \pi_1(H)/\pi_1(G)$.] As promised, we have found that the topological classification of mappings from S^2 into G/H is equivalent to the topological classification of loops in H.

Since we have already seen that the elements of the group $\pi_1(H)$ specify the possible magnetic charges of a configuration with long-range H gauge fields, we suspect, in view of Equation 53, that the topological charge of a finite-energy field configuration coincides with its magnetic charge. To verify this conjecture, we must consider the long-range gauge field of the soliton.

As we saw in Section 3, a finite-energy field configuration must obey

$$D_i\Phi = (\partial_i - ieA_i^a T^a)\Phi = 0 \qquad 54.$$

on the sphere at spatial infinity, where the T^as are the generators of G in the representation according to which Φ transforms. In the gauge constructed above, in which $\Phi = \Phi_0$ is a constant on the sphere, the only gauge fields that can be excited at large distances are the H gauge fields, those associated

with the generators of G that annihilate Φ_0. The gauge transformation (Equation 49) is nonsingular except on the equator, so the gauge fields A^U and A^L defined on each hemisphere are nonsingular in this gauge. But on the equator they differ by the gauge transformation $\Omega(\varphi)$. The winding number of $\Omega(\varphi)$, which we have now seen is the topological charge of the soliton, is also the magnetic charge defined in Section 4.2. So topological charge equals magnetic charge.

We have now verified the claim in Section 3.1, that any unified gauge theory in which $U(1)_{em}$ is embedded in a spontaneously broken semisimple gauge group necessarily contains magnetic monopoles as topological solitons. There are topologically stable finite-energy solutions to the classical field equations associated with each element of $\pi_2(G/H)$. We have now learned that $\pi_2(G/H) = \pi_1(H)$, if G is chosen to be simply connected, and that $\pi_1(H)$ contains the integers if H has a $U(1)_{em}$ factor. Finally, we have found that the integer labeling an element of $\pi_2(G/H)$ is precisely the magnetic charge in Dirac units. [The Dirac magnetic charge is that corresponding to the minimal $U(1)_{em}$ charge occurring in a representation of the simply connected gauge group G.]

We discuss further applications for our topological formalism when we analyze the examples of Section 5.

4.4 *The Kaluza-Klein Monopole*

Kaluza-Klein theories (47), which unify Einstein's theory of gravitation with other gauge interactions, also contain topological solitons that can be identified as magnetic monopoles. The connection between the topological charges and magnetic charges of these objects are described here. This connection is closely analogous to, but not exactly the same as, that discussed in Section 4.3.

The basic hypothesis of Kaluza-Klein theory is that space-time is not really 4-dimensional, but $(4+n)$-dimensional. The $(4+n)$-dimensional space-time is endowed with a metric satisfying a $(4+n)$-dimensional generalization of Einstein's equations, but n dimensions have become spontaneously compactified to a manifold N with radii of order the Planck length. At low energies, far below the Planck mass, the effects of the microscopic compact dimensions cannot be perceived directly, but a remnant of the underlying $(4+n)$-dimensional theory may survive. The metric on the compact n-dimensional manifold N is typically invariant under some group H of isometries, and the massless fields of the theory include, in addition to the four-dimensional metric, spin-one gauge fields associated with H. These gauge fields are components of the $(4+n)$-dimensional metric that have managed to avoid acquiring large masses upon compactification. Low-energy physics is described by an effective four-dimensional field theory that is an H-gauge theory coupled to gravity.

The classical vacuum solution of the Kaluza-Klein theory is assumed to be $M^4 \times N$, the direct product of four-dimensional Minkowski space and the compact manifold N. Any classical field configuration that approaches the vacuum solution at spatial infinity thus defines an N "bundle" (44) over the sphere at spatial infinity S^2. This bundle has the *local* structure of a direct product $S^2 \times N$; that is, the manifold N sits on top of every point of S^2. But it need not be a direct product *globally*. If the N bundle over S^2 cannot be continuously deformed to the global direct product $S^2 \times N$, then the field configuration cannot be continuously deformed to the vacuum solution; it is a topological soliton.

To perform the topological classification of N bundles over S^2, we cut the sphere S^2 into two hemispheres, along the equator. The N bundles over the two hemispheres D^U and D^L are then easily deformed to direct product bundles, $D^U \times N$ and $D^L \times N$, by performing coordinate transformations on each hemisphere. Along the equator, these two coordinate transformations must differ by a transformation that leaves the geometry of the manifold N invariant; that is, an isometry of N. Thus we can associate with every N bundle over S^2 a loop in the isometry group H. The N bundle over S^2 is topologically nontrivial if and only if the loop in H has a nontrivial winding number. We conclude, as before, that all topological solitons have H-magnetic charges.

The Kaluza-Klein monopole solution has been explicitly constructed in the simplest Kaluza-Klein theory, the five-dimensional theory in which N is a circle and H is U(1) (48, 49). It has some interesting properties. In particular, the four-dimensional constant time slices of both the monopole and antimonopole solution have handles; therefore, a monopole-antimonopole pair has a different toplogy from the vacuum, and cannot annihilate classically.

Generally, one expects a Kaluza-Klein monopole to have a mass m of order $(1/e)M_{\text{Planck}}$. In the five-dimensional theory it has been found that

$$m = \tfrac{1}{4}\alpha^{-1/2}M_{\text{Planck}} \sim 5 \times 10^{19}\ \text{GeV}. \qquad 55.$$

As we see in Section 8.2, this is an interesting mass from an astrophysical viewpoint.

4.5 *Monopoles and Global Gauge Transformations*

It was recently discovered (50, 51) that a global gauge transformation cannot be defined in the vicinity of a magnetic monopole with a nonabelian long-range magnetic field, unless the gauge transformation acts trivially on the long-range field. Implications of this result are considered in Section 6. Here we sketch the proof, which is simple and involves topological concepts that we have already encountered.

A classical field f on a sphere surrounding a magnetic monopole is

defined by specifying smooth functions f_U and f_L on the upper and lower hemispheres, and the gauge transformation Ω, which relates f_U and f_L on the equator, where the hemispheres intersect:

$$f_U\left(\vartheta = \frac{\pi}{2}, \varphi\right) = \Omega(\varphi) f_L\left(\vartheta = \frac{\pi}{2}, \varphi\right), \tag{56.}$$

$\Omega(\varphi)$ is a loop in the gauge group H with a nontrivial winding number; the winding number is the magnetic charge of the monopole.

A local gauge transformation of f on the sphere consists of gauge transformations Ω_U and Ω_L on the two hemispheres that preserve the relations (Equation 56):

$$f_U(\vartheta, \varphi) \to \Omega_U(\vartheta, \varphi) f_U(\vartheta, \varphi), \qquad \text{upper}\left(0 \le \vartheta \le \frac{\pi}{2}\right),$$

$$f_L(\vartheta, \varphi) \to \Omega_L(\vartheta, \varphi) f_L(\vartheta, \varphi), \qquad \text{lower}\left(\frac{\pi}{2} \le \vartheta \le \pi\right),$$

$$\Omega_U\left(\vartheta = \frac{\pi}{2}, \varphi\right) = \Omega(\varphi)\Omega_L\left(\vartheta = \frac{\pi}{2}, \varphi\right)\Omega^{-1}(\varphi). \tag{57.}$$

To define an infinitesimal global gauge transformation on the sphere surrounding the monopole, we must specify a set of generators $\{T^a\}$ of the gauge group H at each point of the sphere. If the transformation is globally defined, the commutation relations satisfied by the generators must be independent of the position on the sphere, but we still have the freedom to perform a local redefinition of the generators of the form

$$T^a(\vartheta, \varphi) = \Sigma(\vartheta, \varphi) T^a \Sigma^{-1}(\vartheta, \varphi), \tag{58.}$$

where $\Sigma \in H$. The redefinition of the generators determined by Σ is called an inner automorphism of the Lie algebra $\mathbf{H}$ of the group H. The group aut $\mathbf{H}$ of inner automorphisms (which is the connected component of the group Aut $\mathbf{H}$ of all automorphisms preserving the Lie algebra $\mathbf{H}$) is evidently isomorphic to H/K where K is the center of H, since the elements of K, and only the elements of K, define trivial automorphisms (52).

A global gauge transformation of f on the sphere must be compatible with Equation 56. Therefore, the generators have the form

$$T_U^a(\vartheta, \varphi) = \Sigma_U(\vartheta, \varphi) T^a \Sigma_U^{-1}(\vartheta, \varphi), \qquad \text{upper}\left(0 \le \vartheta \le \frac{\pi}{2}\right),$$

$$T_L^a(\vartheta, \varphi) = \Sigma_L(\vartheta, \varphi) T^a \Sigma_L^{-1}(\vartheta, \varphi), \qquad \text{lower}\left(\frac{\pi}{2} \le \vartheta \le \pi\right), \tag{59.}$$

where

$$T_U^a\left(\vartheta=\frac{\pi}{2},\varphi\right)=\Omega(\varphi)T_L^a\left(\vartheta=\frac{\pi}{2},\varphi\right)\Omega^{-1}(\varphi),$$

or

$$T^a=\Sigma_U^{-1}\left(\frac{\pi}{2},\varphi\right)\Omega(\varphi)\Sigma_L\left(\frac{\pi}{2},\varphi\right)T^a\Sigma_L^{-1}\left(\frac{\pi}{2},\varphi\right)\Omega^{-1}(\varphi)\Sigma_U\left(\frac{\pi}{2},\varphi\right). \quad 60.$$

We see that $\Sigma_U^{-1}(\pi/2,\varphi)\Omega(\varphi)\Sigma_L(\pi/2,\varphi)$ defines a trivial automorphism, and is therefore an element of the center of H. If H is semisimple, its center is discrete, and we have

$$\Omega(\varphi)=\Sigma_U\left(\vartheta=\frac{\pi}{2},\varphi\right)\Omega_0\Sigma_L^{-1}\left(\vartheta=\frac{\pi}{2},\varphi\right) \quad 61.$$

where Ω_0 is a constant element of the center. If we now allow ϑ, the argument of $\Sigma_U(\Sigma_L)$ to vary smoothly from $\vartheta=\pi/2$ to $\vartheta=0(\vartheta=\pi)$ in Equation 61, we find that the loop $\Omega(\varphi)$ in H can be continuously deformed to a point, and therefore has winding number zero.

We are forced to conclude, if H is semisimple, that a global H transformation can be performed on a sphere only if the sphere encloses no H magnetic charge. In the vicinity of a nonabelian monopole, a global nonabelian gauge transformation cannot be implemented!

There is obviously no topological obstacle to defining globally the generators that commute with Ω. So global gauge transformations of the U(1) magnetic monopole can be performed. In general, we can define any global gauge transformation that acts trivially on the long-range gauge field, and hence leaves Ω intact.

5. EXAMPLES

5.1 *A Symmetry-Breaking Hierarchy*

In order to illustrate the topological principles developed in Section 4, we consider various examples of model gauge theories containing magnetic monopoles. In all these examples, it is possible (40) to construct explicit monopole solutions by using suitable generalizations of the spherically symmetric ansatz of Section 3.3. But here we note only the general properties of the monopoles, and do not exhibit explicit solutions.

Our first example illustrates the importance in monopole theory of the global structure of the unbroken gauge group. Consider a model with gauge group $G=SU(3)$ and a scalar field Φ transforming as the adjoint (octet) representation of G: Φ can be written as a hermitian traceless 3×3 matrix,

which, under a gauge transformation $\Omega(x)$, transforms according to

$$\Phi(x) \to \Omega(x)\Phi(x)\Omega^{-1}(x). \tag{62.}$$

Suppose that Φ acquires the expectation value

$$\langle\Phi\rangle = \Phi_0 = (v)\,\mathrm{diag}\,(\tfrac{1}{2},\tfrac{1}{2},-1),$$

where v is the mass scale of the symmetry breakdown, and the diag $(\frac{1}{2},\frac{1}{2},-1)$ notation denotes a diagonal matrix with the indicated eigenvalues.

The unbroken subgroup H of G, the stability group of Φ_0, is locally isomorphic to SU(2) × U(1). "Locally isomorphic" means that H has the same Lie algebra of infinitesimal generators as SU(2) × U(1). The generators of H are the SU(3) generators that commute with Φ_0. These are the SU(2) generators that mix the two degenerate eigenstates of Φ_0, and also the U(1) generator

$$Q = \mathrm{diag}\,(\tfrac{1}{2},\tfrac{1}{2},-1), \tag{63.}$$

which is proportional to Φ_0, and obviously commutes with it. [The eigenvalues of Q are the U(1) electric charges of the members of the SU(3) triplet, in units of e.]

To perform the topological classification of monopole solutions in this model, we need to determine $\pi_2(G/H) = \pi_1(H)$. So it is not sufficient to know that H has the local structure of the direct product SU(2) × U(1); we must know its global structure. For this purpose, we check to see whether the U(1) subgroup of G generated by Q has any elements in common with the unbroken SU(2) subgroup, other than the identity. And, indeed

$$\exp(i2\pi Q) = \mathrm{diag}\,(-1,-1,1) \tag{64.}$$

is the nontrivial element of the center Z_2 of SU(2). We conclude that

$$H = [\mathrm{SU}(2) \times \mathrm{U}(1)]/Z_2, \tag{65.}$$

where "=" denotes a global isomorphism; there are two elements of SU(2) × U(1) corresponding to each element of H.

The topologically nontrivial loops in H consist of loops winding around the U(1) subgroup of H, and also of loops traveling through the U(1) subgroup from the identity to the element in Equation 64, and returning to the identity through the SU(2) subgroup of H. If we had failed to recognize that H is not globally the direct product SU(2) × U(1), we would have missed the latter set of nontrivial loops, and thus missed half of the monopole solutions in this model.

The monopole with minimal U(1) magnetic charge defines a loop that winds only half-way around U(1); it necessarily also has a Z_2 nonabelian magnetic charge. We anticipated the existence of this solution in our

discussion of the generalized Dirac quantization condition in Section 2.2. Equation 64 implies that objects with trivial SU(2) "duality" have integer U(1) charge, although objects with nontrivial duality can have half-integer charge. According to the discussion of Section 2.2, a monopole can exist with the Dirac U(1) magnetic charge $g_D = 1/2e$, provided that it also carries an SU(2) magnetic charge. Alternatively, the charge carried by this monopole can be regarded, in an appropriate gauge, as the U(1)′ charge generated by

$$Q' = T^3 + Q = \text{diag}(1, 0, -1), \qquad 66.$$

where

$$T^3 = \text{diag}(\tfrac{1}{2}, -\tfrac{1}{2}, 0), \qquad 67.$$

is an SU(2) generator. The monopole with minimal U(1) magnetic charge defines a closed loop in U(1)′.

In realistic unified gauge theories, spontaneous symmetry breakdown typically occurs at two or more mass scales differing by many orders of magnitude. To illustrate the effect of such a symmetry-breaking hierarchy on magnetic monopoles, let us imagine that the $G = \text{SU}(3)$ gauge symmetry of our model breaks down in two stages, first to $H_1 = [\text{SU}(2) \times \text{U}(1)]/Z_2$ at mass scale v_1, then to $H_2 = \text{U}(1)$ at mass scale $v_2 \ll v_1$,

$$G = \text{SU}(3) \xrightarrow{v_1} H_1 = [\text{SU}(2) \times \text{U}(1)]/Z_2 \xrightarrow{v_2} H_2 = \text{U}(1). \qquad 68.$$

The effect of the second stage of symmetry breakdown on the monopoles generated by the first stage depends on which U(1) subgroup of H_1 remains unbroken at the second stage (53).

First, suppose that H_2 is the U(1) subgroup generated by

$$Q_2 = Q' = \text{diag}(1, 0, -1). \qquad 69.$$

Since this is the same charge as that carried by the monopole associated with the $G \to H_1$ breakdown at mass scale v_1, the breakdown at the much lower mass scale v_2 has no significant effect on the monopole.

