The Asymmetry Between Matter and Antimatter

Phase transitions and massive-neutrino decay are two processes that could lead to the preponderance of matter in the universe. Either way, the standard model for particle physics will have to be modified.

Helen R. Quinn

We live in a universe that is dominated by matter and contains very little antimatter. The laws of physics, however, include an almost exact symmetry between matter and antimatter. That symmetry is not the simple charge conjugation (C) that relates a particle and its corresponding antiparticle. Rather, it is CP, the product of C with coordinate inversion, or parity (P).

Before the Dirac equation was introduced in 1928, there was no concept of a symmetry relating matter and antimatter; indeed, antimatter had not been conceived. There was simply a conservation law, the conservation of matter. All the matter in the universe—all the stuff with mass—must always have been there. The prevailing view of Western science, and of Western religions and philosophy, was that we live in a static, unchanging universe. The constancy of matter in such a universe is no puzzle.

Hubble's law, the linear relationship between galactic redshift and distance, provided the first key evidence for an expanding universe. That discovery, made a year after the Dirac equation, initiated the field of cosmology. And with that new field came inevitable questions about the role of matter and of antimatter in cosmic evolution.

An entirely new kind of particle

If matter is conserved, its presence—even in an evolving universe—can only be understood as an initial condition of the universe's evolution. Beginning with the Dirac equation, though, physicists started down a path that led to radical changes in their ideas about matter conservation and the role of initial conditions.

The equation was the fruit of an effort of Paul Dirac (seen in figure 1) to obtain a relativistic equation of motion describing a particle, such as the electron, whose intrinsic spin is $\frac{1}{2}\hbar$. The equation had one very successful feature: It gave the correct magnetic moment for an electron. That was, Dirac said in 1977, "an unexpected bonus, completely unexpected."[1] But the equation contained an enigma, which first manifested itself as the existence of negative-energy states.

Such states are clearly unphysical. At best, they suggest one has misidentified the ground state of the theory. At worst, they say the theory is incurably sick, having no lowest-energy state. For electrons, the exclusion principle offers a cure. A better ground state can be found, in which all the negative-energy states are filled. Even then, Dirac's equation remains enigmatic. It has excitations that must be interpreted as positively charged particles (holes in the negative-energy sea) in addition to the negatively charged electron states. What possible physical interpretation could be made of those objects?

At the time of the Dirac equation, only two fundamental matter particles were known: electrons and protons. Understanding the nature of nuclei, particularly the concept of neutrons, was a work in process; evidence for neutrons was not deciphered until 1932. When the Dirac equation appeared, physicists did not readily postulate new particle types to explain new phenomena, let alone a peculiar result in a newly postulated equation. It seemed, thus, that the positively charged states must be protons. That interpretation had one obvious deficiency, and a second, fatal flaw that took a little longer to be noticed.

The obvious problem was that the equation required the proton and the electron to have the same mass. Indeed, all other properties of the positively and negatively charged states were the same. Dirac was aware of that invariance, as his 1929 letter to Niels Bohr makes clear. "As long as one neglects interactions," he wrote, "one has complete symmetry between electrons and protons. . . . However, when the interaction between the electrons is taken into account, this symmetry is spoilt. I have not yet worked out mathematically the consequences of the interaction. . . . One can hope, however, that a proper theory of this will enable one to calculate the ratio of the masses of protons and electrons."

I can only guess at what Dirac meant by "the interaction." Perhaps he was thinking of interactions between the particles filling the negative-energy sea (although we now understand that there are no such interactions). Whatever he was thinking, his wild hope that the symmetry could be removed by interactions was completely wrong.

Hermann Weyl (seen in figure 2) knew a symmetry and its consequences when he saw one. He stated, in November 1930, that "the mass of the proton should be the same as the mass of the electron; furthermore . . . this hypothesis leads to the essential equivalence of positive and negative electricity under all circumstances. . . . The dissimilarity of the two kinds of electricity thus seems to hide a secret of nature which lies deeper than the dissimilarity of past and future . . . a new crisis of quantum physics. . . ."

