PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Review

**Cite this article:** Wilczek F. 2016 Unification of force and substance. *Phil. Trans. R. Soc. A* **374**: 20150257.
http://dx.doi.org/10.1098/rsta.2015.0257

Accepted: 19 April 2016

One contribution of 7 to a discussion meeting issue 'Unifying physics and technology in light of Maxwell's equations'.

**Subject Areas:**
electromagnetism, gauge theory, high energy physics, particle physics, quantum physics

**Keywords:**
gauge invariance, unification, magnetic monopole, supersymmetry

**Author for correspondence:**
Frank Wilczek
e-mail: wilczek@mit.edu

Speech presented at *Unifying physics and technology in light of Maxwell's equations*, Discussion Meeting at the Royal Society, London, 16–17 November 2015, celebrating the 150th anniversary of Maxwell's equations.

THE ROYAL SOCIETY
PUBLISHING

# Unification of force and substance

Frank Wilczek

Center for Theoretical Physics, MIT, Cambridge, MA 02139, USA

FW, 0000-0002-6489-6155

Maxwell's mature presentation of his equations emphasized the unity of electromagnetism and mechanics, subsuming both as 'dynamical systems'. That intuition of unity has proved both fruitful, as a source of pregnant concepts, and broadly inspiring. A deep aspect of Maxwell's work is its use of redundant potentials, and the associated requirement of gauge symmetry. Those concepts have become central to our present understanding of fundamental physics, but they can appear to be rather formal and esoteric. Here I discuss two things: the physical significance of gauge invariance, in broad terms; and some tantalizing prospects for further unification, building on that concept, that are visible on the horizon today. If those prospects are realized, Maxwell's vision of the unity of field and substance will be brought to a new level.

This article is part of the themed issue 'Unifying physics and technology in light of Maxwell's equations'.

## 1. Introduction

As we celebrate the sesquicentennial of Maxwell's equations, there are many facets to celebrate: the triumph of the field concept, the roots of relativity, the cornerstone of a host of technologies. In many ways, the Maxwell equations mark the beginning of modern physics, and indeed of the modern world.

Here I will discuss a special, hidden theme of Maxwell's work, primarily important for fundamental physics, whose profundity only emerged gradually, and may still not stand fully revealed even today. This is the concept of gauge symmetry. Gauge symmetry is perhaps the supreme realization of Heinrich Hertz's famous tribute to the Maxwell equations:

> One cannot escape the feeling that these mathematical formulae have an independent existence and an intelligence of their own,

> that they are wiser than we are, wiser even than their discoverers, that we get more out of them than was originally put into them.

## 2. Gauge symmetry

*Local gauge symmetry* arises when one formulates the Maxwell equations using potentials. It states that one can subject the potentials to an enormous group of transformations, parametrized by a function of space and time, while leaving their physical consequences unchanged. On the surface this seems like a very odd principle. It begs the question, why one uses the (largely arbitrary) potentials at all, rather than working in terms of invariants. Yet this principle, suitably generalized, has proved to be a reliable guide to formulating successful laws of fundamental physics. In different forms, it is central to our theories of the strong, weak and gravitational interactions as well as electromagnetism. It may point the way to more profound, unified understanding of all those forces, as I shall discuss.

How did Maxwell get to this concept? Why does Nature use it? Those are the questions I would like to address now.

### (a) Maxwell, Faraday and Hamilton

Maxwell's mature presentation of his equations, in the *Treatise* [1], emphasized the unity of electromagnetism and mechanics, subsuming both in the concept of (Hamiltonian) 'dynamical system'. In Article §567, he wrote

> In formulating the ideas and words relating to any science, which, like electricity, deals with forces and their effects, we must keep constantly in mind the ideas appropriate to the fundamental science of dynamics, so that we may, during the first development of the science, avoid inconsistency with what is already established . . .

At the same time, Maxwell was determined to embody the physical intuitions of Faraday. In the preface to the first edition of the *Treatise* he wrote, contrasting Faraday's conceptions to the action-at-a-distance theories then more popular among mathematical physicists

> . . . Faraday, in his mind's eye, saw lines of force traversing all space where the mathematicians saw centres of force attracting at a distance: Faraday saw a medium where they saw nothing but distance: Faraday sought the seat of the phenomena in real actions going on in the medium, they were satisfied that they had found it in a power of action at a distance impressed on the electric fluids.

and a paragraph later

> The whole theory, for instance, of the potential, considered as a quantity which satisfies a certain partial differential equation, belongs essentially to the method which I have called that of Faraday.

Faraday's concept of 'electrotonic state' was a non-mathematical anticipation of the vector potential, which Maxwell took over in formulating his equations. For Faraday, it was a way of expressing how his space-filling fields influence electric currents. It is perfectly adapted to Maxwell's program of founding the description of electromagnetic phenomena on the mechanical notion of local energy density, since it allows one to express the energy density $\boldsymbol{A} \cdot \boldsymbol{j}$ associated with a current $\boldsymbol{j}$.

The arbitrariness of the potentials is reflected in the possibility that one can make a wide class of transformations upon them, without changing their physical content. The fields

$$\left.\begin{aligned} & & \boldsymbol{E} &= -\boldsymbol{\nabla}\phi - \frac{\partial \boldsymbol{A}}{\partial t} \\ &\text{and} & \boldsymbol{B} &= \boldsymbol{\nabla} \times \boldsymbol{A} \end{aligned}\right\} \tag{2.1}$$

are manifestly invariant under the transformations

$$\left.\begin{aligned} & & \phi' &= \phi + \frac{\partial \Lambda}{\partial t} \\ &\text{and} & \boldsymbol{A}' &= \boldsymbol{A} - \boldsymbol{\nabla}\Lambda. \end{aligned}\right\} \tag{2.2}$$

The invariance of the work expressions is less obvious. We have

$$\begin{aligned} \rho\phi' - \boldsymbol{j}\cdot\boldsymbol{A}' &= \rho\phi - \boldsymbol{j}\cdot\boldsymbol{A} + \rho\frac{\partial \Lambda}{\partial t} + \boldsymbol{j}\cdot\boldsymbol{\nabla}\Lambda \\ &= \rho\phi - \boldsymbol{j}\cdot\boldsymbol{A} + \{\partial_t(\rho\Lambda) + \nabla\cdot(\boldsymbol{j}\Lambda)\} - \{\{\Lambda(\partial_t\rho + \boldsymbol{\nabla}\cdot\boldsymbol{j})\}\} \\ &\approx \rho\phi - \boldsymbol{j}\cdot\boldsymbol{A}, \end{aligned} \tag{2.3}$$

where the term in single braces vanishes upon integration, as long as $\Lambda$ vanishes at infinity. The term in double braces vanishes if, and in general only if, charge is locally conserved.

The relationship between gauge invariance and charge conservation becomes especially transparent in relativistic notation. The invariance of the coupling (energy) expression $\int A_\mu j^\mu$ under $A_\mu \to A_\mu + \partial_\mu\Lambda$ is equivalent, upon integration by parts, to the conservation equation $\partial_\mu j^\mu = 0$.

In keeping with my theme of unification, it is appropriate to add that Hamilton was led to his formulation of mechanics in large part by a desire to unite the mathematical treatment of mechanics and optics, including wave optics. Following the experimental discoveries of quantum behaviour, those alternative descriptions enabled a smooth implementation of wave-particle duality. Indeed, the mathematical blueprint for passing from Hamiltonian ‘wave’ mechanics, epitomized in the Hamilton–Jacobi equation, to the Schrödinger equation is astoundingly simple.

## (b) Quantum theory

### (i) Dirac

Dirac spelled out the abstract ‘particle’ formulation of Hamiltonian dynamics for electromagnetism, which Maxwell had gestured towards (but did not work out in detail). In that formulation, the modes of the electromagnetic field appear as independent harmonic oscillators. Dirac could then apply his general quantization procedure, replacing Poisson brackets by commutators, to get a quantum theory of the electromagnetic field [2]. This theory has proved to be tremendously accurate. It has been supplemented, but not replaced, within our current best understanding of Nature.