But if H_2 is the U(1) subgroup generated by

$$Q_2 = Q = \text{diag}(\tfrac{1}{2}, \tfrac{1}{2}, -1), \qquad 70.$$

the monopole is significantly affected, for the only monopole solutions now have twice the U(1) magnetic charge allowed by the $G \to H_1$ breakdown.

What would happen to the minimal G/H_1 monopole if we varied the parameters of the model so as smoothly to turn on the second symmetry-breaking scale v_2? This question is not entirely academic, because the H_1 symmetry is expected to be restored at sufficiently high temperature,

$T \gg v_2$. As the temperature is lowered, a phase transition occurs at $T \sim v_2$ in which H_1 becomes spontaneously broken. We might be interested in what happens to the minimal G/H_1 monopoles during this phase transition, especially since a phase transition like this one may have occurred in the very early universe.

A reasonable guess is that pairs of minimal G/H_1 monopoles or monopole-antimonopole pairs become connected by magnetic flux tubes, and form composite objects with either twice the minimal U(1) magnetic charge or zero magnetic charge. To verify this guess, we need a mathematical criterion to determine when such flux tubes occur.

A magnetic flux tube in three spatial dimensions, a static solution to the field equation with finite energy per unit length, may be regarded as a topological soliton in two spatial dimensions with finite energy. The topological classification of these solitons is very similar to the classification of monopoles in Section 4.3; so similar that we need only sketch the analysis.

In a gauge theory with gauge group G and unbroken group H, the finite-energy two-dimensional field configurations define mappings from the circle S^1 at (two-dimensional) spatial infinity into the vacuum manifold G/H, and are classified by the first homotopy group $\pi_1(G/H)$. To facilitate the calculation of $\pi_1(G/H)$, we cut the circle open at $\varphi = 0$, and find a gauge transformation $\Omega(\varphi) \in G$ that rotates the order parameter $\Phi(\varphi)$ to the standard values Φ_0 for all φ. The discontinuity of this gauge transformation at $\varphi = 0$ is an element of H,

$$\Omega(\varphi = 0)\Omega^{-1}(\varphi = 2\pi) = \Omega_0 \in H. \tag{71.}$$

We thus obtain a group homomorphism from $\pi_1(G/H)$ into a group called $\pi_0(H)$. The elements of $\pi_0(H)$ are equivalence classes of elements of H, defined such that $\Omega_1, \Omega_2 \in H$ are in the same class if there is a continuous path in H from Ω_1 to Ω_2. Group multiplication in H defines the group structure in $\pi_0(H)$.

It is easy to see that the homomorphism from $\pi_1(G/H)$ into $\pi_0(H)$ has a trivial kernel, if G is simply connected, and is onto, if G is connected. So we have the isomorphism

$$\pi_1(G/H) = \pi_0(H), \tag{72.}$$

which holds if G is connected and simply connected.

To apply this result to the symmetry-breaking pattern (Equation 68), with the H_2 generator Q_2 given by Equation 70, we note that the U(1) factor of H_1 is not affected by the second stage of symmetry breakdown, so that the flux tubes are classified by

$$\pi_1[\mathrm{SU}(2)/Z_2] = \pi_0(Z_2) = Z_2. \tag{73.}$$

As we expected, there are Z_2 flux tubes, to which the nonabelian SU(2) magnetic flux becomes confined, generated by the spontaneous breakdown of the SU(2) gauge symmetry. The thickness and energy per unit length of the flux tubes are determined by the lower symmetry-breaking scale v_2; the thickness is of order $(ev_2)^{-1}$, and the energy per unit length is of order v_2^2 (54).

The flux tubes link each G/H_1 monopole with minimal H_1 magnetic charge to either another monopole or an antimonopole, since the monopole and antimonopole carry the same Z_2 charge. The bound pairs of monopoles have the minimal H_2 magnetic charge allowed by the Dirac quantization condition.

Finally, suppose that the unbroken U(1) group H_2 is generated by

$$Q_2 = T_3 = \text{diag}(\tfrac{1}{2}, -\tfrac{1}{2}, 0). \tag{74.}$$

In this case H_2 is contained in $\text{SU}(2) \subset H_1$ and the symmetry breakdown $H_2 \to H_1$ can be represented by

$$\begin{array}{ccc} H_1 = \text{SU}(2) & \times & \text{U}(1) \\ \downarrow & & \downarrow \\ H_2 = \text{U}(1) & & 1 \end{array}. \tag{75.}$$

The flux tubes associated with the breakdown of H_1 are classified by

$$\pi_1[\text{U}(1)] = Z. \tag{76.}$$

These are Z flux tubes to which the U(1) magnetic flux becomes confined, and therefore no heavy monopoles with mass of order v_1/e can survive when v_2 turns on; all heavy monopoles become bound to antimonopoles by the flux tubes. Since $\pi_2(G/H_2) = Z$, there must still be stable, but light (mass of order v_2/e), monopoles associated with the symmetry breakdown $H_1 \to H_2$.

We see that magnetic monopoles generated at a large symmetry-breaking mass scale may be affected by a small symmetry-breaking mass scale in various ways. The monopoles may survive intact, may become bound by flux tubes into monopole-antimonopole pairs, or may become bound into both monopole-antimonopole pairs and clusters of n monopoles. And, of course, new monopoles might also be generated at the smaller mass scale.

5.2 *A Z_2 Monopole*

We encountered above a monopole carrying both a U(1) magnetic charge and a nonabelian magnetic charge. Of course, it is also possible for a monopole to carry only a nonabelian charge.

For example, consider a model with gauge group $G = \text{SU}(3)$ and a scalar field Φ transforming as the symmetric tensor representation of G. Φ can be

written as a symmetric 3×3 matrix, which, under a gauge transformation $\Omega(x)$, transforms according to

$$\Phi(x) \to \Omega(x)\Phi(x)\Omega^{T}(x). \tag{77.}$$

If Φ acquires the expectation value

$$\langle \Phi \rangle = \Phi_0 = v\mathbb{1}, \tag{78.}$$

then G is spontaneously broken to $H = SO(3)$. The monopoles of this model are classified by

$$\pi_2(G/H) = \pi_1[SO(3)] = Z_2. \tag{79.}$$

They are Z_2 monopoles carrying SO(3) magnetic charges. The monopole and antimonopole are indistinguishable.

It is interesting to examine the fate of these monopoles if there is a symmetry-breaking hierarchy of the form (55)

$$G = SU(3) \xrightarrow{v_1} H_1 = SO(3) \xrightarrow{v_2} H_2 = U(1), \tag{80.}$$

where $H_2 = U(1) \subset SO(3)$ is generated by

$$Q = \text{diag}(\tfrac{1}{2}, -\tfrac{1}{2}, 0). \tag{81.}$$

There will, of course, be $\pi_2(H_1/H_2)$ monopoles generated by the second stage of symmetry breakdown. These are light monopoles, with core radius of order $(ev_2)^{-1}$ and mass of order v_2/e, which define topologically nontrivial loops in H_2 that can be contracted to a point in H_1.

But the light monopoles are not all the monopoles of this model; $\pi_2(G/H_2)$ is larger than $\pi_2(H_1/H_2)$, because there are topologically nontrivial loops in H_2 that cannot be contracted to a point in H_1, but are contractible in G. Thus, there are monopoles with half the magnetic charge of the minimal $\pi_1(H_1/H_2)$ monopole that are generated by the first stage of symmetry breakdown. These are heavy monopoles with a core radius of order $(ev_1)^{-1}$ and a mass of order v_1/e. They are just the Z_2 monopoles, which have been converted into Z monopoles with the Dirac magnetic charge by the physics of the second stage of symmetry breakdown. If we turn on v_2 smoothly, the Z_2 monopole, which is equivalent to its antiparticle, must choose the sign of its U(1) magnetic charge at random (55).

The heavy monopole has two cores, and most of its mass resides on its tiny inner core. But if two heavy monopoles are brought together, their inner cores can annihilate, and only the outer cores need survive. So the doubly charged light monopole can be regarded as a very tightly bound composite state of two, singly charged, heavy monopoles.

5.3 *The SU(5) and SO(10) Models*

The monopoles of the realistic grand unified models based on the gauge groups $G = \mathrm{SU}(5)$ and $G = \mathrm{SO}(10)$ have many features in common with the simpler examples considered above.

The SU(5) model (4) is the simplest gauge theory uniting the $\mathrm{SU}(3)_c$ gauge group of the strong interactions with the $[\mathrm{SU}(2) \times \mathrm{U}(1)]_{ew}$ gauge group of the electroweak interaction. This model undergoes symmetry breakdown at two different mass scales,

$$G = \mathrm{SU}(5) \xrightarrow{v_1} H_1 = \{\mathrm{SU}(3)_c \times [\mathrm{SU}(2) \times \mathrm{U}(1)]_{ew}\}/Z_6$$

$$\xrightarrow{v_2} H_2 = [\mathrm{SU}(3)_c \times \mathrm{U}(1)_{em}]/Z_3. \qquad 82.$$

Here $v_2 \sim 250$ GeV is the mass scale of the electroweak symmetry breakdown, and $v_1 \sim 10^{15}$ GeV is the mass scale of unification.

The order parameter for the symmetry breakdown at mass scale v_1 is a scalar field Φ transforming as the adjoint representation of G, which acquires the expectation value

$$\langle \Phi \rangle = \Phi_0 = v_1 \,\mathrm{diag}(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}, -\tfrac{1}{2}, -\tfrac{1}{2}). \qquad 83.$$

The stability group H of G is locally isomorphic to $\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)$, where SU(3) acts on the three degenerate eigenvectors of Φ_0/v_1 with eigenvalue $\frac{1}{3}$, and SU(2) acts on the two degenerate eigenvectors with eigenvalue $-\frac{1}{2}$. The unbroken U(1) is generated by

$$Q = \mathrm{diag}(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}, -\tfrac{1}{2}, -\tfrac{1}{2}), \qquad 84.$$

and, since

$$\exp(i2\pi Q) = \mathrm{diag}[\exp(i2\pi/3), \exp(i2\pi/3), \exp(i2\pi/3), -1, -1]. \qquad 85.$$

we see that this U(1) contains the center of $\mathrm{SU}(3) \times \mathrm{SU}(2)$, so that the unbroken group is actually $H_1 = [\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)]/Z_6$.

Equation 85 ensures that any object with trivial SU(3) triality and SU(2) duality has integer U(1) charge, in units of e. Thus, there exists a magnetic monopole in this model with the Dirac U(1) magnetic charge $g_D = 1/2e$, which also carries a Z_3 color magnetic charge and a Z_2 SU(2) magnetic charge. In an appropriate gauge, we may regard the magnetic charge carried by the monopole to be a U(1)′ charge generated by

$$Q' = Q + Q_{\mathrm{weak}} + Q_{\mathrm{color}} = \mathrm{diag}(0, 0, 1, 0, -1), \qquad 86.$$

where

$$Q_{\mathrm{weak}} = \mathrm{diag}(0, 0, 0, \tfrac{1}{2}, -\tfrac{1}{2}), \qquad 87.$$

is an SU(2) generator and

$$Q_{\text{color}} = \text{diag}(-\tfrac{1}{3}, -\tfrac{1}{3}, \tfrac{2}{3}, 0, 0) \qquad 88.$$

is an SU(3) generator. Since Q' has integer eigenvalues, a monopole with U(1)′ magnetic charge $g = g_D = 1/2e$ is consistent with the Dirac quantization condition.

The electroweak symmetry breakdown at mass scale v_2 leaves unbroken the $U(1)_{em}$ subgroup of $[SU(2) \times U(1)]_{ew}$ generated by

$$Q_{em} = Q + Q_{weak} = \text{diag}(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}, 0, -1). \qquad 89.$$

Since $\exp(i2\pi Q_{em})$ is a nontrivial element of the center of $SU(3)_c$, the unbroken subgroup is $H_2 = [SU(3) \times U(1)]/Z_3$, and the monopole with minimal $U(1)_{em}$ magnetic charge still carries the U(1)′ charge generated by Q'.

The structure of the SU(5) monopole is not much affected by the electroweak symmetry breakdown, because the magnetic charge carried by the monopole is not changed by this breakdown. There are no W and Z fields excited inside an electroweak core with a radius of order $(ev_2)^{-1} \sim M_W^{-1}$, at least in the classical approximation. The true core of the monopole has a radius of order $(ev_1)^{-1} \sim 10^{-28}$ cm and the mass of the monopole is of order $(v_1/e) \sim 10^{16}$ GeV.

That the electroweak SU(2) × U(1) gauge symmetry is restored within a distance M_W^{-1} of the center of the monopole has some important consequences, though. For one thing, two monopoles with a separation much less than M_W^{-1} may orient their magnetic charges in orthogonal directions in SU(3) × SU(2) × U(1), and reduce their Coulomb repulsion to zero. For an appropriate choice of parameters, it is then possible for the attractive force between the monopoles generated by scalar exchange to cause a stable two-monopole bound state to form, with twice the minimal $U(1)_{em}$ magnetic charge (56). Also the quantum mechanical fluctuations of the W and Z fields within a distance M_W^{-1} of the center of the monopole influence the scattering of fermions by monopoles, as we see in Section 7.

The SO(10) model (57) is the next simplest realistic grand unified theory, after the SU(5) model. There are several possible choices for the symmetry-breaking hierarchy of the SO(10) model, and the properties of its monopoles depend on this choice. Rather than enumerate all the possibilities, let us focus on one particularly interesting case.

The group SO(10) is not simply connected, but has the simply connected covering group Spin(10). The center of Spin(10) is Z_2, and its 16-dimensional spinor representation is a double-valued representation of $SO(10) = \text{Spin}(10)/Z_2$. All representations of Spin(10) can be constructed from direct products of 16s.

Let us suppose that the order parameter for the first stage of symmetry

breakdown in the SO(10) model is a scalar field Φ that transforms as the 54-dimensional representation of SO(10): Φ can be written as a traceless symmetric 10×10 matrix transforming according to

$$\Phi(x) \to \Omega(x)\Phi(x)\Omega^{T}(x), \tag{90.}$$

where $\Omega(x) \in \mathrm{SO}(10)$. If Φ acquires the expectation value

$$\langle \Phi \rangle = \Phi_0 = v_1 \operatorname{diag}(2,2,2,2,2,2,-3,-3,-3,-3), \tag{91.}$$

then the unbroken subgroup H is locally isomorphic to $\mathrm{SO}(6) \times \mathrm{SO}(4)$. This group is, in turn, locally isomorphic to the direct product of SU(4), the covering group of SO(6), and $\mathrm{SU}(2) \times \mathrm{SU}(2)$, the covering group of SO(4).

To determine the global structure of the unbroken group, we check for nontrivial elements of $\mathrm{SU}(4) \times \mathrm{SU}(2) \times \mathrm{SU}(2)$ that act trivially in Spin(10). Since the fundamental spinor representation of Spin(10) transforms under $\mathrm{SU}(4) \times \mathrm{SU}(2) \times \mathrm{SU}(2)$ as

$$\mathbf{16} \to (4,1,2) + (\bar{4},2,1), \tag{92.}$$

we see that the element $(-\mathbb{1}_4, -\mathbb{1}_2, -\mathbb{1}_2)$ of $\mathrm{SU}(4) \times \mathrm{SU}(2) \times \mathrm{SU}(2)$ does act trivially on the spinor. Thus, the symmetry-breaking pattern is (58)

$$G = \mathrm{Spin}(10) \xrightarrow{v_1} H_1 = [\mathrm{SU}(4) \times \mathrm{SU}(2) \times \mathrm{SU}(2)]/Z_2. \tag{93.}$$

The monopoles arising from this symmetry breakdown are Z_2 monopoles carrying SU(4) and $\mathrm{SU}(2) \times \mathrm{SU}(2)$ magnetic charges, classified by $\pi_2(G/H_1) = \pi_1(H_1) = Z_2$.