The crisis deepened when Robert Oppenheimer and Igor Tamm independently noticed the second problem. They saw that the Dirac equation allowed particle–hole annihilation. Their realization was the kiss of death for the proton interpretation of the holes. Although consideration

Helen Quinn *is a research scientist at SLAC in Menlo Park, California.*

Figure 1. Paul Dirac (1902–84). Dirac's puzzling relativistic equation of motion led him to suggest a new particle having the same mass as the electron but opposite electric charge. That new particle, the positron, was the first antiparticle to be observed. (Courtesy of the AIP Emilio Segrè Visual Archives, Fankuchen Collection.)

of interactions might possibly fix the mass problem, it could not eliminate annihilation, a disastrous process that destroys all possibility of stable matter. So, in May 1931, Dirac made what he later called a small step forward. He declared that "a hole, if there were one, would be an entirely new kind of particle, unknown to experimental physics, having the same mass, and opposite charge of the electron." The evasive phrase "if there were one" is interesting: It seems that the idea of predicting a new type of particle was still overwhelmingly bold to Dirac.

Pauli's objection to hole theory

Not much more than a year later, the particle envisioned by Dirac was observed. The observation of the positron raised a new question: Why is the world populated with electrons but not positrons? The cosmological implications of that question were recognized by Wolfgang Pauli (see figure 3), who wrote, in a remarkable June 1933 letter to Werner Heisenberg, "I do not believe in the hole theory, since I would like to have the asymmetry between positive and negative electricity in the laws of nature (it does not satisfy me to shift the empirically established asymmetry to one of the initial state)." As far as I know, Pauli's was the first statement of the view, held today by many particle physicists and cosmologists, that it is unsatisfactory to appeal to initial conditions to explain the dominance of matter over antimatter in the universe. Pauli also pointed out that, to avoid having to appeal to initial conditions, one must somehow remove the symmetry built into Dirac's equation between matter and antimatter.

Experiments soon added more particles. Mesons, discovered in 1947, eluded classification as either matter or antimatter—today we know them as equal mixtures of both, with a basic substructure of a quark and an antiquark. Antiprotons (discovered in 1955) and antineutrons (1957) were the first observed antimatter partners of baryons, particles whose basic substructure is now known to be three quarks. Antibaryons are made from three antiquarks.

The discovery of antimatter forced physicists to modify the law of conservation of matter. Annihilation and production of matter could occur, but only with annihilation or production of a matching amount of antimatter. The law of conservation of matter was replaced by new laws: the conservation of baryon number (the number of baryons minus the number of antibaryons) and the conservation of lepton number (the number of particles minus antiparticles of a given lepton type).

By choosing the term "antimatter" to describe the new particles, physicists changed the definition of the word "matter." One could no longer say, "Matter is all that has mass." Some 70 years after the discovery of positrons, most people in the US know only that old definition, and think antimatter just exists in science fiction. Publicity over recent results on the production of antihydrogen atoms may help change that perception (see PHYSICS TODAY, November 2002, page 17, and January 2003, page 14). But for a long time now, antimatter production has not been rare in our high-energy laboratories! Perhaps we should be teaching about antimatter in our schools.

Physicists considered baryon and lepton number conservation to be exact laws as late as the 1970s. And, Pauli's reservations notwithstanding, they assumed C was exact until well into the 1950s. Given those exact symmetries, the observed asymmetry of matter and antimatter in the universe could only be accounted for by an initial condition. Quantum electrodynamics, the direct descendant of the Dirac equation, does separately conserve C, P, and time-reversal invariance (T). Moreover, all local field theories conserve the product symmetry CPT. Most physicists of the early 1950s expected the symmetries conserved in QED to also be conserved in the weak interactions. In 1956, T.-D. Lee and C. N. Yang pointed out that there was no evidence for or against parity conservation in weak interactions. Within a year, experiments showed that weak interactions violated both C and P symmetry.