The quantum theory, through its *reliance* on potentials, sheds new light on a basic question which the Maxwell equations suggest. There is no difficulty in introducing magnetic charges and currents into the classical Maxwell equations, at the level of field strengths. The system of equations including magnetic charges and currents

$$\left.\begin{aligned} & & \boldsymbol{\nabla}\cdot\boldsymbol{B} &= \rho_{\mathrm{m}}, \\ & & \boldsymbol{\nabla}\cdot\boldsymbol{E} &= \rho, \\ & & \boldsymbol{\nabla}\times\boldsymbol{B} &= \frac{\partial \boldsymbol{E}}{\partial t} + \boldsymbol{J} \\ &\text{and} & \boldsymbol{\nabla}\times\boldsymbol{E} &= -\frac{\partial \boldsymbol{B}}{\partial t} - \boldsymbol{J}_m \end{aligned}\right\} \tag{2.4}$$

supports a conservation law for magnetic charge

$$\frac{\partial \rho_{\mathrm{m}}}{\partial t} + \boldsymbol{\nabla} \cdot \boldsymbol{J}_{\mathrm{m}} = 0 \tag{2.5}$$

and is no less consistent than the original Maxwell equations. It even appears more symmetric. From this perspective, the observed absence of free magnetic charge seems puzzling. Why doesn't Nature make use of that possibility?

In quantum theory, however, the picture is different. Quantum theory, in all known formulations, requires not only field strengths, but also vector potentials. Once we introduce vector potentials, to write field strengths in the form (now using relativistic notation)

$$F_{\alpha\beta} = \partial_\alpha A_\beta - \partial_\beta A_\alpha \tag{2.6}$$

with

$$\boldsymbol{E}_j = -F_{0j} \tag{2.7}$$

and

$$\boldsymbol{B}_j = \tfrac{1}{2}\epsilon_{jkl}F_{kl}, \tag{2.8}$$

we find that the first Maxwell equation in its usual form, stating the absence of magnetic charge

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \tag{2.9}$$

holds true as an identity, as does Faraday's law of induction.

It is remarkable that Faraday's great empirical discovery, which came as a great surprise in its time, now appears as a mathematical identity, required by the logic of profound theoretical principles.

Dirac later demonstrated [3] that if we loosen the framework, by allowing certain kinds of singularities in the vector potentials, then we can introduce magnetic charge while maintaining many of the basic principles of relativistic quantum theory. There are significant—possibly insuperable—difficulties in accommodating such singularities within a fully dynamical theory of pure electrodynamics. But in a final twist to this story, 'tHooft [4] and Polyakov [5] showed that if electrodynamics is embedded within larger gauge theories, including the possibly realistic ones described below, then magnetic monopoles can appear without requiring singular potentials. The extra degrees of freedom that resolve the singularity of the purely electromagnetic potentials are predicted to be very massive, so the magnetic monopoles are predicted to be very heavy—out of the reach of accelerators.[1] From a larger perspective, it is a major achievement of Hamiltonian mechanics, and of the quantum theory that builds upon it, to have explained the pronounced *asymmetry* between electric and magnetic charge, by emphasizing the need to introduce potentials.

### (ii) Longitudinal modes

Given that potentials are desirable, in order to formulate electromagnetic field theory as a dynamical system, we may ask why Nature chooses to make them largely redundant, by enforcing gauge symmetry. Relativistic quantum theory supplies a profound answer to that question.

Gauge transformations, as shown in equation (2.2), allow us to change the longitudinal (in the $3+1$ dimensional sense) part of the potential field at will. Gauge symmetry, therefore, basically states that the longitudinal modes have no direct physical significance. They appear in the equations, but not in the consequences of those equations. In common jargon, we say that longitudinal photons decouple.

That decoupling of longitudinal photons is an important consistency requirement, because special relativity forces a relativistically invariant norm to have opposite sign for photons with time-like versus space-like polarization. If we are going to identify the norm with a probability,

[1] The absence of residual magnetic monopoles emerging from the big bang is a separate issue. Their observed absence was one of the original motivations for introducing the idea of cosmic inflation.

we must make sure that the time-like photons do not leak into our world. Gauge invariance insures this, as I will now indicate more precisely and mathematically.

Using local charge conservation, we can decompose the electric current at spatial momentum $\boldsymbol{k}$ into longitudinal and transverse parts, according to

$$\boldsymbol{j} = -\omega \frac{\boldsymbol{k}}{k^2}\rho + \boldsymbol{j}^{\perp}. \tag{2.10}$$

Then we have

$$\rho_1\rho_2 - \boldsymbol{j}_1 \cdot \boldsymbol{j}_2 = \rho_1\rho_2\left(1 - \frac{\omega^2}{k^2}\right) - \boldsymbol{j}_1^{\perp} \cdot \boldsymbol{j}_2^{\perp} \tag{2.11}$$

and so for the photon-mediated coupling, allowing all four polarizations,

$$\frac{\rho_1\rho_2 - \boldsymbol{j}_1^{\perp} \cdot \boldsymbol{j}_2^{\perp}}{\omega^2 - k^2} = -\frac{\rho_1\rho_2}{k^2} - \frac{\boldsymbol{j}_1^{\perp} \cdot \boldsymbol{j}_2^{\perp}}{\omega^2 - k^2}. \tag{2.12}$$

The two surviving terms are an instantaneous Coulomb force and a coupling to *transverse* dynamical radiation—or, in the quantum theory, to transverse photons—only. Those effects form the core of physical electrodynamics.

### (iii) Working backwards

As the preceding manipulations indicate, one can work backwards, 'deriving' the Maxwell equations as a consequence of gauge symmetry and special relativity. (And as we've seen, gauge invariance itself is closely tied to the possibility of formulating a consistent quantum theory.) That cluster of ideas remains valid, and becomes more precise and rigorous, in the formal treatment of relativistic quantum field theory. One finds that gauge invariance and special relativity, together with general principles of locality and renormalizability, lead one directly and uniquely to modern quantum electrodynamics, with Maxwell's equations as its heart.

## (c) Local symmetry

Gauge invariance is a vast symmetry. It allows transformations that depend freely on time and space. We say that such transformations define a local symmetry group.

Similar to how Maxwell's electrodynamics can be derived from local gauge invariance, Einstein's general relativity can be derived from symmetry under local—that is, space–time dependent—Lorentz transformations. However, that understanding of general relativity—and of electrodynamics—reverses the historical order of discovery. Einstein, guided by experimental facts (equivalence principle, Newtonian limit) arrived at general relativity through a mixture of geometrical ideas and inspired intuition [6]. Only later did Weyl [7] clarify and emphasize the local symmetry aspect, in an attempt to understand both gravity and electrodynamics on a similar footing. (See below.)

On the other hand, Yang & Mills [8], in their pioneering attempt to generalize Maxwell electrodynamics, consciously started from local symmetry. They showed how one could construct interesting relativistic field theories accommodating more complicated gauge transformations, where the parameter $\Lambda(x)$ is not simply a numerical function, but a function taking values in a compact group. Amazingly, their mathematical construction forms the core of our present theories of the weak and strong interaction.

Although in fundamental physics the central significance of local symmetry is a relatively recent discovery, artists in many cultures have employed it for millennia, especially in decorative art. In such art [9], one often finds large-scale patterns built up from symmetric shapes, such as equilateral triangles, squares and circles. The symmetry of such objects arises from the fact that we can rotate them through appropriate angles around their centres, without changing their overall form. But of course each object in the pattern has its own distinct centre, and so to realize the full symmetry we must allow transformations of space that vary from place to place. Many design patterns also employ transformations of colour, as well as shape, whose symmetry is local.