Now suppose that, at a lower mass scale v_2, the symmetry breakdown

$$\begin{aligned} H_1 &= [\mathrm{SU}(4) \times \mathrm{SU}(2) \times \mathrm{SU}(2)]/Z_2 \xrightarrow{v_2} H_2 \\ &= [\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)]/Z_6 \end{aligned} \tag{94.}$$

occurs. [The order parameter could be a scalar field transforming as the 16-dimensional spinor representation of SO(10).] H_2 is exactly the same as the unbroken gauge group of the SU(5) model, and the monopole with the minimal U(1) magnetic charge in this SO(10) model also carries SU(3) and SU(2) magnetic charges, just like the monopole of the SU(5) model.

But, as in the example of Section 5.2, the doubly charged monopole in this model is lighter than the monopole with minimal charge (59). The minimal monopole defines a loop in H_2 that cannot be contracted to a point in H_1, but can be in G. So the core of this monopole has a radius of order $(ev_1)^{-1}$, and its mass is of order (v_1/e). The doubly charged monopole, however, has no SU(2) magnetic charge, and it defines a loop in H_2 that can be contracted to a point in H_1. It arises from the breakdown of H_1 to H_2,

and has a core radius of order $(ev_2)^{-1}$ and a mass of order (v_2/e). Neither the minimal monopole nor the doubly charged monopole is much affected by the subsequent breakdown of H_2 to $H_3 = [SU(3) \times U(1)]/Z_3$.

In general, a grand unified theory with a complicated symmetry-breaking hierarchy may possess several stable monopoles with widely disparate masses, the monopole of minimal $U(1)_{em}$ charge being the heaviest. The SO(10) model described here is the simplest realistic example illustrating this possibility.

5.4 *Monopoles and Strings*

In Section 5.1, we saw that there are model gauge theories in which magnetic flux becomes confined to topologically stable tubes, or "strings." In some models, the strings can end on magnetic monopoles, and cause monopoles and antimonopoles to form bound pairs connected by strings. In other models, the strings cannot end; they become either infinite open strings or closed loops of string.

Our last example is a model containing both monopoles and strings (60). Although the strings in this model cannot end on monopoles, they have interesting long-range interactions with monopoles. A monopole that winds once around a string becomes an antimonopole! The model has gauge group $G = SO(3)$ and a scalar order parameter Φ transforming as the 5-dimensional representation of G: Φ can be written as a traceless symmetric 3×3 matrix transforming as

$$\Phi(x) \to \Omega(x)\Phi(x)\Omega^{T}(x), \tag{95.}$$

where $\Omega(x) \in SO(3)$. If Φ acquires the expectation value

$$\langle\Phi\rangle = \Phi_0 = v\ \mathrm{diag}(1, 1, -2), \tag{96.}$$

then the unbroken subgroup H is locally isomorphic to the SO(2) subgroup of rotations about the "z-axis," generated by

$$Q = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{97.}$$

But H actually has a disconnected component, because

$$\Phi_0 = \Omega_0\Phi_0\Omega_0^{T}, \tag{98.}$$

where

$$\Omega_0 = \mathrm{diag}(1, -1, -1) \tag{99.}$$

is a 180° rotation about the "x-axis." The symmetry-breaking pattern is

$$G = SU(2) \xrightarrow{v} H = U(1) \times Z_2,$$

and the vacuum manifold G/H is topologically equivalent to a two-dimensional sphere with antipodal points identified. [The unbroken subgroup of the SO(10) model of Section 5.3 has a similar Z_2 factor, which we did not bother to point out there (61).]

Since $\pi_2(G/H) = Z$, this model has magnetic monopoles, just like the monopoles of the SO(3) gauge theory discussed in Section 3. But these monopoles have a peculiar new feature. Since a 180° rotation about the x-axis changes the sense of a rotation about the z-axis, we have

$$\Omega_0 Q \Omega_0^{\mathrm{T}} = -Q. \qquad 100.$$

Therefore, the sign of an electric or magnetic charge can be changed by a gauge transformation in H, and there is no gauge-invariant way to distinguish a monopole from an antimonopole. A "hedgehog" is not different from an "antihedgehog," because the order parameter is a "headless" vector in three-dimensional space, identified with the vector pointing in the opposite direction. We can, however, distinguish a pair of monopoles (or antimonopoles) from a monopole-antimonopole pair; the ambiguity afflicts only the sign of total charge, not the relative charge of two objects (19).

This model also contains topologically stable strings, because $\pi_1(G/H) = \pi_0(H) = Z_2$. If we perform a gauge transformation $\Omega(\varphi)$ that rotates Φ to Φ_0 at all points on the circle at spatial infinity enclosing a string, as described in Section 5.1, then this gauge transformation must have a discontinuity, at some value of φ, by an element of H in the connected component of Ω_0. The two-dimensional cross section of a string is indicated in Figure 5, where the order parameter is represented by an arrow, with the understanding that arrows pointing in opposite directions represent the same value of the order parameter.

According to Equation 100, the magnetic charge of a monopole changes sign when it crosses the discontinuity in $\Omega(\varphi)$. The location of the discontinuity is of course gauge dependent. But any monopole trajectory that winds once around the string must cross the discontinuity an odd number

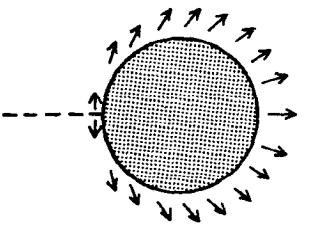

Figure 5 Cross section of a string.

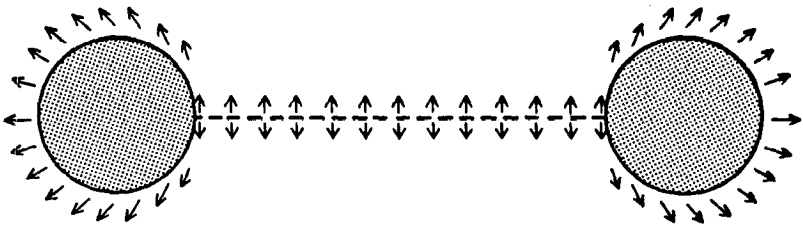

Figure 6 Cross section of a magnetically charged loop of string.

of times. We thus obtain the gauge-invariant result that a monopole that winds once around the string becomes an antimonopole (60).

There is a local criterion for distinguishing between a pair of monopoles (or antimonopoles) and a monopole-antimonopole pair; we can bring the two objects together and see whether they will annihilate or not. But this criterion is not globally well defined if strings are present. Whether they annihilate or not depends on how many times the monopoles wind around the strings before they are brought together.

Magnetic charge is conserved, so the magnetic charge lost by a monopole that winds around a string cannot disappear; it must be transferred to the string. If the string is open, the magnetic charge is transmitted to infinity along the string. But if the string is a closed loop, a finite magnetic charge density remains on the string, after it interacts with the monopole.

A cross section of a magnetically charged loop of string is sketched in Figure 6; the order parameter on a large sphere surrounding this loop is in a hedgehog configuration. The loop is a peculiar highly excited monopole, whose core has been distorted into a ring of radius R and thickness $(ev)^{-1}$. Its energy is of order v^2R, plus a magnetic excitation energy of order $1/e^2R$. The string tension of the loop causes it to oscillate with a period of order R.

The electric charge of a particle that winds around a string must also change sign, so a loop of string must be capable of supporting electric charge excitations, as well as magnetic charge excitations. The electric charge excitations of a string loop arise in much the same way as the dyonic excitations of a monopole, which are the subject of Section 6.

6. DYONS

6.1 *Semiclassical Quantization*

In Section 3, we constructed the time-independent monopole solution to the classical field equations in an SU(2) gauge theory. Now we consider the semiclassical quantization of this soliton.

The semiclassical expansion is an expansion in $\hbar$, where $\hbar^{-1}$ is a

parameter multiplying the whole action. By rescaling the fields, we can write the Lagrangian (Equation 17) as

$$\mathscr{L} = \frac{1}{e^2}\left[-\frac{1}{4}F^a_{\mu\nu}F^{\mu\nu a} + \frac{1}{2}D_\mu\Phi^a D^\mu\Phi^a_a - \frac{1}{8}\left(\frac{\lambda}{e^2}\right)(\Phi^a\Phi^a - M_W^2)^2\right], \qquad 101.$$

where $F_{\mu\nu}$ and $D_\mu\Phi$ no longer depend on the gauge coupling e. We thus see that the semiclassical expansion is an expansion in e^2 with M_W and λ/e^2 fixed. In the classical limit $\hbar \to 0$, the size of the monopole remains fixed while its mass diverges like $\hbar^{-1}$.

Semiclassical quantization is carried out (in the gauge $A_0 = 0$), by expanding the Hamiltonian about the stable time-independent monopole solution. In order e, the monopole possesses a spectrum of positive-frequency vibrational excitations, which can be interpreted as meson states in the vicinity of the monopole.

Expanding about the classical solution, we also discover zero-frequency modes; these are associated with unbroken exact symmetries of the theory that act nontrivially on the solution. The time-independent monopole solutions form a degenerate set, and the zero-frequency modes are infinitesimal displacements in the manifold of degenerate solutions.

For example, the monopole solution is not translation-invariant; therefore, it has translational zero modes. The translational modes are easily quantized. To obtain an eigenstate of the Hamiltonian, we construct states that transform as irreducible unitary representations of the translation group; that is, plane wave states labeled by a momentum $\mathbf{p}$. For fixed $\mathbf{p}$, the energy of a monopole plane wave state is $O(e^2)$ in the semiclassical expansion, because the monopole mass m is $O(1/e^2)$:

$$E_{\mathbf{p}} = \sqrt{m^2 + p^2} = m + p^2/2m + \cdots = m + O(e^2). \qquad 102.$$

If the classical monopole solution were not rotationally invariant, it would have a moment of inertia of order $1/e^2$, and rotational excitations with energy of order e^2. But, because the monopole solution is rotationally invariant, there are no such rotational excitations.

A soliton can have zero-frequency modes associated with internal symmetries as well as space-time symmetries. In fact, the monopole solution is not invariant under a global $U(1)_{em}$ charge rotation, because the charged fields $W^\pm$ are excited in the monopole core. (Although the physical states in $A_0 = 0$ gauge are required to be invariant under time-independent local gauge transformations with compact support, they need not be invariant under global gauge transformations.) To quantize the charge rotation degree of freedom, we diagonalize the Hamiltonian by constructing irreducible representations of $U(1)_{em}$; that is, eigenstates of the electric charge Q.

Since $U(1)_{em}$ is compact, $\exp(i2\pi Q) = 1$, the eigenvalues of Q are integers. [Q is the electric charge in units of e. Half-odd-integer charge cannot occur because the monopole is invariant under the center Z_2 of SU(2).] Thus, the quantum-mechanical excitations of the fundamental monopole include *dyons*, particles that carry both magnetic and electric charge. The dyons have arisen automatically, from the semiclassical quantization of the global charge rotation degree of freedom of the monopole (62).

To determine the energies of the dyon states, we must compute the "moment of inertia" I of the monopole associated with a global charge rotation. The kinetic energy of a monopole undergoing a time-dependent charge rotation $\Omega(t) = \exp[i\vartheta(t)]$ has the form

$$L = \frac{1}{2e^2} I\dot{\vartheta}^2, \qquad 103.$$

where I is of order M_W^{-1}, or of order one in the semiclassical expansion. [The explicit computation of I involves some technical subtleties, as explained in (63–65).] The electric charge operator, the generator of a charge rotation, is the angular momentum conjugate to ϑ,

$$Q = \frac{\partial L}{\partial \dot{\vartheta}} = \frac{1}{e^2} I\dot{\vartheta} \qquad 104.$$

and the Hamiltonian may be written as

$$\mathcal{H} = \frac{e^2 Q^2}{2I}. \qquad 105.$$

The dyon excitations are split from the monopole ground state by an amount that is of order the electrostatic energy of a charge eQ localized on the monopole core, where Q is an integer.

The monopoles that occur in more complicated models, like those considered in Section 5, also have dyon excitations. One's naive expectation, based on the above discussion, is that these dyon states will transform as irreducible representations of the unbroken gauge group. In a realistic grand unified theory, with unbroken group $SU(3)_C \times U(1)_{em}$, one then expects the dyons to form color multiplets (66).

But we have already seen that this expectation is wrong. In Section 4.5 we found that global color rotations of a monopole that act nontrivially on its long-range field actually cannot be implemented. Therefore, the dyon excitations obtained by semiclassical quantization of a monopole with a color magnetic field need only have definite values of color hypercharge, the $SU(3)_c$ charge that commutes with the magnetic charge. They do not form complete color multiplets (64, 67).

The dyon excitations associated with color rotations of the monopole

that act nontrivially on its long-range magnetic field fail to appear because they cannot be supported by the monopole core. These excitations are carried out to large distances by the nonabelian magnetic field, and are lost in the gluon continuum. They do appear explicitly, however, in the excitation spectrum of a widely separated monopole-antimonopole pair, with energy splittings inversely proportional to the separation of the pair (65).

6.2 *The Anomalous Dyon Charge*

In Section 2 we noted that the Dirac quantization condition permits dyons to have an anomalous electric charge characterized by a *CP*-violating angular parameter ϑ. We have seen that the semiclassical quantization procedure generates dyons with integer electric charge, in a *CP*-conserving theory. One wonders whether it is possible to introduce *CP* violation such that the dyons acquire anomalous charges.

In fact, it is possible (68). Let us consider adding to the Lagrange density of electrodynamics the *CP*-violating term

$$\mathscr{L}_\vartheta = \frac{\vartheta e^2}{4\pi^2}\mathbf{E}\cdot\mathbf{B}, \qquad 106.$$

where ϑ is a free parameter. In the absence of magnetic monopoles, this term is a total divergence, and has no physical consequences. But if a magnetic monopole is present, **B** is not the curl of a nonsingular vector potential, and this term has significant consequences.

Let us consider the effect of Equation 106 on the electric charge of a point monopole fixed at the origin (68, 69). In the $A_0 = 0$ gauge, the extra term modifies the momentum conjugate to A_i, and therefore also modifies the generator of an infinitesimal gauge transformation. In this gauge, physical states are invariant under finite time-independent gauge transformations that act trivially at spatial infinity. Let $\exp(i2\pi Q_\Lambda)$ be the operator that implements the gauge transformation

$$\Omega(\mathbf{r}) = \exp[i2\pi\Lambda(\mathbf{r})], \qquad 107.$$

where

$$\Lambda(\mathbf{r}=0) = 0$$

$$\Lambda(|\mathbf{r}| = \infty) = 1. \qquad 108.$$

Then, acting on physical states, we must have

$$n = Q_\Lambda = \frac{1}{2\pi}\int \mathrm{d}^3r \frac{\partial L}{\partial\partial_0 A_i}\delta A_i = \int \mathrm{d}^3r\left(\mathbf{E}+\frac{\vartheta e^2}{4\pi^2}\mathbf{B}\right)\cdot\left(\frac{1}{e}\nabla\Lambda\right)$$

$$= \int_{r=\infty} \mathrm{d}^2\mathbf{S}\cdot\left(\frac{1}{e}\mathbf{E}+\frac{\vartheta e}{4\pi^2}\mathbf{B}\right) - \frac{1}{e}\int \mathrm{d}^3r\Lambda\nabla\cdot\left(\mathbf{E}+\frac{e^2}{4\pi^2}\vartheta\mathbf{B}\right), \qquad 109.$$

where n is an integer, and the last equality has been obtained from an integration by parts. Since Equation 109 is satisfied by any $\Lambda(\mathbf{r})$ consistent with Equations 108, the volume integral vanishes and we obtain

$$Q = n - \frac{\vartheta}{2\pi} M, \qquad 110.$$

where Q is the electric charge of the monopole in units of e, and M is the magnetic charge in units of $g_D = 1/2e$. We have succeeded in reproducing Equation 16; now, through Equation 106, we have a dynamical interpretation of ϑ.