A modified charge conjugation

Although C and P symmetry were observed to be broken by the weak interactions, a revised matter–antimatter symmetry, CP, still appeared to be exact. Indeed, like Dirac's equation and QED, all known particle theories at the time had CP symmetry. The modern standard model would too, if it had only two generations of quarks and a single Higgs boson.[2] One needs quite a number of different particle types and thus many independent couplings before a field theory has room for CP violation.

In 1957, such a theory would have been wildly speculative. The Fermi theory of the weak interactions fit the observations and preserved CP (and T) symmetry. It did have an inconvenient feature though: Any calculation beyond lowest order gave divergent answers! Because weak interactions are rare processes, the lowest order theory worked quite satisfactorily, so the divergence embarrassment could be set aside for a while at least. As far as I know, no one made the leap that Pauli had earlier made, to the idea that a theory without CP symmetry would be preferable for cosmology, until after the 1964 experiment that demonstrated CP violation.

Figure 2. Hermann Weyl (1885–1955). Writing in 1930, around the time this photograph was taken, Weyl observed that the Dirac equation implied an exact symmetry between positive and negative electricity. Because the positron was not yet discovered, the symmetry was, for Weyl, a "crisis." (Courtesy of the AIP Emilio Segrè Visual Archives, Nina Courant Collection.)

The violation was discovered in neutral K meson decays. Two states of similar mass but very different halflife were generally assumed to be the CP-even and CP-odd combinations of the strange K^0 meson and its CP-conjugate antiparticle of opposite strangeness, the $\bar{K}^0$. If CP were exact, only the CP-even state $(K^0 + \bar{K}^0)/\sqrt{2}$ could decay to two pions. The CP-odd state $(K^0 - \bar{K}^0)/\sqrt{2}$ decays into three pions, but the rate is suppressed by the much smaller phase space available for the decay. CP symmetry thus neatly explained the large difference in halflife for the two neutral K mesons. In 1964, however, James Christenson, James Cronin, Val Fitch, and Rene Turlay observed the decay of the long-lived neutral K meson into two pions. The effect, although small, proved that the definite-mass state is not exactly the CP-odd state; it has a small admixture of the even CP state. If the mass eigenstates are not CP eigenstates, then CP symmetry is broken.

The experimental result was simple, irrefutable, and rapidly confirmed. The result was a surprise and a puzzle too; all then known particle theories conserved CP. The modern theory of particle interactions—the three-generation standard model—allows CP-violating effects such as were discovered with neutral K mesons, but that was not evident until almost 10 years later.

Enter Sakharov

If CP is not conserved, then the matter–antimatter asymmetry of the universe could result from cosmic evolution rather than from the initial condition that Pauli so disliked. One of the first to recognize that possibility was Andrei Sakharov (shown in figure 4). In 1967, he suggested that baryons and antibaryons were present in equal quantities in the early universe, and that baryon–antibaryon asymmetry developed at some later time.[3] Such "baryogenesis" clearly requires processes that change baryon number but, Sakharov noted, CP symmetry must also be violated: In a CP-conserving universe, any baryon-producing process would be balanced by a CP-related process that produces antibaryons.

Sakharov's primary observation was that any generation of net baryon number must occur at a time when the universe is out of equilibrium. The equal masses of particles and their antiparticles—assured by CPT symmetry—imply that if a particle and its antiparticle are in equilibrium, their populations are equal. Once baryon–antibaryon asymmetry has developed, baryon-number-changing processes must be rare if the asymmetry is to be preserved. Sakharov's was a revolutionary paper; when it was published, the notion of baryon-number conservation was still firmly established in the theorist's canon.