## (d) Weyl's gauge symmetry

### (i) Gauge symmetry as geometry

The historical origin of the term 'gauge' in gauge theory, and of several important ideas in modern physics, is an incorrect geometric theory of electromagnetism proposed by Hermann Weyl in 1918. Weyl was inspired by Einstein's then-new theory of gravity. The leading idea of Weyl's theory is that the fundamental laws of physics should be invariant under position-dependent scale transformations. We implement those by changing the metric, according to

$$g'_{\mu\nu}(x) = \mathrm{e}^{2\sigma(x)} g_{\mu\nu}(x). \tag{2.13}$$

Note that this is a different operation than using an expanded coordinate system. It implements real changes in size, or scale, or alternatively *gauge*, rather than changes in how we label things.

The term 'gauge symmetry' is a relic of Weyl's original theory.

Weyl, in his original theory, postulated local-scale symmetry. That is, he postulated that one could change the size of objects independently at every point in space-time—and still get the same behaviour! To make that outrageous idea viable, he had to introduce a 'gauge' connection field. The gauge connection field tells us how much we must adjust our scale of length, or re-gauge our rulers, as we move from one point to another. Weyl made the remarkable discovery, that this gauge connection field, in order to do its job of implementing local-scale symmetry, must satisfy the Maxwell equations. Dazzled by that apparent miracle, Weyl proposed to identify his ideal mathematical connection field with the real physical electromagnetic field.

Let me very briefly indicate how that proposal works.

In order to have motion of particles or fields, we must have derivatives. And if we want the derivatives to transform simply under gauge transformations, so that we can package them into invariant equations, we need to adjust them. Specifically, if we consider the simplest non-trivial transformation law for a field $\phi(x)$,

$$\phi'(x) = \mathrm{e}^{-\sigma(x)} \phi(x) \tag{2.14}$$

the derivative $\partial_\mu \phi$ picks up terms involving the derivative of $\sigma$, and transforms horribly. To fix it, Weyl introduced a vector field $\omega_\mu$ and a covariant derivative operation $\nabla_\mu \phi = (\partial_\mu - \omega_\mu)\phi$. Then we will have a good transformation law for the covariant derivative, *viz*.

$$\nabla'_\mu \phi' = (\nabla_\mu \phi)' \tag{2.15}$$

if $\omega$ transforms according to

$$\omega'_\mu = \omega_\mu - \partial_\mu \sigma. \tag{2.16}$$

But this, mathematically, is identical to the standard gauge transformation law for electromagnetic potentials. And, as we have seen, that gauge transformation law (plus special relativity) leads us pretty directly to the Maxwell equations.

Although Weyl's connection field is a necessary ingredient of local-scale symmetry, it is not sufficient to ensure that symmetry. Other properties of matter, such as the dimensions of a proton, give us objective scales of length that do not change as we move from point to point. Einstein and others did not fail to notice this shortcoming of Weyl's theory. Despite its visionary brilliance, Weyl's theory seemed destined for oblivion.

The situation changed, however, with the emergence of quantum theory. In that context electric charge associated with the phase of wave-functions, which we can regard as defining a circular one-dimensional property space, living on top of space–time [10]. In 1929, Weyl exploited this new space, to revive his gauge theory in a modified form. In the new theory, the local symmetry transformations are no longer space-time dependent changes in the scale of length of space–time, but rather rotations in the new dimension, whose coordinate is dual to electric charge. (Note that multiplication by a phase faithfully implements rotations of a circle.) After that modification gauge symmetry leads to a satisfactory theory of electromagnetism, as I have sketched above.

### (ii) Scale transformations as a conceptual tool

There is a close connection between Weyl scaling equation (2.13) with constant $\sigma$ and the scaling transformations

$$x'^{\mu} = \mathrm{e}^{-\sigma} x^{\mu} \tag{2.17}$$

with *fixed*, flat metric that we often consider in high energy and condensed matter physics. If we perform the scaling equation (2.17) as a general coordinate transformation, and then follow it with the Weyl scaling equation (2.13), we get a transformation that leaves the metric fixed. So if general covariance is a valid symmetry, then the 'dimensional analysis' re-scalings considered in high-energy and condensed matter physics are equivalent to constant Weyl transformations.

In modern physics local Weyl transformations, though interesting, do not appear to be a good candidate to provide new fundamental symmetries, or new gauge interactions. There are major barriers to supposing that local-scale invariance holds in Nature. As we already observed, scale invariance would forbid basic properties we seem to need, like particle masses. That difficulty might not be insuperable, as we have become accustomed to exploiting hidden (spontaneously broken) symmetries of equations that are not manifest in their stable solutions. More difficult to finesse, it seems, is the fact that local-scale symmetry is in tension with the basic structure of interacting quantum field theory, due to the need for renormalization—running of couplings. This effect brings in distance dependence, even when the underlying classical theory is scale invariant.

Remarkably, that difficulty may well prove to be a blessing in disguise. For there are situations where that quantum effect, running of the couplings, is the main correction to scale symmetry. Then we can gain insight from an appropriately corrected version of scale symmetry. In the remainder of this lecture, I will indicate how pursuing that line of thought enables us to quantify and refine an attractive proposal for the unification of force and substance.

## 3. Unification of quantum numbers

The 'standard model' gauge theory SU(3) × SU(2) × $U(1)$ gives a successful, economical account of fundamental interactions based on far-reaching symmetry postulates, within the framework of relativity and quantum theory (for extensive documentation and reviews, see especially http://pdg.lbl.gov). General relativity is also readily accommodated using a minimal coupling procedure. As this theory is close to Nature's last word, we should take its remaining aesthetic imperfections seriously.

One class of imperfections arises in the account of masses and mixings, where free parameters proliferate. Unfortunately, there are few really compelling ideas that address this issue. Even the most basic question: Why is there a threefold 'family' repetition of quarks and leptons?—remains wide open. (And if cosmic landscape ideas are right, it may be just an accident, with no deeper explanation.)

A second class of imperfections concerns the core of the standard model that is the symmetry structure. Here there is a compelling idea. The product gauge symmetry structure SU(3) × SU(2) × $U(1)$ practically begs to be embedded into a larger, encompassing symmetry. The electroweak theory, which breaks $\mathrm{SU}(2) \times U(1)_Y \rightarrow U(1)_Q$, shows how the symmetry of fundamental equations can be hidden in their low-energy solutions, by the influence of cosmic fields or condensations (Higgs mechanism). Slightly more elaborate versions of the same mechanism can implement $\mathrm{SU}(5) \rightarrow \mathrm{SU}(3) \times \mathrm{SU}(2) \times U(1)$ or $\mathrm{SO}(10) \rightarrow \mathrm{SU}(3) \times \mathrm{SU}(2) \times U(1)$, as we will presently discuss.

An important test for the hypothetical expanded symmetries is whether they act naturally on the quarks and leptons. Indeed, another 'imperfection' of the symmetry of the standard model is that it classifies the quarks and leptons into several unrelated multiplets, even within one family. If we allow for the right-handed neutrino $N$, needed to give a smooth theory of neutrino masses, there are six (if not, five). Moreover, the $U(1)_Y$ hypercharge quantum numbers we need to assign to those multiplets are funny fractions, determined phenomenologically.

As we will see, the extended symmetries—especially $SO(10)$—do a brilliant job of organizing the fermion multiplets into a single one, and also of explaining the funny fractions.

Although I will not develop it here, I should mention that one can constrain the hypercharges in an alternative way, by demanding anomaly cancellation. That approach does not address the unification of couplings. Nor does it explain the multiplet structure nearly so neatly: in particular, it does not predict the existence of the right-handed neutrino $N$, which plays a central role in the theory of neutrino masses.

## (a) Standard model multiplets

To get started, let us briefly review the multiplet structure of quarks and leptons in SU(3) × SU(2) × $U(1)$. To keep things focused, I will pretend there is just one family and use the notations appropriate to the lightest quarks and leptons.