Since the charge spectrum (Equation 110) is unchanged when ϑ increases by 2π, one is tempted to interpret ϑ as an angular variable, and claim that the dyons parametrized by (n, ϑ) and $(n+1, \vartheta+2\pi)$, which have the same charge, are actually the same object. It is easy to see that this interpretation is correct (67, 68). Quantization of the charge rotation degree of freedom of the monopole in a theory with the term in Equation 106 is evidently equivalent to quantization in a theory without such a term, but subject to the condition

$$\exp(i2\pi Q_\Lambda) = \exp(i\vartheta), \qquad 111.$$

where $\exp(i2\pi Q_\Lambda) = \exp(i2\pi Q)$ implements a gauge transformation satisfying Equation 108. So we can think of $\exp(i\vartheta)$ as an arbitrary phase by which physical states are multiplied when acted on by a "large" gauge transformation with $\Lambda(|\mathbf{r}| = \infty) = 1$. Obviously, ϑ is an angle.

An angle ϑ can be associated with any gauge group; the dyon excitations of nonabelian monopoles, as well as abelian monopoles, may carry anomalous electric charges (67, 70). This observation seems paradoxical at first, because we know that the nontrivial commutation relations satisfied by the generators of a nonabelian gauge group require the eigenvalues of the generators to be quantized. But we have already seen how this paradox is resolved. Global gauge transformations of a nonabelian monopole cannot be defined; therefore, the dyon excitations need not form complete representations of the gauge group, and the peculiar values of the electric charge are allowed (67).

The discovery of the anomalous electric charge of the dyon has led to deep insights into the interactions of dyons and fermions, as discussed in Section 7.

6.3 *Composite Dyons*

The dyons we have considered so far are quantum mechanical excitations of a fundamental monopole. Another type of dyon is a composite state of a magnetic monopole and an electrically charged particle. A composite dyon

has a peculiar property—it can carry half-odd-integral orbital angular momentum.

To understand this phenomenon, we consider a monopole (magnetic charge g) fixed at the origin interacting with a charged particle (electric charge e), which moves in the x–y plane. Let us imagine that the magnetic charge of the monopole turns on gradually (71). Then the z-component of the orbital angular momentum of the charged particle changes, according to Faraday's law, by

$$\Delta L_z = -(e/2\pi)\Delta\Phi_z. \qquad 112.$$

Here $\Delta\Phi_z$ is the change in the magnetic flux through a surface bounded by a circular loop in the x–y plane, centered on the monopole. But we may choose the surface bounded by the loop to pass either above or below the monopole, and the fluxes through the two surfaces differ by $4\pi g$ (not counting, of course, the flux carried by the Dirac string).

Since the choice of a surface bounding the loop is arbitrary, we must demand that the spectrum of L_z levels not depend on the choice. The spacings between L_z levels are integers; therefore, the ambiguity in L_z is undetectable if it is an integer; that is, if

$$n = (e/2\pi)(4\pi g) = 2eg. \qquad 113.$$

In yet another way, we have found that the quantum mechanics of a charged particle interacting with a magnetic monopole is consistent only if the Dirac quantization condition is satisfied.

We have also found that, if the monopole carries the Dirac magnetic charge ($n = 1$), then the L_z values are shifted up or down by half a unit. The magnetic flux through a surface bounded by a loop in the x–y plane is half the total flux emanating from the monopole. The orbital angular momentum of the charged particle is half-odd-integral. [Another way to reach this conclusion is to note that the electromagnetic field of the monopole and charged particle has an angular momentum of magnitude $eg = \frac{1}{2}$ (21).]

We conclude that the composite of an integer-spin monopole and an integer-spin charged particle can be a dyon with half-odd-integer spin (72). According to the usual connection between spin and statistics, a composite of two bosons can be a fermion!

Does the usual spin-statistics connection really hold for these objects? One might expect that the interchange of two identical composite dyons could be accomplished by merely interchanging their constituents. However, the interchange of the two dyons should in fact be performed by transporting each dyon covariantly in the gauge potential of the other (71, 73); this procedure corresponds to interchanging the electromagnetic fields of the dyons, as well as their constituents.

It is trivial to perform the covariant interchange of the dyons, if we first choose a gauge in which the vector potential vanishes. The velocity-dependent interaction of two dyons, each with magnetic charge g and electric charge e, can obviously be represented by

$$-e\mathbf{v}\cdot[\mathbf{A}(\mathbf{r})-\mathbf{A}(-\mathbf{r})], \tag{114.}$$

where $\mathbf{r}$ is the separation of the dyons, $\mathbf{v}$ the relative velocity, and $\mathbf{A}$ the monopole vector potential

$$\mathbf{A}(\mathbf{r})\cdot d\mathbf{r} = g(1-\cos\vartheta)\,d\varphi. \tag{115.}$$

The first term in the brackets in Equation 114 is due to the interaction of the electric charge of dyon 1 and the magnetic charge of dyon 2; the second term is due to the interaction of the electric charge of dyon 2 with the magnetic charge of dyon 1. Now, since

$$A_\varphi(\vartheta,\varphi)-A_\varphi(\pi-\vartheta,\varphi+\pi) = 2g = \frac{1}{ie}(\partial_\varphi\Omega)\Omega^{-1}, \tag{116.}$$

where

$$\Omega = \exp(i2eg\varphi), \tag{117.}$$

the gauge interaction, Equation 114, between the dyons can be removed by performing the gauge transformation

$$\Psi(\mathbf{r}) \to \Omega(\mathbf{r})\Psi(\mathbf{r}), \tag{118.}$$

on the two-dyon wave function Ψ.

In this gauge, the dyons may be interchanged naively, by replacing $\mathbf{r}$ by $-\mathbf{r}$. But the gauge transformation shown in Equation 118 has changed the symmetry of the wave function; when the dyons are interchanged, $\varphi \to \varphi + \pi$, Ω changes by the phase

$$\Omega \to \exp(i2\pi eg)\Omega. \tag{119.}$$

This phase is precisely what is needed to restore the usual connection between spin and statistics (73). It is $(-1)^{2l}$, where l is the orbital angular momentum of the composite dyon. It really is possible to obtain a fermion as a composite of two bosons, if $2eg$ is odd.

6.4 *Dyons in Quantum Chromodynamics*

In quantum chromodynamics, as in any gauge theory, a term of the form

$$\mathcal{L}_\vartheta = \frac{\vartheta e^2}{8\pi^2}\mathbf{E}^a\cdot\mathbf{B}^a \tag{120.}$$

can occur in the Lagrange density, where ϑ is an arbitrary parameter. In the real world, ϑ is known to be very close to zero, but it is nonetheless interesting to ask how the strong interactions would behave for different values of ϑ. For one thing, by studying the ϑ-dependence of the theory, we can gain a deeper understanding of quark confinement. For another, there may exist other "super-strong" interactions, not yet known, for which ϑ is not close to zero.

The discovery of the anomalous dyon charge, $Q = -\vartheta/2\pi$, has led to some interesting insights into the ϑ-dependence of quark confinement in quantum chromodynamics (70).

What do magnetic monopoles and dyons have to do with quark confinement? It should be possible to understand quark confinement by considering the dynamics of pure Yang-Mills theory, without quarks. If, in this theory, it is dynamically favored for color-electric flux to collapse to a tube with a characteristic width of order the hadronic size, then confinment is explained. A distantly separated quark-antiquark pair would become connected by a color-electric flux tube carrying constant energy per unit length, and the potential energy of the pair would rise linearly with the separation.

A similar phenomenon occurs in a superconductor. Magnetic flux is expelled from a superconductor (the Meissner effect), and hence collapses to a flux tube. As a result, magnetic monopoles in a superconductor would be "confined."

The Meissner effect arises because of the condensation of electrically charged Cooper pairs in the ground state of a superconductor. It is natural to suggest that quark confinement arises in quantum chromodynamics because of the condensation of color-magnetic monopoles in the vacuum state of Yang-Mills theory (70, 74). The monopole condensate would cause the Yang-Mills vacuum to expel color-electric flux and screen color-magnetic flux.

But what are the monopoles of Yang-Mills theory? We can understand their origin by choosing an appropriate gauge (70). In SU(N) Yang-Mills theory, a gauge transformation can be performed that diagonalizes, for example, the gauge field F_{12} at each point of space-time. This gauge condition generically specifies the gauge transformation up to a diagonal element of SU(N), and hence reduces the theory to an abelian $U(1)^{N-1}$ gauge theory. However, the ambiguity in the gauge transformation is a nondiagonal element of SU(N) at the isolated points in three-dimensional space where two eigenvalues of F_{12} coincide. At these isolated points, the embedding of $U(1)^{N-1}$ in SU(N) is ill defined, and magnetic monopoles appear, carrying $U(1)^{N-1}$ magnetic charges. It is the condensation of these monopoles that presumably accounts for quark confinement.

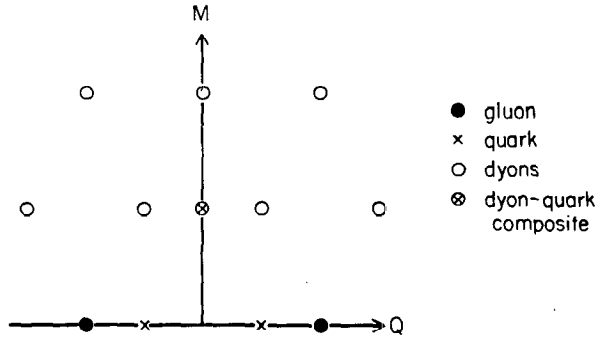

Figure 7 Electric (Q) and magnetic (M) charges in an SU(2) gauge theory with $\vartheta = \pi$.

For simplicity, consider an SU(2) gauge theory, which has only one diagonal U(1) charge and hence only one type of monopole. If $\vartheta = 0$, these monopoles carry no U(1) electric charge. Such monopoles condense, and, as a result, U(1) electric charge is confined, while U(1) magnetic charge is screened.

But, as ϑ varies from 0 to 2π, the monopole acquires an electric charge $Q = -\vartheta/2\pi$. When ϑ reaches 2π, there is again an electrically neutral monopole that condenses (Figure 3). Evidently, the object that condenses must change discontinuously for at least one value of ϑ.

In particular, for $\vartheta = \pi$, there is a monopole with twice the minimal magnetic charge (a bound state of two dyons) that is electrically neutral, and it is plausible that this object condenses, instead of the fundamental dyon (Figure 7). Thus, both elementary dyons and elementary quarks, which have electric charge $Q = \pm\frac{1}{2}$ in SU(2), are confined, but electrically neutral composites of quarks and dyons exist, which are not confined.

On the basis of this picture, one is tempted to conjecture that quarks are unconfined in SU(2) gauge theory at $\vartheta = \pi$. In fact, the liberated quark-dyon composites have orbital angular momentum 1/2, and so are bosons rather than fermions, as discussed in Section 6.3.

Similar phenomena can occur for other values of N. The discovery of the anomalous electric charge of the dyon has led us to expect a highly nontrivial dependence on the parameter ϑ in nonabelian gauge theories.

7. MONOPOLES AND FERMIONS

7.1 *Fractional Fermion Number on Monopoles*

Surprising and qualitatively new phenomena arise when we consider the quantum mechanics of electrically charged fermions interacting with magnetic monopoles.

A first indication of the subtlety of monopole-fermion interactions is obtained if we reconsider the derivation of the dyon charge spectrum (Section 6.2), including the effect of an electron coupled to the electromagnetic field. The derivation appears to go through as before, so we conclude that the allowed electric charges of the dyon are

$$Q = n - (\vartheta/2\pi)M, \tag{121.}$$

where n is an integer.

But it is known that, because of the axial anomaly (75), it is possible to rotate ϑ, the coefficient of the CP-violating $\mathbf{E} \cdot \mathbf{B}$ term in the Lagrangian, to zero, at the cost of introducing a CP-violating phase into the electron mass term (76),

$$\mathscr{L}_{\mathrm{m}} = -me^{i\vartheta}\bar{\psi}_{\mathrm{L}}\psi_{\mathrm{R}} + \mathrm{h.c.} \tag{122.}$$

Therefore, it should be possible to understand the origin of the anomalous dyon charge by carefully inspecting the electron vacuum polarization cloud surrounding the monopole.

Indeed, this is possible. When the Dirac equation is solved for a fermion with a complex mass in the field of a point monopole, it is found that the Dirac sea is distorted for nonzero ϑ, so that the ground state of the monopole-fermion system carries fermion number $(\vartheta/2\pi)M$ and electric charge $Q = -(\vartheta/2\pi)M$ (77). Actually, solving the Dirac equation for an electron in the field of a point monopole involves a further subtlety; it is necessary to impose a boundary condition on the electron wave function at the pole (78, 79). The above remark really holds only if the boundary condition is chosen to be CP conserving, so that the phase ϑ is the only source of CP violation in the problem. Both singular and nonsingular monopoles have the ability to carry a fractional fermion number, in a CP-nonconserving theory (80, 81).

If the mass m of the electron vanishes, its phase is not well defined, and the parameter ϑ must become unobservable. The electric charge of the monopole ground state, which is $-(\vartheta/2\pi)M$ for any nonzero m, must vanish discontinuously in the limit $m \to 0$. This behavior is not so puzzling once we realize that the anomalous charge is carried by the electron vacuum polarization cloud of the monopole. The electric charge radius of the monopole is of order m^{-1}, the electron Compton wavelength, so any observer a finite distance from the monopole center thinks that the electric charge disappears as $m \to 0$ (15, 16, 69).

In fact, all the dyonic excitations of the monopole have an electric charge radius of order m^{-1}, since it is very much preferred energetically to deposit the charge in the electron cloud, instead of on the very tiny monopole core. These excitations are split from the dyon ground state by an energy of order m, and can be regarded as (unstable) dyon-electron composites.

7.2 *Monopole-Fermion Scattering*

The interactions of magnetic monopoles and fermions have another, even more spectacular property. When a monopole and charged fermion collide at low energy, compared to the inverse size M_X of the monopole core, the outcome is strongly dependent on the structure of the core (15, 16). In particular, in a typical grand unified theory there are heavy gauge bosons with masses of order M_X and couplings that violate baryon-number conservation; in such a theory the cross section for baryon-number-changing scattering of a fermion by a monopole at low energy is large, and independent of M_X.

This result seems to violate a cherished principle of quantum field theory, the decoupling principle (82), which asserts that the effects of the very short-distance physics must be suppressed at low energy by a power of the short-distance scale. In this respect, monopole-fermion scattering appears to be a unique phenomenon.

To begin to understand the peculiar features of monopole-fermion scattering, recall that if a particle with electric charge e moves in the field of a point monopole with magnetic charge g, the electromagnetic field carries angular momentum

$$\mathbf{J}_{\text{em}} = -eg\hat{r} \qquad 123.$$

where $\hat{r}$ is the unit vector pointing toward the charged particle from the monopole (21, 83). If the charged particle were to pass through the monopole, this contribution to the angular momentum would change discontinuously. Therefore, conservation of angular momentum forbids the particle to pass through the pole, unless its charge or intrinsic spin can change discontinuously as it does so.

The above remark has a quantum mechanical counterpart, as is seen by solving the Dirac equation for a massless electron in the field of an abelian point monopole with $eg = \frac{1}{2}$. The wave function ψ of the electron is defined (42), as described in Section 4.5, by specifying smooth functions ψ_{U} and ψ_{L} on the upper and lower hemispheres surrounding the monopole, which satisfy a matching condition at the equator of the form shown in Equation 56. Because there is a conserved angular momentum (84)

$$\mathbf{J} = \mathbf{r} \times (-i\nabla - e\mathbf{A}) - \tfrac{1}{2}\hat{r} + \tfrac{1}{2}\boldsymbol{\sigma} \qquad 124.$$

in this problem, the eigenstates of the Hamiltonian can be chosen to be eigenstates of J^2 and J_z. For the states with $J = 0$, the Dirac equation reduces to the radial equation

$$i\gamma_5 \frac{\mathrm{d}}{\mathrm{d}r}\chi(x) = E\chi(r), \qquad 125.$$

where

$$\psi(r, \vartheta, \varphi, t) = \frac{1}{r}\chi(r)\eta_0(\vartheta, \varphi) \exp(-iEt), \qquad 126.$$

and η_0 is the $J = 0$ "monopole harmonic" (42, 78).