The standard model is a fully developed theory of particle interactions. At high temperature, it includes baryon-number-changing terms as well as terms that violate CP symmetry. The model can satisfy all the baryogenesis con-

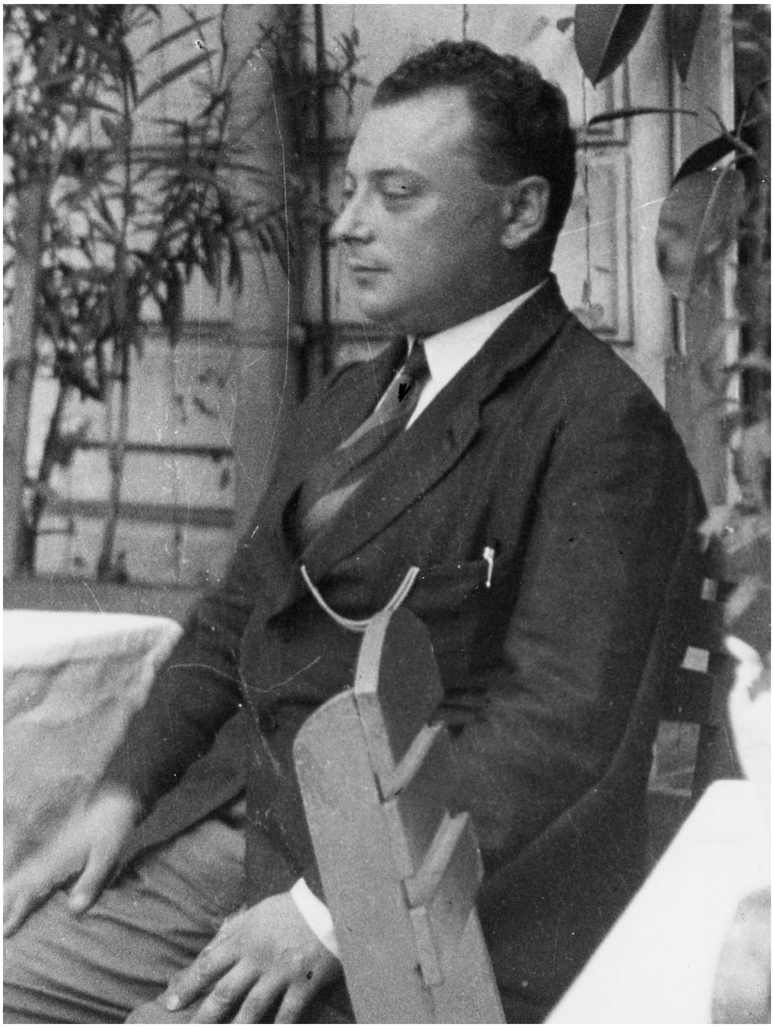


Figure 3. Wolfgang Pauli (1900–58) photographed in Odessa, Ukraine, in the late 1930s. A few years before this photograph was taken, Pauli expressed his view that dynamical generation of matter–antimatter asymmetry was more satisfying than positing such an asymmetry as an initial condition. (Photograph by Francis Simon. Courtesy of the AIP Emilio Segrè Visual Archives.)

ditions that Sakharov enumerated. The question that must be addressed is whether standard-model baryogenesis gives the observed cosmic imbalance.

The early version of the model was a two-generation theory, with four quarks and four leptons. In 1973, Makoto Kobayashi and Toshihide Maskawa considered *CP* symmetry breaking in that theory.[2] They pointed out that the theory with two generations conserves *CP*, and showed that a three-generation generalization allows *CP* violation. In the generalized theory, *CP* violation is described in terms of a single parameter, a relative phase in the matrix of couplings of the W boson between any up-type (charge $+2/3$) quark and any down-type (charge $-1/3$) quark. The three-generation matrix generalizes the Cabibbo matrix of the two-generation theory and is called the CKM matrix (for Nicola Cabibbo, Kobayashi, and Maskawa). At a time when most physicists were skeptical of the hypothesis of a fourth quark, adding two more seemed a flight of fancy. And, as Steven Weinberg observed, there was another way to introduce *CP* violation to the two-generation theory: One could add more Higgs bosons instead of adding quarks.[4]