Here is what we have got, in the usual presentation of the standard model:

$$Q_\mathrm{L} = \begin{pmatrix} u_\mathrm{L} & u_\mathrm{L} & u_\mathrm{L} \\ d_\mathrm{L} & d_\mathrm{L} & d_\mathrm{L} \end{pmatrix}^{1/6}, \tag{3.1}$$

$$u_\mathrm{R} = \begin{pmatrix} u_\mathrm{R} & u_\mathrm{R} & u_\mathrm{R} \end{pmatrix}^{2/3}, \tag{3.2}$$

$$d_\mathrm{R} = \begin{pmatrix} d_\mathrm{R} & d_\mathrm{R} & d_\mathrm{R} \end{pmatrix}^{-1/3}, \tag{3.3}$$

$$L_\mathrm{L} = \begin{pmatrix} \nu_\mathrm{L} \\ e_\mathrm{L} \end{pmatrix}^{-1/2} \tag{3.4}$$

and

$$e_\mathrm{R} = (e_\mathrm{R})^{-1}. \tag{3.5}$$

Here:

- — The L and R subscripts denote left and right helicity, respectively.
- — The threefold repetition of quark fields indicates three colours. I dispensed with distinguishing indices. I also let the SU(3) act horizontally on the fundamental **3** representation, for typographical clarity (whereas SU(2) acts vertically).
- — The numbers attached to the multiplets are their $U(1)_Y$ hypercharges. These are given by the average electric charge of the particles in the multiplet, up to an overall normalization.

We can also write this in a less picturesque but more flexible way, using indices:

$$Q_\mathrm{L} \equiv Q^{\alpha a}_{L(1/6),} \tag{3.6}$$

$$u_\mathrm{R} \equiv u^{\beta}_{R(2/3),} \tag{3.7}$$

$$d_\mathrm{R} \equiv d^{\gamma}_{R-1/3,} \tag{3.8}$$

$$L_\mathrm{L} \equiv L^{d}_{L-1/2} \tag{3.9}$$

and

$$e_\mathrm{R} \equiv e_{\mathrm{R}-1}. \tag{3.10}$$

Here the Latin indices run from 1 to 2, and denote SU(2) quantum numbers, while the Greek indices run from 1 to 3, and denote SU(3) quantum numbers (colours).

For purposes of unification, we will want to have fermions all of the same helicity. We can do that by going to anti-particles. In this way, we arrive at

$$Q \equiv Q^{\alpha a}_{1/6}, \tag{3.11}$$

$$\bar{u} \equiv \bar{u}_{\beta -2/3}, \tag{3.12}$$

$$\bar{d} \equiv \bar{d}_{\gamma (1/3)}, \tag{3.13}$$

$$L \equiv L^{d}_{-1/2} \tag{3.14}$$

and

$$\bar{e} \equiv \bar{e}_1. \tag{3.15}$$

Here all the fields are understood to have left-handed helicity. $\bar{u}$ and $\bar{d}$ are anti-triplets $\bar{\mathbf{3}}$ of SU(3).

## (b) *N*

The theory of neutrino masses works most smoothly if we include an additional SU(3) × SU(2) × $U(1)$ singlet field $N$. It (or rather, to be more precise, $\bar{N}$) can be considered as the right-handed neutrino.

## (c) Unification in SU(5) [11]

It is easy to visualize how the action of SU(5) can incorporate SU(3) × SU(2) × $U(1)$. Indeed, a $5 \times 5$ matrix contains $3 \times 3$ and $2 \times 2$ blocks as follows:

$$\begin{pmatrix} *** & *** & *** & 0 & 0 \\ ***** & *** & *** & 0 & 0 \\ ***** & *** & *** & 0 & 0 \\ 0 & 0 & 0 & ** & ** \\ 0 & 0 & 0 & ** & ** \end{pmatrix}. \tag{3.16}$$

Now we need to implement the '$S$' part of SU($n$)—getting unit determinants. Within SU(5), but not within SU(3) × SU(2), we have the $U(1)$ subgroup

$$\begin{pmatrix} \mathrm{e}^{\mathrm{i}2\theta} & 0 & 0 & 0 & 0 \\ 0 & \mathrm{e}^{\mathrm{i}2\theta} & 0 & 0 & 0 \\ 0 & 0 & \mathrm{e}^{\mathrm{i}2\theta} & 0 & 0 \\ 0 & 0 & 0 & \mathrm{e}^{-\mathrm{i}3\theta} & 0 \\ 0 & 0 & 0 & 0 & \mathrm{e}^{-\mathrm{i}3\theta} \end{pmatrix}. \tag{3.17}$$

This identifies, essentially uniquely, SU(3) × SU(2) × $U(1)$ as a subgroup of SU(5).

Now let us see how the fermions might fit in this picture. There are not an enormous number of them—15 or 16, depending on whether you include $N$—relative to 5. So we must look to low-dimensional representations, specifically to vectors ($\mathbf{5}, \bar{\mathbf{5}}$) or two-index antisymmetric tensors ($\mathbf{10}, \bar{\mathbf{10}}$). In making the initial comparison, we will allow the normalization of the hypercharge to float, using

$$Y = c \begin{pmatrix} -2 & 0 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix} \tag{3.18}$$

with $c$ a free parameter. The value of $c$ is an important element in the quantitative aspect of unification, as we will see.

A little experimentation reveals that we get the right representation and charge spectrum by choosing the representations $\bar{\mathbf{5}} \oplus \mathbf{10}$ and $c = \frac{1}{6}$. Let us spell that out.

First, a simple but vital observation: in SU(2), doublets and anti-doublets are equivalent. Indeed, we can trade one for the other using the (invariant) antisymmetric symbol $\epsilon$, in the style

$$\left.\begin{aligned} & r^a \leftrightarrow \epsilon^{ab} r_b \\ \text{and} \quad & r_b \leftrightarrow -\epsilon_{bc} r^c \end{aligned}\right\} \tag{3.19}$$

The first three components of $\bar{\mathbf{5}}$ make a colour anti-triplet weak singlet, and the hypercharge is $-(-2)c = \frac{1}{3}$. That matches the quantum numbers of $\bar{d}$. The last two components make a colour singlet weak doublet with hypercharge $-3c = -\frac{1}{2}$. (Recall that SU(2) doublets and anti-doublets are equivalent.) That matches the quantum numbers of the lepton multiplet $L$.

The components of **10** with two early indices—that is, $T^{ab}$ with $a, b = 1, 2, 3$—make a colour *anti-triplet* weak singlet with hypercharge $((-2) + (-2))c = -\frac{2}{3}$. That matches the quantum numbers of $\bar{u}$. The components with one early and one late index make a colour triplet weak doublet with hypercharge $\frac{1}{6}$. That matches the quantum numbers of the quark multiplet $Q$. Finally, the component $T^{45}$ with two late indices defines a colour singlet weak singlet with hypercharge 1. That matches the quantum numbers of $\bar{e}$.

Thus all the quarks and leptons have been accommodated neatly, with no loose ends, and the mishmash of funny hypercharges has been rationalized into the one number $c$, with the consistent value

$$c = \frac{1}{6}, \tag{3.20}$$

$N$ can be brought in as an SU(5) singlet, since it is an SU(3) × SU(2) × $U(1)$ singlet, with hypercharge 0.

## (d) Unification in $SO(10)$ [11]

The spinor representation is central to $SO(10)$ unification, so let me recall (or reveal) how that works.

It is convenient to begin with the Clifford algebra

$$\{\gamma_j, \gamma_k\} = 2\delta_{jk}, \tag{3.21}$$

where $j, k$ run from 1 to 10. This algebra can be realized in a useful form, very familiar to physicists, by defining

$$\left.\begin{aligned} & a_j = \tfrac{1}{2}(\gamma_{2j-1} - \mathrm{i}\gamma_{2j}) \\ \text{and} \quad & a_j^\dagger = \tfrac{1}{2}(\gamma_{2j-1} + \mathrm{i}\gamma_{2j}), \end{aligned}\right\} \tag{3.22}$$

for the $a$s behave as a set of five fermion creation and destruction operators:

$$\left.\begin{aligned} & \{a_j, a_k^\dagger\} = 2\delta_{jk}, \\ & \{a_j, a_k\} = 0 \\ \text{and} \quad & \{a_j^\dagger, a_k^\dagger\} = 0. \end{aligned}\right\} \tag{3.23}$$

We can of course also invert the definition equation (3.22) to get the Clifford algebra from fermions.