The solutions to Equation 125 have an odd property; the positive-helicity ($\gamma_5 = -1$) solution is purely an outgoing wave, and the negative-helicity ($\gamma_5 = +1$) solution is purely an incoming wave. (For a positron, the helicities of the solutions are reversed.) Both solutions are singular at the origin, the location of the pole, and the Dirac equation provides no criterion for matching up the incoming and outgoing solutions. The Hamiltonian defined by the Dirac equation is therefore not self-adjoint; probability is not conserved unless the Hamiltonian is supplemented by a boundary condition at the origin (the location of the pole) relating the incoming and outgoing waves (79).

This trouble can be traced back to the unusual term $\mathbf{J}_{em} = -\frac{1}{2}\hat{r}$ in the expression for the angular momentum. An incoming (outgoing) electron must have negative (positive) helicity to be in a state with $\mathbf{J} = \mathbf{J}_{em} + \boldsymbol{\sigma} = 0$. For a positron, $\mathbf{J}_{em}$ has the opposite sign, and the helicities are reversed.

The boundary condition at the origin determines the fate of a left-handed electron, which scatters from a monopole in the $J = 0$ partial wave. But there are only two options; it becomes either a right-handed electron or a left-handed positron, because these are the only available outgoing modes with $J = 0$. The boundary condition must therefore either violate chirality (which is otherwise a good symmetry of the Hamiltonian) or require the monopole to absorb electric charge. If the charge-conserving boundary condition is chosen, then the chirality-changing $J = 0$ cross section will saturate the unitarity limit (15, 16).

The need for a boundary condition to determine the final state of an electron scattering from a point monopole is the crucial feature of monopole-fermion scattering that results in the violation of the decoupling principle. The decoupling principle leads one to expect that the amplitude for monopole-fermion scattering at energies much less than the inverse size of the monopole core does not depend on the structure of the core, except for power corrections that vanish as the size of the core goes to zero. Up to power corrections, the amplitude should be calculable in a low-energy "effective theory" in which the core is regarded as pointlike and its properties need not be specified. This expectation fails because monopole-fermion scattering is inherently ambiguous when the monopole is pointlike. Information about the core of the monopole survives in the low-energy effective theory as a boundary condition needed to specify the outcome of a scattering event; a low-energy fermion with $J = 0$ can penetrate to the core

of the monopole, and be strongly influenced by its structure. In particular, the boundary condition may violate a symmetry (like baryon number) that would otherwise be a good symmetry of the low-energy effective theory.

We now see that the analysis of the scattering of a low-energy fermion by a nonsingular monopole with nonvanishing core size can be divided into two parts. First, we decide what boundary conditions must be imposed as the limit of zero core size is taken. Then the interaction of a point monopole with fermions satisfying the appropriate boundary conditions is studied. The second step is highly nontrivial. Fermion pair creation effects, which are responsible for smearing out the electric charge of the dyonic excitations of the monopole over a region with radius of order the fermion Compton wavelength, must be taken into account as fully as possible. But Rubakov (15) and Callan (16) suggested that, since only $J = 0$ fermions can penetrate to the core of the monopole, the problem can be reasonably approximated by an effective $(1+1)$-dimensional quantum field theory describing the $J = 0$ partial wave, in which the spatial coordinate is the radial coordinate r. The qualitative features of this $(1+1)$-dimensional theory are most easily glimpsed if it is converted into an equivalent "bosonized" theory (85) in which the fermions are represented by solitons. This soliton picture of monopole-fermion scattering is especially convenient when we try to understand the effects of fermion masses.

Returning to the problem of finding the appropriate boundary conditions satisfied by the fermions, let us consider, for concreteness, the case of the SU(5) grand unified model with a single generation of fermions. The magnetic charge of the SU(5) monopole is a linear combination of ordinary magnetic charge and color magnetic charge. At a distance from the monopole center much less than the characteristic hadronic size 10^{-13} cm and much greater than the radius of the core, the only fermions that interact with the monopole are those carrying Q', the corresponding combination of electric charge and color electric charge, given, in an appropriate gauge, by Equation 86. The right-handed quarks and leptons carrying nonzero Q' are

$$\begin{aligned} Q' &= +1 : e_R^+ \bar{d}_{3R} u_{1R} u_{2R} \text{ (incoming)} \\ Q' &= -1 : d_{3R} e_R^- \bar{u}_{2R} \bar{u}_{1R} \text{ (outgoing)}, \end{aligned} \qquad 127.$$

where e denotes the electron, u and d denote up and down quarks, and 1, 2, 3 are color indices. The behavior of these fermions in the field of the SU(5) monopole is identical to the behavior of an electron or positron in the field of an ordinary Dirac monopole. If fermion masses are ignored, then the right-handed (left-handed) fermions with $Q' = +1$ in the $J = 0$ partial wave are incoming (outgoing) only; for $Q' = -1$ the helicities are reversed. The new feature is that there are now four Dirac fermions interacting with the

monopole, and the boundary condition at the origin causes these fermions to mix in a manner determined by the structure of the core of the monopole.

One can attempt to determine the boundary condition by solving the Dirac equation in the field of the nonsingular SU(5) monopole with finite core radius (16, 80, 86). The result is that the helicity of the incoming fermion is preserved; incoming and outgoing states are matched up as in Equations 127. We see that two units of Q' are transferred to the monopole, exciting its dyon degree of freedom.

But if we now investigate the consequences of this boundary condition, taking proper account of pair creation effects, we realize that the picture in which the incoming fermion falls to the core and deposits charge there, suggested by the solution to the Dirac equation, is not very accurate. An enormous Coulomb barrier prevents charge from being deposited on the core. It is energetically favored for the charge to be spread out over a region with a radius of order a fermion Compton wavelength. As a result, our original procedure for finding the correct boundary condition is called into question. It seems that a more appropriate boundary condition is one forbidding charge to accumulate at the origin (87).

Fortunately and remarkably, in the case of the SU(5) monopole we can obtain quite nontrivial information about the scattering process by merely demanding that none of the charges coupling to massless gauge bosons accumulate on the core (88). This constraint is especially powerful because W and Z bosons must be regarded as effectively massless at distances from the center of the monopole much less than M_Z^{-1}. Since left-handed and right-handed fermions with the same electric charge have different values of the charge $(T_3)_{\text{weak}}$, which couples to the W_3 boson, simple chirality-violating processes such as

$$e_L^- + M \rightarrow e_R^- + M \qquad 128.$$

are forbidden for massless fermions. If, for example, two $J = 0$ u quarks scatter from the monopole, there is only one possible final state of two $J = 0$ fermions; the allowed process is

$$u_{1R}u_{2R} + M \rightarrow \bar{d}_{3L}e_L^+ + M. \qquad 129.$$

The most general final fermion state consistent with conservation of all gauged charges is $\bar{d}_{3L}e^+$ accompanied by an indefinite number of the pairs $\bar{u}_R u_L$, $\bar{d}_L d_R$, and $e_L^+ e_R^-$, which carry no favor quantum numbers. So baryon-number nonconservation is forced on us, if we ignore the masses of the fermions, and the cross section for baryon-number-changing scattering is not suppressed by the small size of the monopole core.

Presumably, then, SU(5) monopoles are able to catalyze the decay of a nucleon at a characteristic strong-interaction rate (15, 16). The only

property of the SU(5) model that we needed to invoke was the existence of a monopole that couples to the charge Q', at distances from the monopole center less than M_Z^{-1}, where the weak-interaction symmetries are effectively restored. In any realistic model containing an $[SU(3) \times SU(2) \times U(1)]/Z_6$ monopole (see Section 5.3) and light fermions with the standard charge assignments, the amplitude for the process in Equation 129 will be unsuppressed by the small size of the monopole core.

This process (Equation 129), with two fermions in the initial state, must occur in two steps. It is natural to inquire about the intermediate state produced when u_{1R} scatters from the monopole. What one finds (89–91) is rather subtle and bizarre; the intermediate state consists of four "semitons," each with fermion number $\frac{1}{2}$. The reaction

$$u_{1R} + M \rightarrow \tfrac{1}{2}(u_{1L}\bar{u}_{2R}\bar{d}_{3L}e_L^+) + M, \qquad 130.$$

changes baryon number and lepton number by $-\frac{1}{2}$ unit.

The semitons are destabilized by fermion mass terms or the effects of the strong color interaction. At a distance from the monopole center where these effects become important, the intermediate state in Equation 130 evolves into a final state with baryon and lepton number differing by an integer (possibly zero) from that of the initial state. One possibility is that the semitons in Equation 130 evolve into u_{1L}; chirality-violating processes like that in Equation 128 are allowed if the fermions have masses.

The evolution of semitons into "final-state" quarks and leptons is not yet understood in quantitative detail. But it is reasonable to expect that the semiton intermediate state can evolve with a probability of order one into a final state with a baryon number different from the initial state (89–91). It is also expected that adding more generations of fermions will have no qualitative effect on the baryon-number-changing processes. The main new feature in the many-generation case is that the boundary conditions and hence the scattering amplitudes depend on generalized Cabibbo-like mixing angles (92).

The above considerations strongly suggest that the baryon-number-changing cross section for a quark of energy E scattering from a monopole is of order E^{-2}, if E^{-1} is much greater than the radius of the monopole core, and much less than both the Compton wavelength of the quark and the size of a hadron. But we are really interested in the cross section for *nucleon* decay catalyzed by the monopole. There are actually two questions of experimental interest. One is, what is the cross section for capture of nuclei by the monopole? It is probably large (93), and the capture rate might conceivably control the catalysis rate in terrestrial experiments. The other is, what is the cross section for the catalysis process itself? In spite of some ambitious attempts (94), techniques do not yet exist for doing detailed

quantitative calculations of the catalysis cross section. The best guess is that it is roughly geometrical, $\sigma\beta \sim 10^{-27}$ cm^2, and that the most likely final state is a positron accompanied by a pion.

8. MONOPOLES IN COSMOLOGY AND ASTROPHYSICS

8.1 *Monopoles in the Very Early Universe*

We have seen that the existence of magnetic monopoles is a very general consequence of the unification of the fundamental interactions. But to say that monopoles must exist in grand unified theories is not necessarily to say that we have a reasonable chance of observing one. If the monopole mass is really as large as 10^{16} GeV, then there is no hope of producing monopoles in accelerator experiments in the foreseeable future.

However, it is likely that the universe was once extremely hot, so hot that processes occurred that were sufficiently energetic to produce monopoles. If there are any monopoles around today, they are presumably relics of the very brief, very energetic epoch immediately following the "big bang." So it is evidently interesting to consider how many monopoles might have been produced in the very early universe (8, 9).

As the universe cooled, it is expected to have undergone a phase transition at a critical temperature T_c of order the unification mass scale M_X. When the temperature T was above T_c, the full, grand unified gauge symmetry was restored (95); the scalar field Φ, which acts as an order parameter for the breakdown of the gauge symmetry, had a vanishing expectation value. But monopoles can exist only when the gauge symmetry is spontaneously broken, so no monopoles were present when T was above T_c. When T fell below T_c, the expectation value of Φ turned on, and monopole production became possible.

Because monopoles, unlike the other superheavy particles in grand unified theories, are stable, the density of monopoles per comoving volume established in the phase transition at T of order T_c could subsequently be reduced only by annihilation of monopole-antimonopole pairs. As the universe rapidly expanded, monopoles and antimonopoles had an increasingly more difficult time finding each other, and an appreciable density of monopoles per comoving volume might have persisted.

Thus, the problem of estimating the monopole abundance may be separated into two parts. We must estimate the initial density of monopoles established during the phase transition, and we must determine to what extent the monopole density was subsequently reduced by pair annihilation.

Let us first consider the production of monopoles during the phase transition. The detailed mechanism by which monopoles were produced

depends on the nature of the phase transition; in particular, on whether it was a second-order (or weakly first-order) transition, in which large fluctuations occurred, or a strongly first-order transition, in which supercooling occurred. In either case, there is no reason to believe that the monopole abundance was ever in thermal equilibrium.

In the case of a second-order (or weakly first-order) phase transition, the scalar field Φ underwent large random fluctuations when T was near T_c. As the universe expanded and cooled, the scalar field was rapidly quenched, and a large density of topological defects became frozen in; these defects are the monopoles and antimonopoles. The quenching process may be described in the following way (7): At the time when the monopoles are being produced, the scalar field Φ is uncorrelated over distances larger than some characteristic correlation length ξ. We may thus regard Φ as having a domain structure, with ξ the characteristic size of a domain. At the intersection point of several domains, each with a randomly oriented scalar field, there is some probability p that the scalar field orientation is topologically nontrivial; if so, a monopole or antimonopole must form at the intersection point. The probability p depends on the detailed structure of the monopole, but it is not very much less than one. According to this picture, the density of monopoles n established in the phase transition is

$$(n)_{\text{initial}} \sim p\xi^{-3}. \qquad 131.$$

This argument sounds suspicious, because it relies on the notion of a scalar field domain structure, even though it is always possible to make a uniform scalar field look wiggly by performing a gauge transformation. However, there is no fundamental difficulty. We can fix the gauge in an appropriate way so that the idea of a scalar field domain makes sense, and we have reached a conclusion about the density of topological defects, which is a gauge-invariant quantity.

In a second-order (or weakly first-order) phase transition, the correlation length ξ becomes large as T approaches T_c, but in the early universe causality places a limit on how much ξ can grow (96). The scalar field Φ must remain uncorrelated over distances exceeding the horizon length d_H, the largest distance any signal could have traveled since the initial singularity; thus $\xi < d_H$. In terms of the temperature T, d_H may be written as (97)

$$\xi < d_H \sim Cm_p/T^2, \qquad 132.$$

where $m_p \sim 10^{19}$ GeV is the Planck mass, which determines the expansion rate of the universe, and $C = (0.60)N^{-1/2}$, where N is the effective number of massless spin degrees of freedom in thermal equilibrium at temperature T. (In a minimal grand unified theory, $C \sim 1/20$.) Combining Equations 131

and 132, we conclude that the initial value of the dimensionless ratio n/T^3 is bounded by

$$(n/T^3)_{\text{initial}} \gtrsim p(T_c/Cm_p)^3. \quad 133.$$

In a typical grand unified theory, with $T_c \sim 10^{15}$ GeV, $Cm_p \sim 10^{18}$ GeV, and $p \sim 1/10$, we obtain $(n/T^3)_{\text{initial}} \gtrsim 10^{-10}$.

In the case of a strongly first-order phase transition, supercooling occurs. The phase with unbroken grand unified gauge symmetry becomes thermodynamically unstable when $T < T_c$, but nonetheless persists for a while, until bubbles of the stable broken-symmetry phase eventually begin to nucleate. These bubbles expand, collide, and coalesce, filling the universe with the stable phase (98). Inside each bubble, the scalar field Φ is quite homogeneous, so that each bubble contains a negligible number of monopoles. But when the expanding bubbles collide, monopoles can be produced.

Although it is not easy to calculate in detail the initial density of monopoles produced by bubble collisions, we can obtain a lower bound on the monopole density by invoking an argument similar to the one applied above to the case of a second-order transition. Now each bubble can be regarded as a scalar field domain, and the density of monopoles produced by the collisions must exceed the probability factor p times the density of bubbles at the time they collide. Since bubbles cannot expand faster than the speed of light, each bubble must be smaller in radius than the horizon site d_H, and the density of bubbles must be greater than d_H^{-3}. We again conclude, therefore, that $n > pd_H^{-3}$, and the bound in Equation 133 still applies, except that T_c is replaced by a temperature at which bubble nucleation becomes probable.