A major change in attitude occurred in 1974–75. One reason was the discovery of particles containing charm quarks, which confirmed the fourth-quark hypothesis of the original standard model. Another new particle, the tau lepton, was produced together with its antiparticle in the same energy region as the charm-containing particles. So by the time the second generation of particles was complete, physicists had also seen a third lepton—a third generation was no longer a fringe suggestion. Kobayashi and Maskawa's three-generation theory became the standard model, and with it, *CP* violation became an effect that could be considered theoretically.

Is the single parameter describing *CP* violation in the Kobayashi–Maskawa scheme the full story? To date, all quark weak decays are consistent with a single set of standard-model parameters. New measurements and new calculations continue to test and refine the model. However, calculations show that the standard model does not provide a satisfactory framework for describing the cosmological evolution of matter–antimatter asymmetry! Some modification of the theory is needed.

Many theorized extensions suggest additional Higgs particles along with the now well-established three generations. In all those theories, it is the rich variety of possible couplings of the Higgs particles that, as well as giving the quarks masses, also gives the pattern of quark weak couplings and the *CP*-violating effects in them. Another modification, demanded by a growing body of evidence,[5] is to give mass to the neutrinos. Massive neutrinos allow for *CP*-violating effects in the lepton sector in addition to the usual violations of the standard-model quark sector.

Phase transitions

The violation of *CP* symmetry was only one of Sakharov's conditions for the generation of cosmic matter–antimatter asymmetry. Equally striking was the requirement that baryon number not be a conserved quantity. The standard model predicts that baryon- (and lepton-) number-changing processes are possible. Those processes are nonperturbative multiparticle effects and are very rare at temperatures significantly lower than the weak-scale temperature (about 100 GeV) below which the weak and electromagnetic forces became appreciably different. The rarity of processes that change baryon number is consistent with the observed halflife of the proton—greater than 10^{32} years. At the extremely high temperature of the early universe, however, baryon-number-changing processes are frequent and, given thermal equilibrium, assure that baryons and their antibaryon partners exist in equal numbers.

One way in which the lack of thermal equilibrium necessary for matter–antimatter imbalance can develop is when the universe undergoes a phase transition. Such a scenario can lead directly to baryon-number generation. Phase transitions occur when the universe cools to certain critical temperatures in the course of its expansion. The transition of interest in standard-model baryogenesis occurs at the weak scale, when the universe cools enough so that the effective potential for the Higgs field develops a global minimum at nonzero field (see figure 5).[6] All particle masses (except that of the Higgs) are generated as a consequence of the newly developed minimum. Once the universe has cooled below the critical temperature, the large mass of the W and Z bosons greatly suppresses baryon-number-changing processes. Thus, any matter–antimatter imbalance produced during the phase transition persists.

Figure 4. Andrei Sakharov (1921–89). In 1967, Sakharov enumerated the conditions necessary for baryon–antibaryon asymmetry to be a consequence of cosmic evolution. (VNIIEF Museum and Archive. Courtesy of the AIP Emilio Segrè Visual Archives.)

Baryogenesis at the weak-scale transition requires a first-order phase transition in which a bubble region forms with a distinct bubble wall. Outside the bubble, the average value of the Higgs field is zero; but inside, the Higgs field has its low-temperature, nonzero value. The bubble rapidly grows. A surplus of matter over antimatter inside the bubble can arise because *CP* violation in Higgs–fermion couplings gives different transmission probabilities for quarks and antiquarks to pass through the bubble wall as the wall sweeps through space. Outside the bubble, quarks and antiquarks are still in thermal equilibrium, so the external region has no net baryon number.