The connection of the Clifford algebra to SO(10) is that the commutators

$$R_{jk} = \tfrac{1}{4}[\gamma_j, \gamma_k] \tag{3.24}$$

satisfy, by virtue of the Clifford algebra, the commutation relations of infinitesimal rotations. If we define $R_{jk}$ to represent an infinitesimal rotation in the $jk$ plane, we get the same Lie algebra as we get from equation (3.24). I will leave the detailed verification as an exercise, but here only check the corresponding logic for SO(3). In that case, we can realize the Clifford algebra using

Pauli spin matrices

$$\gamma_j = \sigma_j. \tag{3.25}$$

Our recipe gives, for example,

$$R_{12} = \frac{1}{4}[\sigma_1, \sigma_2] = \mathrm{i}\frac{\sigma_3}{2}, \tag{3.26}$$

which indeed is the standard generator of rotations in the 12 plane—i.e. around the three axis—in the theory of spin $\frac{1}{2}$.

Going back to the fermion representation, we see that the SO(10) generators generally contain pieces of the type $a_j a_k$ and their Hermitean conjugates that change fermion number by two units. On the other hand, these cancel in the rotations involving $a$s with a fixed index, i.e. $R_{12}, R_{34}, R_{56}, R_{78}, R_{9(10)}$. Indeed, for example,

$$R_{12} = \mathrm{i}(a_1^\dagger a_1 - \tfrac{1}{2}), \tag{3.27}$$

so these essentially count the occupancy of the different fermion states.

More generally, we find that the number-changing terms cancel in combinations like

$$R_{13} + R_{24} \propto [a_1 + a_1^\dagger, a_2 + a_2^\dagger] - [a_1 - a_1^\dagger, a_2 - a_2^\dagger] \tag{3.28}$$

and $R_{14} - R_{23}$, to give us the possibility of exchanging fermions of types 1 and 2. In this way, we generate the full $U(5)$ symmetry of five distinguishable but equivalent free elementary fermions. The $U(1)$, which simply counts total number, is of course essentially $R_{12} + R_{34} + R_{56} + R_{78} + R_{9(10)}$. We can, alternatively, define our $U(5)$ as the centralizer of this transformation, i.e. the number-conserving transformations within SO(10).

The minimal realization of fermions gives us a $2^5 = 32$ dimensional state space on which SO(10) acts. That action is not quite irreducible, however, because SO(10) conserves fermion number modulo 2. So we have two invariant 16-dimensional state spaces wherein the fermion number is $(0, 2, 4)$ or $(1, 3, 5)$, respectively. These turn out to be irreducible (and inequivalent). For unification purposes, with our realization of SU(5) as transformations among elementary fermions, we want the former. Indeed, since the fermion label transforms as a vector under SU(5), and the states are antisymmetric in that label (fermionic), $(0, 2, 4)$ will give us a singlet, an antisymmetric 2-index tensor **10**, and an anti-vector $\bar{\mathbf{5}}$. As we saw previously, that is just what we need to define the observed quarks and leptons.

In this construction, all the directly observed fermions are combined into a single irreducible representation, together with $N$. $N$ has been promoted from a desirable option to a necessary feature. Now it forms part of the operating system.

## 4. Visual summary

Figure 1 captures the main result of the preceding discussion in a striking, compact visual form. The black column of symbols on the left-hand panel shows up (u) and down (d) quarks, the electron (e) and the electron neutrino ($\nu$). Each of those particles can have a left-handed (L) or right-handed (R) helicity; the left-handed particles are organized in doublets, so the column displays six distinct entities. The left side of the panel displays the properties of those entities that account for their strong, weak and electromagnetic interactions. Gluons, which mediate the strong interaction, respond to three strong 'colour' charges, here indicated by red, green and blue. The weak interactions, which only act upon left-handed particles, respond to two weak colour charges, here indicated as yellow and purple. Electromagnetism is incorporated through couplings to electric charge. On the left side of the panel, the average charge within each entity, also called its hypercharge ($Y$), is indicated by a numerical subscript.

The right-hand panel shows how the observed, scattered pattern of particles can be deduced from a unified template, derived from higher symmetry. The far-right column of the panel again gives the names of the particles. Here, everything has left-handed helicity; right-handed particles in the earlier description are now represented through their left-handed anti-particles, so that $u^{\mathrm{R}}$ becomes $-u$ and so forth. The far-left table shows all possible assignments of full or empty

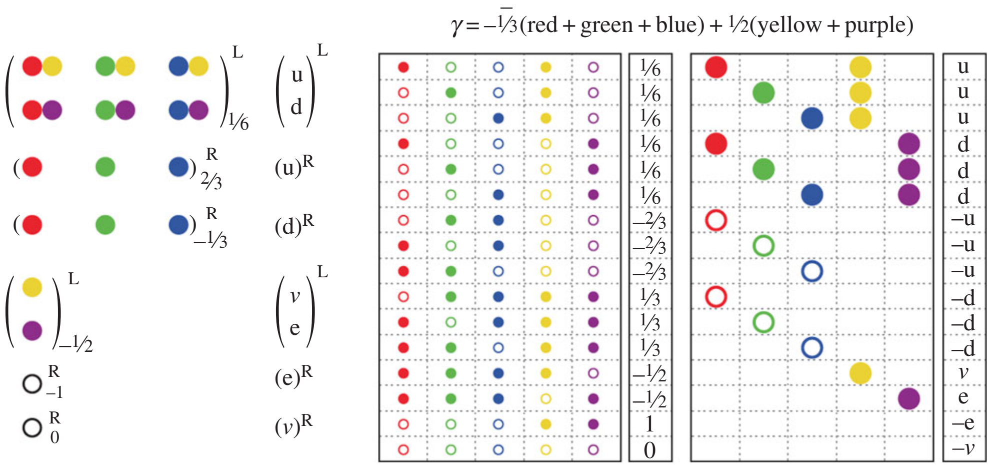

**Figure 1.** This figure, explained in detail in the text, shows how the rather untidy organization of fundamental particles in our present day core theories of the strong, weak and electromagnetic interactions falls neatly into, and can be explained by, a highly symmetric unified theory.

circles of strong and weak colours, subject to the constraint that the number of full circles is even. A full circle is interpreted as a positive half-unit of the corresponding charge, an empty circle as a negative half-unit. The hypercharge values, which appear in the middle column, are now generated from the weak and strong colour charges according to the formula on top.

To recover the properties displayed in the left panel, I invoke the rule that adding equal amounts of all the strong colour charges, or equal amounts of all the weak colour charges, does not influence the strong or weak interactions. That allows me to change the colour scheme on the left table into that on the right table. Note that the colour charges are now in full units, and—remarkably—that the colours and hypercharges match the content of the right-hand panel, given that anti-particles have the opposite colour charges and hypercharges from their corresponding particles.

## 5. Unification of coupling strengths

In the preceding section, we have seen how the messy multiplet structure of quarks and leptons in the standard model, including their hypercharges, comes to look much nicer when viewed within the context of SU(5) or especially SO(10) unification.

Those higher symmetries realize the separate SU(3) × SU(2) × $U(1)$ gauge symmetries of the standard model as different aspects of encompassing symmetry. The glory of local (gauge) symmetry, however, is that it controls not only bookkeeping but also dynamics. For a simple (in the technical sense) gauge group such as SU(5) or SO(10), symmetry dictates all the couplings of the gauge bosons, up to a single overall coupling constant.