Regardless of the nature of the phase transition, reasonably copious production of monopoles seems to be inevitable. Moreover, the monopole abundance cannot be significantly reduced by pair annihilation (8, 9, 99). The annihilation rate is determined by the monopole-antimonopole capture rate; once a bound pair forms, it quickly cascades down emitting many photons and gluons, and finally annihilates into a burst of superheavy scalar particles and X-bosons. But the capture process is relatively inefficient because the monopoles are so heavy; it fails to keep pace with the expansion of the universe if the monopole abundance n is smaller than (9)

$$(n/T^3) \sim 10^{-9}(m/10^{16}\ \text{GeV}), \quad 134.$$

where m is the mass of the monopole. (The quantity n/T^3 is convenient to consider, because it remains constant if the expansion of the universe is adiabatic, and no monopoles are created or destroyed.) Once the monopole abundance is comparable to Equation 134 or smaller, monopole-

antimonopole annihilation cannot further reduce the monopole density per comoving volume.

Using the standard estimate $m \sim 10^{16}$ GeV, we see that, if the smallest possible initial monopole abundance consistent with the bound (Equation 133), $(n/T^3) \sim 10^{-10}$, is established in the phase transition, this abundance is not further reduced by annihilation at all. The only way to reduce n/T^3 further is through nonadiabatic effects that increase the entropy density, but such effects cannot dilute the monopole abundance by many orders of magnitude without at the same time diluting the baryon-number density of the universe. Neglecting generation of entropy, we conclude that the density of magnetic monopoles today is $n \sim 10^{-10}T^3$, which is comparable to the density of baryons. This conclusion is clearly absurd, if the mass of a monopole exceeds the mass of a baryon by a factor of order 10^{16}.

We have uncovered the "monopole problem," an apparently serious conflict between grand unified theories and standard big-bang cosmology. Various attempts have been made to resolve this conflict. By far the most appealing resolution of the monopole problem is offered by the inflationary universe scenario (10–12, 139).

In this scenario, a positive effective cosmological constant causes the universe to "inflate" exponentially as a function of time, after the appearance of bubbles or fluctuation regions in which the scalar field Φ is nonzero. Some monopoles are produced when bubbles or fluctuation regions first form, but they are subsequently "inflated away"; in the course of the exponential expansion, the monopole abundance is reduced to a negligible value. Eventually, after many e-foldings of expansion, the cosmological constant that drove the inflation is rapidly converted to radiation, and the universe "reheats." Its subsequent evolution is well described by the standard cosmological model.

The inflationary universe scenario is more appealing than other possible solutions to the cosmological monopole problem (22, 23, 25) because inflation solves other cosmological problems as well. It explains why the universe is nearly homogeneous and isotropic, and why the mass density of the universe today is close to the critical density required to cause it to recollapse (10). It may also explain the origin of the primordial density fluctuations that led to galaxy formation (100). As presently formulated, the inflationary-scenario is not free of flaws, but it seems likely to be essentially correct. So it is plausible that inflation is the mechanism by which the cosmological abundance of monopoles became suppressed.

Not only can inflation reduce the monopole abundance to an acceptably small level, it can easily reduce the abundance to so low a level that a monopole will never be seen (23, 101, 102). Fortunately, this last statement is not a firm prediction. Monopoles may be produced either during

inflation (103) or during (104) and after (23, 101, 102) the reheating of the universe. Until the details of the scenario are better known, it will not be possible to predict accurately the monopole abundance within the context of the inflationary universe scenario.

One suggestion for suppressing the monopole abundance within the context of the standard cosmological scenario is worthy of mention. It is possible that the universe entered a superconducting phase as it cooled (105). A superconductor tries to expel magnetic flux, so monopole-antimonopole pairs would have become connected by flux tubes and annihilated rapidly (106). As the universe cooled further, it might have eventually returned to a normal nonsuperconducting phase, but only after the monopole abundance had been significantly reduced.

An interesting feature of this superconductor scenario is that a potentially interesting number of monopoles could have survived until the universe re-entered the normal phase. Although the positions of monopoles and antimonopoles are strongly correlated when monopoles are first produced, the flux tubes do not pair up monopoles and antimonopoles perfectly. Some monopoles, unable to decide which antimonopole to pair up with, may get left behind. It thus seems possible that the superconductor scenario predicts a detectable abundance of monopoles (107).

The monopole problem has exerted a healthy influence on the development of cosmology during the past few years. And if the monopole abundance is ever measured, it will severely constrain our speculations about the very early universe. But, for now, cosmology does not offer much guidance to the prospective monopole hunter; there is no definite cosmological prediction for the monopole abundance.

8.2 *Astrophysical Constraints on the Monopole Flux*

Although cosmological considerations provide no definite prediction for the monopole abundance, both the inflationary scenario and the superconductor scenario offer the possibility that the monopole abundance is both small enough to be acceptable and large enough to be detectable. Theoretical cosmology should not discourage an experimenter from looking for monopoles.

People have been looking for magnetic monopoles for a long time. But traditional monopole searches do not place significant constraints on superheavy monopoles with mass m of order 10^{16} GeV. The traditional searches have relied on the strong ionization power of a relativistic monopole (108, 109), or have sought monopoles trapped in the Earth's crust (110, 111). But a superheavy monopole would be expected to be slowly moving and very penetrating; it need not ionize heavily or stop in the earth (9, 112, 113).

How slowly moving? The monopole can be accelerated by either gravitational fields or magnetic fields; which effect is more important depends on the mass of the monopole. From gravitational fields alone, the monopole would acquire a typical galactic infall velocity of order $10^{-3}c$, regardless of its mass. To determine the effect of the magnetic field in our galaxy, recall that the field has a strength B of order 3×10^{-6} gauss and a coherence length L of order 10^{21} cm (114). A monopole with the Dirac charge g_D crossing one coherence length is accelerated to

$$v = \left(\frac{2g_D BL}{m}\right)^{1/2} \sim 10^{-3}c\left(\frac{10^{17}\ \text{GeV}}{m}\right)^{1/2}. \qquad 135.$$

This magnetic acceleration is therefore more important than the gravitational acceleration for $m \leq 10^{17}$ GeV. The monopole does not attain a relativistic velocity for $m > 10^{11}$ GeV.

How penetrating? The stopping power of slowly moving magnetic monopoles remains a rather controversial subject, about which a little more is said in Section 9. But the energy loss in rock of a monopole with $v/c \lesssim 10^{-2}$ surely does not much exceed (115)

$$\frac{dE}{dx} \sim 100\left(\frac{v}{c}\right)\ \text{GeV cm}^{-1}. \qquad 136.$$

Thus, the range in rock of a monopole with $m \sim 10^{16}$ GeV is larger than 10^{11} cm; the monopole passes through the Earth without slowing down.

Although superheavy monopoles are not easily stopped or detected, astrophysical arguments can be used to place severe limits on the flux of magnetic monopoles in cosmic rays. These limits offer valuable guidance to the prospective monopole hunter.

One stringent limit [the "Parker limit" (13)] on the monopole flux is obtained by noting that, because the magnetic field of our galaxy accelerates monopoles, the energy density $U = B^2/8\pi$ stored in the field is dissipated at the rate $dU/dt \sim \langle gnv \cdot \mathbf{B} \rangle$, where n is the monopole density. By demanding that the field energy is not substantially depleted in a time τ, of order 10^8 years, required to regenerate the field, we obtain the bound

$$F = \frac{nv}{4\pi} \lesssim \frac{B}{32\pi^2 g\tau} \sim 10^{-16}\ \text{cm}^{-2}\ \text{s}^{-1}\ \text{sr}^{-1}\left(\frac{B}{3 \times 10^{-6}\text{G}}\right)\left(\frac{10^8\ \text{y}}{\tau}\right). \qquad 137.$$

A nice feature of this flux limit is that it appears to be independent of the mass m of the monopole.

However, it is implicitly assumed in the derivation of Equation 137 that gravitational effects on the trajectory of the monopole are negligible, and we have already argued that this is not so for $m \gtrsim 10^{17}$ GeV. If a monopole

enters a coherent domain of the galactic magnetic field with incident energy $\frac{1}{2}mv^2 > gBL$, then the energy ΔE it extracts from the domain is a second-order effect, due to the deflection of the monopole trajectory as it crosses the domain; on the average it is

$$\Delta E \sim (gBL)^2/\tfrac{1}{2}mv^2. \qquad 138.$$

Therefore, the rate of dissipation of magnetic field energy scales like $1/m$ for $m \gtrsim 10^{17}$ GeV, and the flux limit becomes (14)

$$F \lesssim 10^{-16}\ \text{cm}^{-2}\ \text{s}^{-1}\ \text{sr}^{-1}(m/10^{17}\ \text{GeV}),\ m \gtrsim 10^{17}\ \text{GeV}. \qquad 139.$$

For $m \gtrsim 10^{20}$ GeV, a more stringent limit than Equation 139 can be obtained, which is based solely on the enormous mass of the monopole and has nothing to do with its magnetic charge (9, 14, 112, 113). The total number of monopoles in our galaxy must not exceed the mass of the galaxy. By demanding that the mass of a spherical monopole galactic halo with radius of order 30 Kpc not exceed 10^{12} solar masses, and taking the typical monopole velocity to be of order $10^{-3}c$, we obtain the flux limit (14)

$$F \lesssim 10^{-13}\ \text{cm}^{-2}\ \text{s}^{-1}\ \text{sr}^{-1}(10^{20}\ \text{GeV}/m). \qquad 140.$$

Since this limit on the flux crosses the one in Equation 139 for $m \sim 10^{20}$ GeV, we have also the mass-independent bound

$$F \lesssim 10^{-13}\ \text{cm}^{-2}\ \text{s}^{-1}\ \text{sr}^{-1}. \qquad 141.$$

(See Figure 8.)

While monopoles are undoubtedly rare, they may play an important role in the dynamics of galaxies. The above reasoning does not exclude the possibility that monopoles make up the dark matter of galactic halos, for

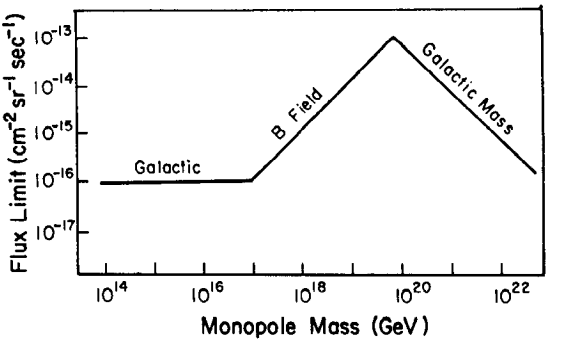

Figure 8 Astrophysical limits on the monopole flux as a function of monopole mass. "Galactic B field" labels the limit based on the energetics of the galactic magnetic field. "Galactic mass" labels the limit based on the total mass of the galaxy.

$m \gtrsim 10^{20}$ GeV. A monopole mass much larger than 10^{20} GeV seems unlikely, but a mass of order 10^{20} GeV is not that implausible; this is the typical sort of monopole mass expected in a Kaluza-Klein theory. And a monopole flux close to the limit in Equation 141 ought to be detectable; it corresponds to about one event per year in a few m^2 of detector.

The derivation of the Parker limit, Equation 139, is subject to one potentially serious criticism—the feedback of the accelerated monopoles on the galactic magnetic field has been ignored. If the monopole abundance were sufficiently *large*, then the period of magnetic plasma oscillations would be less than the time required for a monopole to cross a coherent domain of the magnetic field. The galactic magnetic field might then drive weakly damped plasma oscillations, rather than irreversible magnetic currents, and the Parker limit might be evaded by a significant margin if coherent oscillations could be maintained over many cycles (14, 116, 117). It seems likely, however, that small-scale gravitational and magnetic inhomogeneities, which are inevitably present, would destroy coherence and damp such oscillations rapidly.

In the hope of evading the Parker limit, it has also been suggested that the local flux of magnetic monopoles in the solar system greatly exceeds the ambient flux in the galaxy (118). This suggestion seems implausible on purely kinematic grounds (119).

Even more powerful limits on the monopole flux can be obtained by considering the astrophysical implications of the catalysis by monopoles of nucleon decay. The most interesting implication concerns the effects of monopoles in neutron stars (120, 121). A monopole striking a neutron star gets captured inside the star, if $m \lesssim 10^{22}$ GeV. Then, surrounded by matter at nuclear density, it catalyzes nucleon decay at a furious rate. A modest number of monopoles in a neutron star would cause the star to heat up, and emit a substantial flux of ultraviolet or x-ray photons. From observational limits on the ultraviolet and x-ray luminosity of old pulsars, it is therefore possible to derive a bound on the monopole flux F; conservatively, this bound is (120–122)

$$F \lesssim 10^{-22}\ \mathrm{cm}^{-2}\ \mathrm{s}^{-1}\ \mathrm{sr}^{-1}(\sigma\beta/10^{-27}\ \mathrm{cm}^2)^{-1}, \qquad 142.$$

where σ is the cross section for catalysis of nucleon decay by a monopole, and $\beta \sim 0.3$ is the relative velocity of the nucleon and monopole.

We infer that, if catalysis really proceeds at a strong-interaction rate, then the monopole flux must be smaller than Parker's limit (Equation 137) by at least six orders of magnitude. Monopoles must be so rare that there is little hope of observing one directly. The best way to find evidence for their existence would be by observing their effect on the luminosity distribution of neutron stars.

Of course, the limit (Equation 142) is nullified if monopoles do not catalyze nucleon decay, or do so at an insignificant rate. We expect the catalysis phenomenon to occur only if the new interactions associated with the monopole core fail to conserve baryon number, and it is surely possible to construct models for which this is not the case. An example is the SO(10) model of Section 5.3. The light monopoles associated with the $H_1 \to H_2$ breakdown in that model do not catalyze nucleon decay (59), although the heavy monopoles do, and it is easy to imagine a cosmological scenario in which these light monopoles are much more abundant than the heavy monopoles. (Note that, as is typical of monopoles that do not catalyze nucleon decay, the light monopoles carry twice the Dirac charge.) For monopoles that do not catalyze nucleon decay, the best flux limits we have are Equations 137, 139, and 140.

Still, for the experimenter who dreams of catching a monopole in the act of catalyzing nucleon decay, the bound in Equation 142 is very discouraging. An even more stringent limit can be obtained if capture of monopoles by the main sequence progenitor of the neutron star is taken into account (122). This stronger limit is more mass sensitive, however; Planck-mass monopoles, for example, would rarely be captured by main sequence stars.

Once captured by a neutron star, a monopole must be accelerated to a velocity of order c to escape. In the hope of evading the bound (Equation 142), it is worthwhile to consider whether there is any possible mechanism by which monopoles can be efficiently ejected from neutron stars at relativistic velocities (123).

It is generally believed that the core of a neutron star is a type II superconductor in which Cooper pairs of protons have condensed (124). Because the superconducting core expels magnetic flux, monopoles entering the star will eventually come to rest on the surface of the core. Typically, many magnetic flux tubes will have been trapped in the core when it went superconducting, and a monopole floating on the surface of the core will occasionally encounter the opening of a tube and drop in, penetrating the core (125).