Experimental lower limits on the Higgs mass, obtained within the past two years,[7] have destroyed the viability of the standard-model phase-transition scenario: The relevant phase transition is not first order. The failure, though, instead of being in the phase-transition idea, may be in the underlying particle physics of the standard model. For example, if one adds a second Higgs field to the standard model, the conditions for a first-order phase transition might be recovered. Furthermore, an extra Higgs field can also give additional *CP*-violating effects at the phase boundary. Such effects are necessary if the bubble scenario is to match observations.

One can investigate more complex extensions of the standard model, among them supersymmetric and grand unified theories. The list of ideas is long; only further particle physics experiments can narrow it down. If their mass reach is sufficient, direct searches can find the additional particles predicted in the extensions. Indirect searches via consistency tests of standard-model predictions (for example, in the decay of B mesons) can also provide clues. Experiments severely constrain, but do not rule out, theories that give baryogenesis at a phase transition.

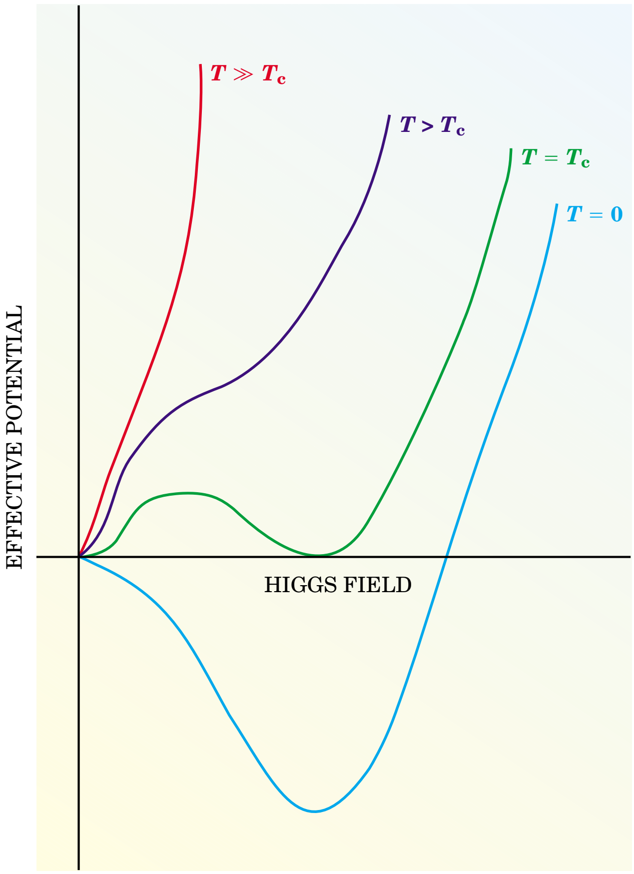
Massive neutrino decay

A second mechanism for the out-of-equilibrium generation of matter–antimatter asymmetry posits that a very weakly interacting massive particle was produced in the hot early universe and decayed late enough that the inverse production processes were no longer probable. *CP* violation in the decay of such a particle could have then produced a matter–antimatter imbalance in its decay products. Scenarios of this type are plausible—indeed are of great interest to particle physicists—now that we know neutrinos have tiny masses. Theories that explain those small masses typically give additional very massive, essentially noninteracting neutrino types. The decay of the massive neutrinos generates net lepton number, a process known as leptogenesis.

How do physicists know that neutrinos have masses? The weak-interaction-decay eigenstates and mass eigenstates of the light neutrinos are not the same. An observable consequence is neutrino oscillation. Neutrinos are produced as particles with definite flavor but, due to the superposition of the different mass eigenstates in the definite flavor neutrino, they appear to oscillate to some other flavor (and back) as they travel. Oscillation effects have been observed both for neutrinos produced in the Sun and for those produced from the decay of cosmic-ray secondary particles in the upper atmosphere. (See PHYSICS TODAY, July 2002, page 13 and August 1998, page 17.)