Thus, unification predicts relationships among the strong, weak and hypercharge couplings. Basically—up to the group-theoretic task of normalization—it predicts that the three couplings for SU(3) × SU(2) × $U(1)$ must be equal. As observed, of course, they are not. But the two great dynamical lessons of the standard model—namely symmetry breaking through field condensation (Higgs mechanism), and running of couplings (asymptotic freedom)—suggest a way out [12]. We can imagine that the symmetry breaking $G \to \mathrm{SU}(3) \times \mathrm{SU}(2) \times U(1)$ occurs through a big condensation, at a high mass scale. In the symmetric theory, appropriate to the description of processes at large mass scales, there was only one unified coupling. But we make our observations at a much lower mass scale. To get to the unified coupling, we must evolve the observed couplings up to high energy, taking into account vacuum polarization. Note that throughout that evolution the unified symmetry is violated, so the three SU(3) × SU(2) × $U(1)$ couplings evolve differently.

Before entering the details, let us pause to consider a soft-focus view of what we can expect from this sort of calculation. Our input will be the observed couplings, plus some hypothesis $\mathcal{H}$ about the spectrum of virtual particles we need to include in the vacuum polarization. Our output should be the unified coupling strength, and the scale of unification. For any given $\mathcal{H}$, we have three inputs—the observed couplings—and two outputs—the scale and coupling at unification. So there will be a consistency condition. If the calculation works, we will have reduced the number of free parameters in the core of the standard model by one, from three to two.

There are additional *physical* consistency conditions, concerning the value of the unification scale, which are quite significant. Also, in case of success, we will need to discuss the plausibility of our hypothesis $\mathcal{H}$. We will return to these important points later, after we have done the central calculation.

## (a) Normalization of hypercharge

The value of non-abelian couplings can be fixed absolutely, because the generators obey nonlinear commutation relations. It is common practice, for SU($n$) or SO($n$) groups, to choose the coupling constant to multiply, in the fundamental representation, generators the trace of whose square $\frac{1}{2}$. Thus, for SU(2) we have the covariant derivative

$$\nabla_\mu = \partial_\mu + \mathrm{i}\, g_2 \frac{\sigma_a}{2} A^a_\mu \tag{5.1}$$

for isospinors and so forth. (Of course, we understand here that the $A^a$ appear in a Maxwell-like (Yang–Mills) action, in such a way that the plane wave 'generalized photons' are canonically normalized.)

That fixes the normalizations for $g_3$ and $g_2$, the couplings associated with SU(3) and SU(2), respectively, and also the normalization of $g_5$ in SU(5).[2]

The normalization of the hypercharge generator in SU(5) is also thereby fixed, but its numerical relationship to the conventional normalization of the hypercharge of $U(1)_Y$, as it appears in electroweak phenomenology, requires discussion.

We have seen that in SU(5) the fundamental representation (actually its conjugate, but that makes no difference here) is implemented on $(\bar{d}, \bar{d}, \bar{d}, L)$. The trace of the square of the hypercharge generator, times the square of the coupling constant, acting on this is therefore $g_5^2/2$. On the other hand, if we evaluate the same thing in the standard electroweak notation, we get

$$(g')^2(3 \times (\tfrac{1}{3})^2 + 2 \times (\tfrac{1}{2})^2) = (g')^2 \times \tfrac{5}{6}. \tag{5.2}$$

Equating these two evaluations of the same thing, we have

$$\left.\begin{aligned} & g_5^2 \times \tfrac{1}{2} = (g')^2 \times \tfrac{5}{6} \\ \text{and} \qquad & g_5^2 \times \tfrac{3}{5} = (g')^2. \end{aligned}\right\} \tag{5.3}$$

It will be convenient, for later purposes, to express this result as the definition

$$g_1^2 \equiv \tfrac{5}{3}(g')^2. \tag{5.4}$$

In the unified theory, it is $g_1$ that should evolve to become equal to $g_2$ and $g_3$ (and $g_5$).

Before including any running of the couplings, we get the two 'predictions'

$$\left.\begin{aligned} & g_3^2 = g_2^2 (= g_5^2) \\ \text{and} \qquad & \sin^2\theta_W \equiv \frac{(g')^2}{g_2^2 + (g')^2} = \frac{3/5}{1+3/5} = \frac{3}{8}. \end{aligned}\right\} \tag{5.5}$$

They are way off.

[2] There is no need, here, to discuss SO(10) separately.

## (b) Structure of coupling renormalization

Each of the couplings is affected by vacuum polarization. Thus, the values observed in an experiment will depend on the distance, or equivalently the energy and momentum, characteristic of the measurement process. The logarithmically divergent (before regularization) terms are proportional, in lowest order, to the cube of the coupling.[3] So we find equations for the running couplings of the form

$$\frac{\mathrm{d}\bar{g}_j(Q)}{\mathrm{d}\ln Q} \approx b_j \bar{g}_j^3, \tag{5.6}$$

where $b$ is a number that depends on the spectrum of virtual particles that contribute.

To expose the logic of coupling unification, it is helpful to re-write equation (5.6) as

$$\frac{\mathrm{d}1/\bar{g}_j^2}{\mathrm{d}\ln Q} = -\frac{2}{\bar{g}_j^3}\frac{\mathrm{d}\bar{g}_j(Q)}{\mathrm{d}\ln Q} = -2b_j \tag{5.7}$$

with the solution

$$\frac{1}{\bar{g}_j(Q)^2} = -2b_j \ln\frac{Q}{Q_0} + \frac{1}{\bar{g}_j(Q_0)^2}. \tag{5.8}$$

Here we take $Q_0$ to be an accessible laboratory scale, where we do empirical measurements of the $\bar{g}_j(Q)^2$.

Now the unification condition is that for some $Q$ the effective couplings $\bar{g}_3^2(Q), \bar{g}_2^2(Q), \bar{g}_2^2(Q)$ become equal to a common value, call it $g_5(Q)$. By subtracting the solutions equation (5.8), for $j = 2, 3$, we derive an equation determining the unification scale:

$$2(b_3 - b_2)\ln\frac{Q}{Q_0} = \frac{1}{\bar{g}_3(Q_0)^2} - \frac{1}{\bar{g}_2(Q_0)^2}. \tag{5.9}$$

Since we must derive the same scale from other pairs of couplings, we have the consistency condition

$$\frac{b_3 - b_2}{b_2 - b_1} = \frac{1/\bar{g}_3(Q_0)^2 - 1/\bar{g}_2(Q_0)^2}{1/\bar{g}_2(Q_0)^2 - 1/\bar{g}_1(Q_0)^2}. \tag{5.10}$$

This is the anticipated prediction, supplying a numerical relation among the observed couplings.

Of course, once we have the unification scale, we can go back to determine the value of the unified coupling using equation (5.8).

## (c) Numerics of couplings renormalization

### (i) Renormalization group coefficients

The renormalization group coefficients $b_j$ can be calculated perturbatively, for any combination of spin 0, $\frac{1}{2}$ and 1 (gauge) fields. The result is [13,14]

$$16\pi^2 b = -\tfrac{11}{3}C_A + \tfrac{4}{3}T(R_{1/2}) + \tfrac{2}{3}T(R_0). \tag{5.11}$$

Here $C_{\mathrm{A}}$ is the value of the Casimir operator for the adjoint representation of the gauge group in question. Explicitly, we have

$$C_{\mathrm{A}}(\mathrm{SU}(n)) = n. \tag{5.12}$$

$T(R_{1/2})$ is the trace of the square of a normalized generator acting on the spin-$\frac{1}{2}$ fields in the theory, which may of course include several multiplets. We will only need the basic defining normalization $T = \frac{1}{2}$ for the fundamental representation, and, when we come to consider supersymmetry, $T = C = n$, for the adjoint of SU($n$). The coefficient $\frac{4}{3}$ for fermions holds for complex Dirac fermions. For Weyl fermions (i.e. fermion fields with definite helicity), we get half

[3] This emerges most clearly if we consider the gluon self-coupling. When we consider couplings of gluons to fermions, there appears to be cross-talk between the different interactions. These cross terms turn out not to contribute, due to Ward's identity, which is a diagrammatic manifestation of gauge symmetry. They had better not, because the renormalized non-abelian coupling should be universal—the same for gluons, fermions, scalars, ghosts, ....

that, as we do for Majorana (real) fermions. $T(R_0)$, unsurprisingly, is the trace of the square of a normalized generator acting on the spin-0 fields in the theory, which may of course include several multiplets. The coefficient $\frac{1}{3}$ holds for complex scalars; for real scalars, we would get half that (but that case will not arise below).