It is conceivable that, deep within the core, there is an inner core in which charged pions condense. This pion condensate is also a type II superconductor, but its flux tubes would carry considerably higher energy per unit length than the flux tubes in the proton superconductor. Also, the flux quantum in the pion condensate would be the Dirac magnetic charge carried by the monopole, rather than half the Dirac charge as in the proton superconductor. Thus, two flux tubes in the proton superconductor would coalesce at the surface of the pion condensate, and a monopole drifting down one of them would be rapidly accelerated upon entering the pion condensate, the sizable magnetic field energy stored in the tube being

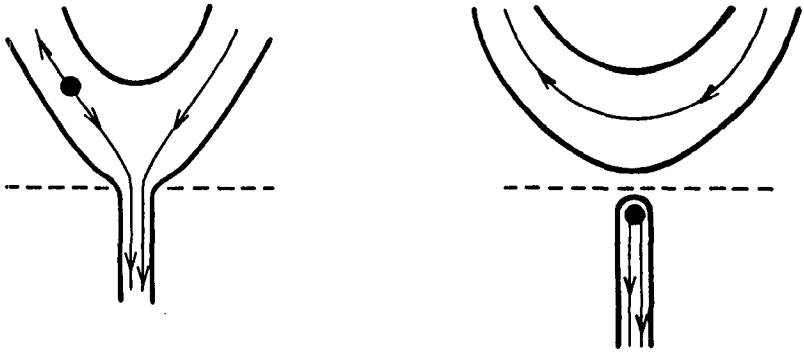

Figure 9 Magnetic monopole in a flux tube near the boundary between a proton pair condensate and a charged pion condensate. In the proton pair condensate (*left*), the magnetic flux in the tube reverses direction at the monopole, and there is no magnetic force on the monopole. In the pion condensate (*right*), the flux tube terminates on the monopole, and the monopole accelerates rapidly.

efficiently converted to monopole kinetic energy (see Figure 9). The monopole could be accelerated to a relativistic velocity, and ejected from the star! Hence, if we take the fullest advantage of our ignorance concerning the interiors of neutron stars, it is possible that the discouraging bound of Equation 142 can be evaded (123).

Even if we throw out the neutron star arguments, it may still be possible to obtain a stringent bound on the monopole flux by considering the effect of catalysis in white dwarfs; the interiors of white dwarfs are less exotic and better understood than those of neutron stars. If it is assumed that all monopoles that strike a white dwarf are captured, then, from observational limits on the luminosities of white dwarfs, we obtain a bound on the monopole flux (126) that is weaker by only about three orders of magnitude than Equation 142. Sufficiently heavy monopoles ($m \gtrsim 10^{20}$ GeV) are not captured by white dwarfs, but such heavy monopoles probably could not be ejected from a neutron star either.

All in all, the uncertaintities in the astrophysical arguments are such that it is barely possible to believe that there is an observable flux of monopoles that catalyze nucleon decay at a strong-interaction rate. Terrestrial searches for such monopoles are not completely pointless.

9. DETECTION OF MONOPOLES

9.1 *Induction Detectors*

If there are magnetic monopoles in the universe, they are surely rare, and they are probably slowly moving. Attempting to detect these monopoles is

a challenging and risky experimental enterprise. But the potential rewards are so great that considerable risk is justified.

Possible techniques for detecting monopoles are briefly described here, but no attempt is made to give a complete review of recent experiments. More comprehensive reviews can be found in (127–129).

In principle, a closed loop of superconducting wire is an ideal monopole detector, because it gives an unambiguous signal whenever a monopole passes through the loop, however slowly (111, 130). To determine the effect on the loop of a monopole passing through, we may use the integrated Maxwell equation

$$\oint_\Gamma \mathbf{E}\cdot d\mathbf{r} = -\frac{d\Phi}{dt} - \frac{dQ_m}{dt}, \qquad 143.$$

which has been suitably modified to take into account the magnetic monopole current. Here Φ is the magnetic flux through a surface S_Γ bounded by the closed path Γ, and dQ_m/dt is the monopole current through the surface S_Γ. Applying Equation 143 to a path Γ entirely contained in the superconducting wire, where $\mathbf{E} = 0$, and integrating over time, we find that the change $\Delta\Phi$ in the magnetic flux linking the loop and the total magnetic charge ΔQ_m that passes through the loop are related by

$$\Delta\Phi = -\Delta Q_m. \qquad 144.$$

In particular, if a magnetic monopole with the Dirac charge g_D passes through the loop, the flux changes by two quantized flux units; the factor of two arises because the electric charge of a Cooper pair is $2e$.

The result (Equation 144) is actually obvious, because the magnetic field cannot penetrate the superconducting wire. The magnetic field lines emanating from the monopole are therefore swept back as the monopole approaches the wire, and break off, forming closed loops around the wire, as indicated in Figure 10 (131). (The field lines are allowed to break and rejoin where the magnetic field vanishes.)

We see that a monopole passing through the loop causes a sudden change in the magnetic flux linking the loop, and a corresponding shift in the dc current level in the wire, with a rise time of order the radius of the loop divided by the velocity of the monopole. For a sufficiently small loop (less than 10 cm in diameter), the shift can be easily detected by a SQUID (superconducting quantum interference device) magnetometer, provided that the loop is adequately shielded from other fluctuating magnetic disturbances.

The sizes of superconducting loop detectors are currently limited by the magnetic shielding requirement and by signal-to-noise problems, but monopole searches have been conducted with such detectors, and an

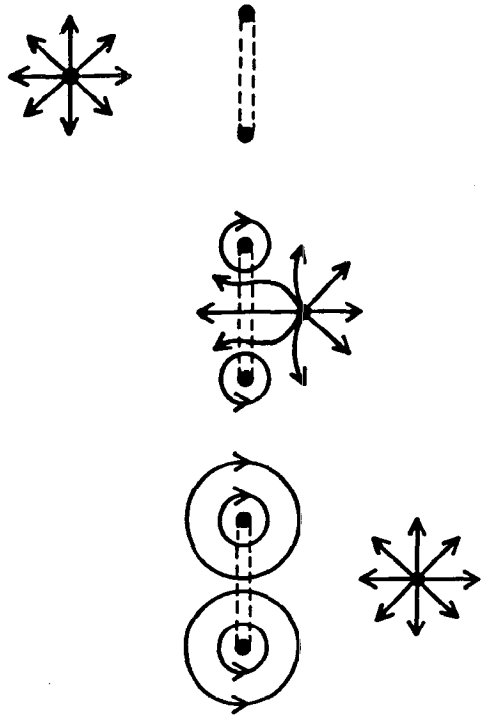

Figure 10 Bending and breaking of magnetic field lines, as a monopole passes through a loop of superconducting wire.

experimental flux limit (132) $F \lesssim 2 \times 10^{-11}$ cm^{-2} s^{-1} sr^{-1} has been obtained at the 90% confidence level. One candidate event has been seen (133), with a magnitude consistent with the Dirac magnetic charge. This event is the basis of the statement in the first paragraph of the introduction. Confirmation is still awaited.

The interpretation of this event as a magnetic monopole is not easily reconciled with the theoretical flux limits of Section 8.2. But no other completely satisfactory interpretation has yet been suggested.

9.2 *Ionization Detectors*

The superconducting induction detector has the significant advantage of being sensitive to monopoles of arbitrarily low velocity. But it will not be easy to construct a detector based on these principles big enough to challenge the theoretical flux limits of Section 8.2.

Larger detectors can be designed to detect the ionization loss of a monopole passing through matter. A very slowly moving ($\beta \sim 10^{-3}$)

particle with significant energy loss would be an unmistakable signal, but a negative result of a search for such events is not so easy to interpret, because the energy loss of slow monopoles is not that well understood.

The energy loss of a fast (electrically or magnetically) charged particle in matter is easily calculated, because electron encounters may be treated in the impulse approximation. But the energy loss of a slow particle depends on the details of atomic and molecular physics. One is inclined to say that the response of an atom or molecule to a very slow ($\beta \lesssim 10^{-3}$) monopole passing nearby is similar to its reponse to a magnetic field that adiabatically turns on and off; therefore, it is unlikely to become excited. But this conclusion is not necessarily correct, because the very strong magnetic field of the monopole greatly distorts the energy levels of the atom or molecule. If the ground state and an excited state closely approach each other, the adiabatic approximation may fail badly. To decide whether this occurs, detailed knowledge of the level structure in the inhomogeneous magnetic field of the monopole is needed.

As an illustration, consider a problem simple enough to be amenable to a sound theoretical analysis; the energy loss of a monopole in atomic hydrogen (134). First, imagine that a very slow monopole is incident on the nucleus of a hydrogen atom with zero impact parameter, and that the nucleus is held fixed, so that atomic recoil is neglected. If the monopole moves along the z-axis, then the z-component of angular momentum

$$J_z = [\mathbf{r} \times (\mathbf{p} - e\mathbf{A}) + \tfrac{1}{2}\boldsymbol{\sigma} - \tfrac{1}{2}\hat{n}]_z \qquad 145.$$

is conserved (84); here $\mathbf{r}$ is the electron coordinate relative to the nucleus and $\hat{n}$ is the unit vector pointing from the monopole to the electron. When the monopole is very distant from the atom, the electron Hamiltonian and angular momentum reduce to those of a simple hydrogen atom. But, because the last term in Equation 145 changes sign as the monopole moves from the far left ($\hat{n}_z = -1$) to the far right ($\hat{n}_z = +1$), we see that the passage of the monopole causes one unit of J_z to be transferred to the electron.

The low-lying levels of the hydrogen atom are sketched in Figure 11 as a function of the separation z between the nucleus and monopole. When $|z|$ is large compared to the Bohr radius a_0, the levels approach those of an unperturbed hydrogen atom; the ground state is a degenerate 1S doublet with $m = \pm\frac{1}{2}$, if we neglect the hyperfine splitting. Since the passage of the monopole increases m by 1, the $m = \frac{1}{2}$ member of the 1S doublet must evolve into the $m = \frac{3}{2}$ member of the 2P multiplet. If the $m = \pm\frac{1}{2}$ ground-state levels are occupied with equal probabilities, then there is a 50% chance that the passage of the monopole will excite the hydrogen atom.

When the positions of the nucleus and monopole coincide ($z = 0$), all three components of $\mathbf{J}$ are again conserved, and the energy levels can be

computed exactly (78, 135); the ground state is an angular momentum singlet, and the first excited state is an angular momentum triplet. Thus, at $z = 0$, the levels with $J_z = -1, 0, 1$ cross. But if the impact parameter of the incident monopole is nonzero, J_z is not conserved and a level crossing cannot occur. Therefore, in the adiabatic limit, excitation of the atom occurs only for zero impact parameter.

However, if a monopole with impact parameter b moves at velocity v, excitation is likely to occur as long as the levels approach within $\hbar\omega$ of one another, where $\omega \lesssim v/b$. Calculation indicates that the hydrogen energy levels approach within a few tenths of an electron volt of one another even if b is of order a_0 (134). The excitation cross section is therefore surprisingly large; the calculated energy loss per unit density in hydrogen is (134)

$$\frac{dE}{dx} = 37(\beta/10^{-4})\text{MeV cm}^2\text{ g}^{-1}, \qquad 146.$$

if atomic recoil is neglected. When atomic recoil is included, there is a kinematic threshold for excitation at $\beta = 1.5 \times 10^{-4}$.

This calculation illustrates that a reasonably large energy loss is possible for β as small as 10^{-4}, but also that a detailed understanding of the atomic levels in the presence of the monopole is necessary before a quantitative calculation of the energy loss can be performed. Nonetheless, a similar qualitative picture is probably applicable to more complicated materials. For example, a monopole incident on a Z-electron atom with zero impact

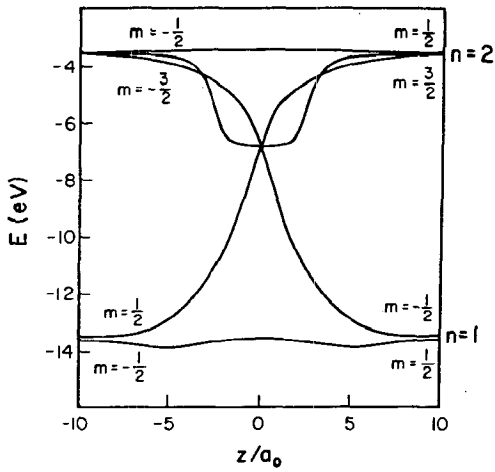

Figure 11 Electronic energy levels of a hydrogen atom in the vicinity of a Dirac monopole. The position of the monopole, relative to the atomic nucleus, is (0, 0, z).

parameter will transfer Z units of angular momentum to the atom, and it seems likely that level crossings will be induced like those that occur in the hydrogen atom. The excited many-electron atom should then be able to autoionize, unlike the excited hydrogen atom, which is in a 2P state and must decay radiatively.

Other calculational schemes have also indicated that monopoles with β down to 10^{-4} have a detectable ionization loss in many materials (115). It is thus probable that existing ionization detectors are capable of detecting monopoles with $\beta \sim 10^{-4}$, and in any case we can be quite confident that monopoles with $\beta \sim 10^{-3}$ are detectable.

Since it is unreasonable to expect most of the cosmic ray monopoles incident on the Earth to have velocities much below $10^{-3}c$ (escape velocity from the galaxy is of order $10^{-3}c$), ionization detector experiments should be able to place useful limits on the cosmic ray monopole flux. The best current limit is $F \lesssim 7 \times 10^{-13}$ cm^{-2} s^{-1} sr^{-1} (136). It seems to be technically feasible to build much larger detectors that can improve this limit by several orders of magnitude.

9.3 *Catalysis Detectors*

Astrophysical arguments place very discouraging constraints on the flux in cosmic rays of magnetic monopoles that catalyze nucleon decay at a strong-interaction rate. It is nonetheless worthwhile to conduct experimental searches for such monopoles.

For one thing, as discussed in Section 8.2, the astrophysical arguments could be wrong. It is also conceivable that catalysis occurs at a rate small enough to be unimportant in neutron stars, but still large enough to be detectable in terrestrial experiments. The point is that the monopole may have a reasonably large cross section for capturing a nucleon; after capture it will hold onto the nucleon until it is able to catalyze its decay. Even if the nucleon must wait 10^{-6} s after capture before finally decaying, the monopole, traveling at 10^{-3} c, will have moved less than a meter. The catalysis would be no harder to observe than if it had followed 10^{-23} s after capture, as long as the detector is more than a meter long (137).

Experiments designed to search for spontaneous nucleon decay are also well suited to detect nucleon decay catalyzed by a monopole. If catalysis (or capture) occurs with a typical strong-interaction cross section, then several nucleon decay events will occur in the detector along the trajectory of the monopole. This distinctive signature allows catalyzed nucleon decay to be distinguished from spontaneous decay. Unsuccessful searches for such multiple nucleon decay events have placed the limit $F \lesssim 7 \times 10^{-15}$ cm^{-2} s^{-1} sr^{-1} on the monopole flux, assuming a catalysis cross section σ greater than 10^{-26} cm^2 (138).

The magnetic monopole continues to be strangely elusive. But just as theorists have continued to explore the wonders of the monopole with undeterred enthusiasm in spite of its elusiveness, so experimenters will continue to stalk the monopole with unfailing determination and ingenuity.

ACKNOWLEDGMENTS

I have learned about magnetic monopoles from many colleagues, especially S. Coleman, P. Nelson, and F. Wilczek. In preparing this manuscript, I have benefited from conversations with H. Sonada and N. Warner.