### (ii) Minimal extrapolation

Taking the particles of the standard model, but allowing (why not?) for $n_f$ families and $n_s$ Higgs doublets, we have

$$16\pi^2 b_3 = -11 + \tfrac{4}{3}n_f, \tag{5.13}$$

since there are two fundamentals of Dirac fermions per family. Similarly,

$$16\pi^2 b_2 = -\tfrac{22}{3} + \tfrac{4}{3}n_f + \tfrac{1}{3}n_s \tag{5.14}$$

(four fundamentals of Weyl fermions!) and, finally,

$$16\pi^2 b_1 = \tfrac{4}{3}n_f + \tfrac{1}{5}n_s. \tag{5.15}$$

The $b_1$ equation can be obtained painlessly by noting that the contribution of complete families must respect the SU(5) symmetry—that remark governs the fermions directly, while the Higgs particle should be fleshed out with a colour triplet that is not there, so its contribution is reduced by the factor

$$\frac{2 \times (1/2)^2}{2 \times (1/2)^2 + 3 \times (1/3)^2} = \frac{3}{5}. \tag{5.16}$$

If we run the couplings using these coefficients, with $n_f = 3$, $n_s = 1$, we get an unsatisfactory result [15].

### (iii) Extrapolation with supersymmetry [16]

Supersymmetry is too big, and too technical, a subject to develop here for a general audience of physicists. I will only insert two brief, broad motivating comments, which I find irresistible.

- — Wave-particle duality blurred the contrast between force, epitomized by electromagnetism and emergent light, and substance, epitomized by electrons. At the level of single-particle quantum mechanics, photons and electrons fall comfortably into a single framework. This changes, however, at the level of multi-particle quantum mechanics, where the contrasting quantum statistics of force (bosons) and substance (fermions) marks a sharp dichotomy of matter. Supersymmetry, by allowing transformations that exchange bosons and fermions, re-establishes unity.
- — Our unification, described above, brings together all the forces, on the one hand, and all the substances, on the other.[4] But even were it consummated, it would leave us with two separate things: force and substance. Supersymmetry bring those together.

To implement low-energy supersymmetry, in a minimal fashion, we must expand the standard model in several ways:

1. We have spin-$\frac{1}{2}$ Majorana fermion partners of the gauge fields, in the adjoint representation. Grouping their contribution with the gluons, has the effect of changing the $-\frac{11}{3}$ in equation (5.11) to $-3$.
2. For each chiral fermion in the theory, we get a complex scalar superpartner. Grouping their contribution with the fermions, this has the effect of changing the $\frac{4}{3}$ in equation (5.11) to 2—still understanding, of course, that we halve this for chiral fermions!
3. Conversely, for each ordinary Higgs doublet, we need a chiral fermion with the same quantum numbers. This changes the $\frac{1}{3}$ in equation (5.11) to 1.
4. Supersymmetry requires, at a minimum, $n_s = 2$ Higgs doublets.

[4] Close readers will appreciate that I have allowed myself some poetic licence here.

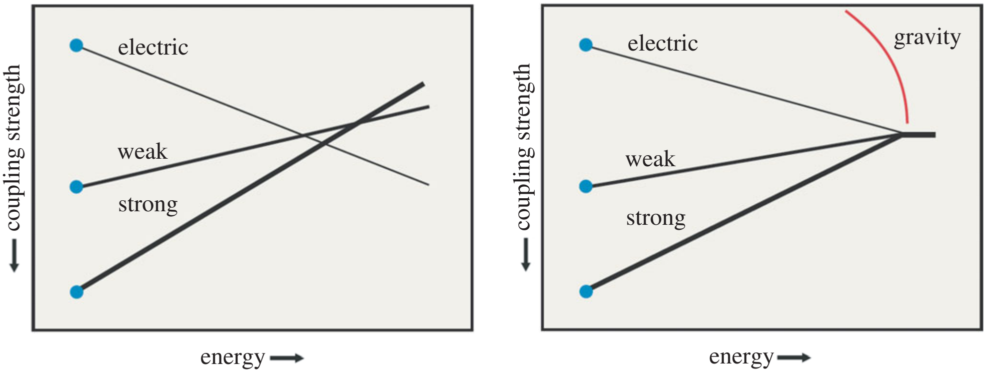

**Figure 2.** The equality of coupling strengths at high energy, suggested by theories of unification, is suggested, but not achieved, by minimal extension of known interactions (left panel). With the additional hypothesis of supersymmetry, one achieves good quantitative agreement. For more detailed discussion, see the main text. (Online version in colour.)

Putting all this together, we find that when the contribution of virtual particles required by supersymmetry is included, we have

$$\left.\begin{aligned} 16\pi^2 b_3 &= -9 + 2n_f, \\ 16\pi^2 b_2 &= -6 + 2n_f + \tfrac{1}{2}n_s \\ \text{and}\qquad 16\pi^2 b_1 &= 0 + 2n_f + \tfrac{3}{10}n_s. \end{aligned}\right\} \tag{5.17}$$

If we run the couplings using these coefficients, with $n_f = 3$, $n_s = 2$, we get a much more satisfactory result [15]. The essence of the matter is that our predictive equation, equation (5.10), is now well satisfied. The unification scale is computed to be $\approx 2 \times 10^{16}$ GeV.

As a measure of the delicacy and resolving power of the calculation, let us note that taking $n_s = 4$ leads to a 15% error in the prediction of the Weinberg angle, if we use strong–weak unification to fix the scale.

### (iv) Visual summary

The preceding discussion of coupling unification has been long and somewhat technical, as befits an issue where quantitative accuracy is both possible and important. It may be helpful, therefore, to capture the essential structure and conclusion in a simple, memorable image (figure 2).

A central consequence of enhanced gauge symmetry is that the basic strengths of the different interactions are equal. The observed strengths are not equal, but we understand that the basic strengths are revealed only at very high energies. To get at them, we must correct for virtual particles, which can screen or anti-screen an interaction—an effect known as vacuum polarization. When we allow for the contributions of all the known particles we find a suggestive approach to equality but a quantitative failure, as illustrated in the panel to the left. But as the right-hand panel shows, when we include the additional contributions from superpartners, as suggested by supersymmetry, we find accurate unification.

Gravity is much weaker than the other forces when it acts between particles at accessible energies. But at the energy where the other forces are unified, its strength is comparable. Viewed from another angle, the hypothesis that all the basic forces are unified at the scale indicated by gauge coupling unification quantitatively explains the observed feebleness of gravity.

### (v) Observations on the generality

*An important general observation*: to the order we have been working, complete SU(5) multiplets affect neither the predicted relation among observed couplings, nor the predicted scale of unification [16]! That striking conclusion follows because complete multiplets contribute equally to $b_1, b_2$ and $b_3$, and in equations (5.10), (5.9) only differences among those coefficients occur. So our successful 'minimal supersymmetry' hypothesis is not so special as might appear at first sight.

For example, one need not postulate a complete desert (apart from supersymmetry) in the mass spectrum between current observations and the unification scale. One can populate it with any number of singlets, or with a modest number of complete families. It is only broken families that are worrisome. We can also allow, within supersymmetry, the masses of squarks and sleptons—the partners of quarks and leptons—to float up to a high scale together, since they form complete SU(5) multiplets. In principle, there could even be different large masses for the different families of superpartners.

The partners of gauge bosons, on the other hand, *do not* form a complete SU(5) multiplet. Their masses cannot be allowed to float very high before ruining the success of our calculation. A recent estimate [17] suggests $M \sim 2\,\mathrm{TeV}$ is the preferred scale.