Literature Cited

1. Dirac, P. A. M. 1931. *Proc. R. Soc. London A* 133:60
2. Polyakov, A. M. 1974. *JETP Lett.* 20: 194
3. 't Hooft, G. 1974. *Nucl. Phys. B* 79:276
4. Georgi, H., Glashow, S. L. 1974. *Phys. Rev. Lett.* 32:438
5. Pati, J. C., Salam, A. 1974. *Phys. Rev. D* 10:275
6. Georgi, H., Quinn, H., Weinberg, S. 1974. *Phys. Rev. Lett.* 33:451
7. Kibble, T. W. B. 1976. *J. Phys. A* 9-1387
8. Zeldovich, Ya. B., Khlopov, M. Y. 1978. *Phys. Lett.* 79*B*:239
9. Preskill, J. 1979. *Phys. Rev. Lett.* 43: 1365
10. Guth, A. H. 1981. *Phys. Rev. D* 23:347
11. Linde, A. D. 1982. *Phys. Lett.* 108*B*: 389; 114*B*:431
12. Albrecht, A., Steinhardt, P. J. 1982. *Phys. Rev. Lett.* 48:1220
13. Parker, E. N. 1970. *Astrophys. J.* 160: 383
14. Turner, M. S., Parker, E. N., Bogdan, T. J. 1982. *Phys. Rev. D* 26:1296
15. Rubakov, V. 1981. *JETP Lett.* 33:644; 1982. *Nucl. Phys. B* 203:311; Rubakov, V. A., Serbryakov, M. S. 1983. *Nucl. Phys. B* 218:240
16. Callan, C. G. 1982. *Phys. Rev. D* 25: 2141; *D* 26:2058; 1983. *Nucl. Phys. B* 212:391
17. Langacker, P. 1981. *Phys. Rep.* 72*C*: 185
18. Ramond, P. M. 1983. *Ann. Rev. Nucl. Part. Sci.* 33:31
19. Coleman, S. 1977. In *New Phenomena in Subnuclear Physics*, ed. A. Zichichi. London: Plenum
20. Goddard, P., Olive, D. 1978. *Rep. Prog. Phys.* 41:1357
21. Coleman, S. 1983. In *The Unity of the Fundamental Interactions*, ed. A. Zichichi. London: Plenum
22. Weinberg, E. J. 1982. In *Particles and Fields—1981: Testing the Standard Model*, ed. C. A. Heusch, W. T. Kirk. New York: Am. Inst. Phys.
23. Preskill, J. 1983. In *The Very Early Universe*, ed. S. W. Hawking, G. W. Gibbons, S. Siklos. Cambridge: Cambridge Univ. Press
24. Turner, M. S. 1983. See Ref. 27, pp. 127–40
25. Lazarides, G. 1983. See Ref. 27, pp. 71–80
26. Preskill, J. 1984. See Ref. 28
27. Carrigan, R. A., Trower, W. P., eds. 1983. *Magnetic Monopoles*. New York: Plenum
28. Stone, J., ed. 1984. *Proceedings of Monopole '83*. New York: Plenum. In press
29. Aharonov, Y., Bohm, D. 1959. *Phys. Rev.* 115:485
30. 't Hooft, G. 1976. *Nucl. Phys. B* 105: 538; Corrigan, E., Olive, D., Fairlie, D., Nuyts, J. 1976. *Nucl. Phys. B* 106: 475
31. 't Hooft, G. 1978. *Nucl. Phys. B* 138:1
32. Strominger, A. 1983. *Commun. Nucl. Part. Phys.* 11:149; Preskill, J. 1983. See Ref. 27, pp. 111–26; Barr, S. M., Reiss, D. B., Zee, A. 1983. *Phys. Rev. Lett.* 50:317; Aoyama, S., Fujimoto, Y., Zhao, Z. 1983. *Phys. Lett.* 124*B*: 185; Pantaleone, J. 1983. *Nucl. Phys. B* 219:367; Rubakov, V. 1983. *Phys. Lett.* 120*B*:191
33. Goddard, P., Nuyts, J., Olive, D. 1977. *Nucl. Phys. B* 125:1
34. Schwinger, J. 1966. *Phys. Rev.* 144: 1087; Zwanziger, D. 1968. *Phys. Rev.* 176:1480, 1489
35. Glashow, S. L. 1961, *Nucl. Phys.*

22:579; Weinberg, S. 1967. *Phys. Rev. Lett.* 19:1264; Salam, A. 1968. In *Elementary Particle Physics*, ed. N. Svartholm. Stockholm: Almquist & Wiksells
36. Georgi, H., Glashow, S. L. 1972. *Phys. Rev. D* 6:2977
37. Kirkman, T. W., Zachos, C. K. 1981. *Phys. Rev. D* 24:999
38. Bogomol'nyi, E. 1976. *Sov. J. Nucl. Phys.* 24:449; Coleman, S., Parke, S., Neveu, A., Sommerfield, C. 1977. *Phys. Rev. D* 15:544
39. Prasad, M., Sommerfield, C. 1975. *Phys. Rev. Lett.* 35:760
40. Wilkinson, D., Goldhaber, A. 1977. *Phys. Rev. D* 21:1221; Weinberg, E. J. 1984. *Phys. Lett.* 136*B*:179
41. Goldhaber, A., Wilkinson, D. 1976. *Nucl. Phys. B* 114:317
42. Wu, T. T., Yang, C. N. 1976. *Nucl. Phys. B* 107:365; 1975. *Phys. Rev. D* 12:3845
43. Lubkin, E. 1963. *Ann. Phys.* 23:233
44. Hilton, P. J. 1953. *An Introduction to Homotopy Theory*. Cambridge: Cambridge Univ. Press; Steenrod, N. E. 1951. *The Topology of Fibre Bundles*. Princeton: Princeton Univ. Press; Nash, C., Sen, S. 1983. *Topology and Geometry for Physicists*. New York: Academic
45. Brandt, R., Neri, F. 1979. *Nucl. Phys. B* 161:253
46. Husemoller, D. 1966. *Fibre Bundles*. New York: McGraw-Hill; Steenrod, N. 1951. See Ref. 44
47. Kaluza, Th. 1921. *Sitzungsber. Preuss. Akad. Wiss. Berlin, Math. Phys. K* 1:966; Klein, O. 1926. *Z. Phys.* 37:895
48. Gross, D., Perry, M. 1983. *Nucl. Phys. B* 226:29
49. Sorkin, R. 1983. *Phys. Rev. Lett.* 51:87
50. Nelson, P., Manohar, A. 1983. *Phys. Rev. Lett.* 50:943
51. Balachandran, A. P., Marmo, G., Mukunda, N., Nilson, J. S., Sudarshan, E. C. G., Zaccaria, F. 1983. *Phys. Rev. Lett.* 50:1553
52. Gilmore, R. 1974. *Lie Groups, Lie Algebras, and Some of Their Applications*. New York: Wiley
53. Bais, F. A. 1981. *Phys. Lett.* 98*B*:437
54. Nielsen, H., Olesen, P. 1973. *Nucl. Phys. B* 61:45
55. Weinberg, E. J., London, D., Rosner, J. L. 1983. *Fermi Inst. Preprint, EFI 83-39*. Batavia, Ill: Fermilab
56. Gardner, C. L., Harvey, J. A. 1984. *Phys. Rev. Lett.* 52:879
57. Georgi, H. 1975. In *Particles and Fields—1974*, ed. C. E. Carlson. New York: Am. Inst. Phys.; Fritzsch, H., Minkowski, P. 1975. *Ann. Phys.* 93:193
58. Lazarides, G., Magg, M., Shafi, Q. 1980. *Phys. Lett.* 97*B*:87
59. Dawson, S., Schellekens, A. N. 1983. *Phys. Rev. D* 27:2119
60. Ginsparg, P., Coleman, S. 1983. Unpublished
61. Kibble, T. W. B., Lazarides, G., Shafi, Q. 1982. *Phys. Rev. D* 26:435
62. Julia, B., Zee, A. 1975. *Phys. Rev. D* 11:2227
63. Tomboulis, E., Woo, G. 1976. *Nucl. Phys. B* 107:221
64. Aboulsaood, A. 1984. *Nucl. Phys. B* 226:309; 1983. *Phys. Lett.* 125*B*:467
65. Coleman, S., Nelson, P. 1983. *Harvard Preprint HUTP-83/A067*. Cambridge, Mass: Harvard Univ.
66. Dokos, C., Tomaras, T. 1980. *Phys. Rev. D* 21:2940
67. Nelson, P. 1983. *Phys. Rev. Lett.* 50:939
68. Witten, E. 1979. *Phys. Lett.* 86*B*:293
69. Wilczek, F. 1982. *Phys. Rev. Lett.* 48:1146
70. 't Hooft, G. 1981. *Nucl. Phys. B* 109:455
71. Wilczek, F. 1982. *Phys. Rev. Lett.* 48:1144
72. Jackiw, R., Rebbi, C. 1976. *Phys. Rev. Lett.* 36:1116; 't Hooft, G., Hassenfratz, P. 1976. *Phys. Rev. Lett.* 36:1119
73. Goldhaber, A. 1976. *Phys. Rev. Lett.* 36:1122
74. Mandelstam, S. 1976. *Phys. Rep.* 23*C*:235
75. Adler, S. L. 1969. *Phys. Rev.* 177:2426; Bell, J. S., Jackiw, R. 1969. *Nuova Cimento* 51:47
76. 't Hooft, G. 1976. *Phys. Rev. Lett.* 37:8; Jackiw, R., Rebbi, C. 1976. *Phys. Rev. Lett.* 37:177; Callan, C. G., Dashen, R. F., Gross, D. J. 1976. *Phys. Lett.* 63*B*:334
77. Grossman, B. 1983. *Phys. Rev. Lett.* 50:464; Yamagishi, H. 1983. *Phys. Rev. D* 27:2383
78. Kazama, Y., Yang, C. N., Goldhaber, A. S. 1977. *Phys. Rev. D* 15:2287
79. Goldhaber, A. S. 1977. *Phys. Rev. D* 16:1815
80. Jackiw, R., Rebbi, C. 1976. *Phys. Rev. D* 13:3398
81. Goldstone, J., Wilczek, F. 1981. *Phys. Rev. Lett.* 47:986
82. Appelquist, T., Carazzone, J. 1975. *Phys. Rev. D* 11:2856
83. Tamm, I. 1931. *Z. Phys.* 71:141
84. Fierz, M. 1944. *Helv. Phys. Acta* 17:27
85. Coleman, S. 1975. *Phys. Rev. D* 11:2088; Mandelstam, S. 1975. *Phys. Rev. D* 11:3026
86. Besson, C. 1982. PhD thesis. Princeton Univ.
87. Kazama, Y., Sen, A. 1983. *Fermilab-*

Pub-83/58-THY. Batavia, Ill: Fermilab; Yan, T.-M. 1983. *Cornell Preprint CLNS-83/563*. Ithaca, NY: Cornell Univ.; Balachandran, A. P., Schechter, J. 1983. *Syracuse Preprint*; Ezawa, Z. F., Iwazaki, A. 1983. Preprint: Yamagishi, H. 1983. *Princeton preprint*. Princeton Univ.

88. Grossman, B., Lazarides, G., Sanda, A. I. 1983. *Phys. Rev. D* 28:2109; Goldhaber, A. S. 1983. *SUNY Preprint ITP-SB-83-30*
89. Callan, C. G. 1983. *Princeton preprint print-83-0306*. Princeton Univ.
90. Sen, A. 1983. *Fermilab-Pub-83/28*. Batavia, Ill: Fermilab
91. Dawson, S., Schellekans, A. N. 1983. *Phys. Rev. D* 28:3125
92. Ellis, J., Nanopoulos, D. V., Olive, K. A. 1982. *Phys. Lett.* 116*B*:127; Bais, F. A., Ellis, J., Nanopoulos, D. V., Olive, K. A. 1983. *Nucl. Phys. B* 219:189
93. Fiorentini, G. 1984. See Ref. 28
94. Callan, C. G. Witten, E. 1984. *Princeton preprint print-84-0054*. Princeton Univ.
95. Kirzhnits, D., Linde, A. 1972. *Phys. Lett.* 42*B*:471; Weinberg, S. 1974. *Phys. Rev. D* 9:3357; Dolan, L., Jackiw, R. 1974. *Phys. Rev. D* 9:3320
96. Guth, A. H., Tye, S.-H. 1980. *Phys. Rev. Lett.* 44:631; Einhorn, M. B., Stein, D. L., Toussaint, D. 1980. *Phys. Rev. D* 21:3295
97. Peebles, P. J. E. 1971. *Physical Cosmology*. Princeton Univ. Press; Weinberg, S. 1972. *Gravitation and Cosmology*. New York: Wiley
98. Coleman, S. 1977. *Phys. Rev. D* 15:2929
99. Dicus, D. A., Page, D. N., Teplitz, V. L. 1982. *Phys. Rev. D* 26:1306
100. Guth, A. H., Pi, S.-Y. 1982. *Phys. Rev. Lett.* 49:1110; Hawking, S. W., Moss, I. G. 1983. *Nucl. Phys. B* 224:180; Bardeen, J. M., Steinhardt, P. J., Turner, M. S. 1983. *Phys. Rev. D* 28:679; Starobinsky, A. A. 1982. Unpublished
101. Turner, M. S. 1982. *Phys. Lett.* 115*B*:95
102. Lazarides, G., Shafi, Q., Trower, W. P. 1982. *Phys. Rev. Lett.* 49:1756
103. Turner, M. S. 1982. Unpublished; Goldhaber, A. S., Guth, A. H., Pi, S.-Y. 1982. Unpublished
104. Collins, W., Turner, M. S. 1983. *Fermi Inst. Preprint 83-41*. Chicago: Fermi Inst.
105. Langacker, P., Pi, S.-Y. 1980. *Phys. Rev. Lett.* 45:1
106. Vilenkin, A. 1982. *Nucl. Phys. B* 196:240; Bais, F. A., Langacker, P. 1982. *Nucl. Phys. B* 197:520
107. Weinberg, E. 1983. *Phys. Lett.* 126*B*:441
108. Fleischer, R. L., et al. 1969. *Phys. Rev.* 184:1393, 1398
109. Price, P. B., et al. 1978. *Phys. Rev. D* 18:1382
110. Kolm, H. H., et al. 1971. *Phys. Rev. D* 4:1285
111. Eberhard, P., et al. 1971. *Phys. Rev. D* 4:3260
112. Lazarides, G., Shafi, Q., Walsh, T. F. 1981. *Phys. Lett.* 100*B*:21
113. Longo, M. J. 1982. *Phys. Rev. D* 25:2399
114. Parker, E. N. 1979. *Cosmical Magnetic Fields*. Oxford: Clarendon
115. Ahlen, S. P., Kinoshita, K. 1982. *Phys. Rev. D* 26:2347
116. Salpeter, E. E., Shapiro, S. L., Wasserman, I. 1982. *Phys. Rev. Lett.* 49:1114
117. Arons, J., Blandford, R. D. 1983. *Phys. Rev. Lett.* 50:544
118. Dimopoulos, S., Glashow, S. L., Purcell, E. M., Wilczek, F. 1982. *Nature* 298:824
119. Freese, K., Turner, M. S. 1983. *Phys. Lett.* 123*B*:293
120. Kolb, E. W., Colgate, S. A., Harvey, J. A. 1982. *Phys. Rev. Lett.* 49:1373
121. Dimopoulos, S., Preskill, J., Wilczek, F. 1982. *Phys. Lett* 119*B*:320
122. Freese, K., Turner, M. S., Schramm, D. N. 1983. *Phys. Rev. Lett.* 51:1625
123. Harvey, J. A., Ruderman, M., Shaham, J. 1983. Unpublished; Harvey, J. A. 1984. See Ref. 28
124. Irvine, J. 1978. *Neutron Stars*. Oxford: Clarendon
125. Harvey, J. A. 1984. *Nucl. Phys. B* 236:255
126. Freese, K. 1983. *Univ. Chicago preprint 84-0573*
127. Giacomelli, G. 1983. See Ref. 27, pp. 41–70
128. Giacomelli, G. 1984. See Ref. 28
129. Barish, B. C. 1984. See Ref. 28
130. Tassie, L. J. 1965. *Nuovo Cimento* 38:1935
131. Cabrera, B. 1982. In *Third Workshop on Grand Unification*, ed. P. H. Frampton, S. L. Glashow, H. van Dam. Boston: Birkhäuser
132. Cabrera, B. 1984. See Ref. 28
133. Cabrera, B. 1982. *Phys. Rev. Lett.* 48:1378
134. Drell, S., Kroll, N., Mueller, M., Parke, S., Ruderman, M. 1983. *Phys. Rev. Lett.* 50:644
135. Malkus, W. V. R. 1951. *Phys. Rev.* 83:899
136. Kajino, F., et al. 1984. *Phys. Rev. Lett.* 52:1373
137. Goldhaber, A. S. 1983. See Ref. 27, pp. 1–17
138. Errede, S., et al. 1983. *Phys. Rev. Lett.* 51:245
139. Einhorn, M. B., Sato, K., 1981. *Nucl. Phys. B* 180:385

Annual Review of Nuclear and Particle Science
Volume 34, 1984

CONTENTS