Raising the squark and slepton masses is an attractive option phenomenologically, because it relieves difficulties with proton decay and flavor violation—processes that otherwise tend to be over-predicted in supersymmetric models. These possibilities have been explored in speculative phenomenology, first under the epithet 'focus point' and more recently also as 'split' and 'mini-split' supersymmetry. (On the other hand, we may not want to raise those masses $>10\,\mathrm{TeV}$, as this leads to difficulties with another attractive consequence of unified theories, namely their quantitative explanation of the mass ratio $m_b/m_\tau$ [15].)

Effects of complete multiplets will show up in more accurate calculations, taken to higher order. They also affect the value of the unified coupling. By increasing the value of $b_j$ in equation (5.8), they make the unified coupling larger. Therefore, we probably do not want to have too many of them.

### (d) Loose ends

Above I have outlined the lowest order calculation. There is a vast technical literature on corrections, both those due to additional couplings and those due to masses [15]. I think it is fair to say that the situation is generally satisfactory, although unfortunately there are many more potentially relevant parameters than experimental constraints, when one attempts precision work.

The Higgs doublet of the standard model, or the two Higgs doublets of its supersymmetric extension, do not fill out unified multiplets. Indeed, there are powerful bounds on the mass of the possible colour triplet partners, since they make it difficult to maintain baryon number conservation as a good approximation. There are several ideas to address this doublet–triplet splitting embarrassment, but no consensus on which (if any) is correct.

# 6. Prospect

## (a) Significance of the scale

### (i) Relation to Planck scale

The Planck energy

$$\mathcal{E}_{\text{Planck}} = \sqrt{\frac{\hbar c^5}{8\pi G_N}} \approx 2.4 \times 10^{18}\,\mathrm{GeV} \tag{6.1}$$

is another famous energy scale that can be constructed from fundamental constants.[5] Here the construction is simple dimensional analysis, based on Newton's gravitational constant $\mathrm{G_N}$ together with $\hbar, c$. On the face of it, Planck units set the scale for effects of quantum gravity; thus

[5] I have quoted the so-called rationalized Planck scale, including the factor $8\pi$ that naturally appears with $G_N$ in the Lagrangian of general relativity.

when we consider basic (technically: 'hard') processes whose typical energies are of order $E$, we expect gravitational effects of order $(E/\mathcal{E}_{\text{Planck}})^2$.

Our scale $\mathcal{E}_{\text{unification}} \approx 2 \times 10^{16}$ GeV is significantly, but not grotesquely, smaller than the Planck scale. This means that at the unification scale the strength of gravity, heuristically and semi-quantitatively, is of order

$$\left(\frac{\mathcal{E}_{\text{unification}}}{\mathcal{E}_{\text{Planck}}}\right)^2 \sim 10^{-4} \tag{6.2}$$

to be compared with the strength $g_5^2/4\pi \sim 10^{-2}$ for the other interactions. The relative smallness of gravity, thus estimated, which is of course further accentuated at lower energies, suggests that our neglect of quantum gravity in the preceding calculations may be justified.

On the other hand, it seems to me remarkable that the comparison comes so close. A classic challenge in fundamental physics is to understand the grotesque smallness of the observed force of gravity, compared to other interactions, as it operates between fundamental particles. Famously, the gravitational interaction is approximately $10^{42}$ times smaller than any of the other forces. Again, however, proper comparison requires that we specify the energy scale at which the comparison is made. Since the strength of gravity, in general relativity, depends on energy directly, it appears hugely enhanced when observed with high-energy probes. At the scale of unification $\mathcal{E}_{\text{unified}} \sim 2 \times 10^{16}$ GeV, the discrepant factor $10^{42}$ is reduced to approximately $10^4$, or even a bit less. While this does not meet the challenge fully, it is a big step in the right direction.

### (ii) Neutrino masses and proton decay

By expanding our theory, unification along the lines we have been discussing brings in additional interactions. Since the unified multiplets combine particles that normally (i.e. with the standard model itself!) do not transform into one another, we find new processes of transformation. The two classic predictions for 'beyond the standard model' interactions are small neutrino masses, leading to neutrino oscillations, and proton decay. The first has been been vindicated; the second not yet. In both cases, the large-scale $\mathcal{E}_{\text{unification}}$ is crucial for explaining the smallness of the new effects. For an authoritative review of these and other aspects of unification, emphasizing the phenomenological issues, with many further references, see [15].

## (b) Conclusion

The unification of quark and lepton quantum numbers in SU(5), and especially SO(10), is smooth and strikingly beautiful. The unification of coupling strengths fails quantitatively if one makes a minimal extrapolation of the standard model, but under the hypothesis of low-energy supersymmetry its success is likewise smooth and strikingly beautiful. The unification of coupling strengths brings a new scale into physics, which has several attractive features including, notably, a big step towards unification with gravity. The discovery (or not) of some superpartners at the LHC will bring this line of thought to fulfillment (or not). If it does, we shall have not only unified the different forces with one another, and the different substances with one another—but also unified force and substance, fulfilling Maxwell's vision.

Competing interests. I declare that I have no competing interests.
Funding. This work is supported by the US Department of Energy under contract no. DE-FG02-05ER41360.

## References

1. Maxwell JC. 1891 *A treatise on electricity and magnetism*, 3rd edn. Oxford, UK: Clarendon Press.
2. Dirac PAM. 1927 The quantum theory of the emission and absorption of radiation. *Proc. R. Soc. Lond. A* **114**, 243–265. (doi:10.1098/rspa.1927.0039)
3. Dirac PAM. 1931 Quantised singularities in the electromagnetic field. *Proc. R. Soc. Lond. A* **133**, 60–72. (doi:10.1098/rspa.1931.0130)

4. ’tHooft G. 1974 Magnetic monopoles in unified gauge theories. *Nucl. Phys. B* **79**, 276–284. (doi:10.1016/0550-3213(74)90486-6)
5. Polyakov A. 1974 Particle spectrum in quantum field theory. *JETP Lett.* **20**, 194–195.
6. Einstein A. 1916 Die Grundlage der allgemeinen relativitätstheorie. *Ann. Phys.* **49**, 769–822. (doi:10.1002/andp.19163540702)
7. Weyl H. 1918 Gravitation and electricity. *Ann. Phys.* **59**, 101.
8. Yang CN, Mills RL. 1954 Conservation of isotopic spin and isotopic gauge invariance. *Phys. Rev.* **96**, 191–195. (doi:10.1103/PhysRev.96.191)
9. Batterham D. 2012 *The world of ornament*. Cologne, Germany: Taschen.
10. Weyl H. 1929 Elektron und gravitation. I. *Z. Phys.* **56**, 330–352. (doi:10.1007/BF01339504)
11. Georgi H, Glashow SL. 1974 Unity of all elementary-particle forces. *Phys. Rev. Lett.* **32**, 438–441. (doi:10.1103/PhysRevLett.32.438)
12. Georgi H, Quinn H, Weinberg S. 1974 Hierarchy of interactions in unified gauge theories. *Phys. Rev. Lett.* **33**, 451–454. (doi:10.1103/PhysRevLett.33.451)
13. Gross DJ, Wilczek F. 1973 Ultraviolet behavior of non-abelian gauge theories. *Phys. Rev. Lett.* **30**, 1343–1346. (doi:10.1103/PhysRevLett.30.1343)
14. Politzer HD. 1973 Reliable perturbative results for strong interactions? *Phys. Rev. Lett.* **30**, 1346–1349. (doi:10.1103/PhysRevLett.30.1346)
15. Raby S. 2015 Grand unified theories. See http://pdg.lbl.gov/2015/reviews/rpp2014-rev-guts.pdf.
16. Dimpoloulos S, Raby S, Wilczek F. 1981 Supersymmetry and the scale of unification. *Phys. Rev. D* **24**, 1681. (doi:10.1103/PhysRevD.24.1681)
17. Bourilkov D. 2015 Strong coupling running, gauge coupling unification and the scale of new physics. *J. High Energy Phys.* **2015**, 117. (http://arxiv.org/abs/1508.04176)