So far, neutrino masses have proved too small to be measured directly, but observations do tell physicists something about neutrino mass differences. Those differences and the weak-interaction mixing parameters are beginning to be explored. As with the CKM matrix that describes quark couplings, the mixing matrix that defines the weak couplings of charged leptons to definite-mass neutrinos can lead to *CP* violation. Perhaps that violation can give the matter–antimatter asymmetry of the universe.

Figure 5. The Higgs field effective potential evolves as the universe cools. At high temperatures (red and purple curves) the potential has a single minimum where the Higgs field vanishes. As the temperature cools, a second local minimum develops; the critical temperature (green curve) is defined as the temperature at which the values of the effective potential at the two minima are equal. At temperatures below the critical value, about 100 GeV for the standard-model weak-scale phase transition, the effective potential is smallest for a nonzero value of the Higgs field. As a consequence, particles have masses. The zero-temperature effective potential (blue curve) has a minimum at a nonvanishing Higgs field.



In addition to violations associated with light neutrinos, *CP* violation can appear in the very massive neutrino sector, allowing those very massive neutrinos—produced in the early universe—to decay via processes that generate net lepton number. The leptons then can produce baryons through standard-model processes that change both baryon and lepton number but not their difference. The result is the observed matter excess. Leptogenesis scenarios have been explored in a number of theoretical models, including grand unified theory extensions of the standard model. The details depend on the particular extension. One can find models consistent with the observed matter–antimatter asymmetry;[8] time will tell if those models are satisfactory in other aspects of their predictions. Certainly, they are an interesting option.

Experimental probes

Particle physicists do not yet know all the details of the asymmetry in the underlying laws of physics between matter and antimatter. Standard-model *CP* violation is not the full story; some new physics is required. We now know that neutrino masses must be added, and that this addition may bring some further *CP*-violating parameters. But is that enough?

Some of the answers can be sought in the high-energy laboratory. Indeed, the large worldwide effort to understand the physics of B mesons is focused on elucidating the breakdown of *CP* symmetry in the quark sector. B physics provides a wonderful laboratory to study *CP* violation because both the B_d and B_s mesons form pairs similar to the neutral K mesons. That is, the flavor eigenstates are mixed to form mass eigenstates, allowing for sensitive probes of *CP* violation. Moreover, because of the large mass of the bottom quark, the B mesons have many possible decay channels. The multiple decays allow physicists to develop redundant determinations of standard-model parameters. If the measurements are all consistent with a single set of four parameters that define the CKM matrix, then physicists will have more precisely determined values for those parameters. If not, further study will point toward one or another possible addition to the standard model—and perhaps help us understand baryogenesis too.

The work has begun well: The SLAC and KEK B factories have produced first results[9] and will generate additional interesting data over the next 5–10 years (See Physics Today, September 2001, page 19). Those data will be complemented by other results from B-physics experiments at hadron colliders. So far, the experiments are consistent with the standard model.[10] There is plenty of room for new discoveries, though: As yet, only a few channels of B decays have sufficient data for sensitive analysis.

Further experiments may also teach physicists much about neutrinos. Not all the parameters of the very massive neutrino sector can be determined in foreseeable experiments. However, the unknown parameters of the light sector can be pinned down. *CP* violation in the light sector will be difficult to detect. How difficult depends on the size of an as-yet-unmeasured mixing parameter. What we learn will provide important constraints on model-building and scenarios for leptogenesis.

I am particularly indebted to the work of Abraham Pais, whose scholarship has been invaluable to this nonhistorian.

References

1. All historical quotes are from Abraham Pais's lecture in *Paul Dirac: The Man and His Work*, P. Goddard, ed., Cambridge U. Press, New York (1998).
2. M. Kobayashi, T. Maskawa, *Prog. Theor. Phys.* **49**, 652 (1973).
3. A. D. Sakharov, *J. Exp. Theor. Phys. Lett.* **5**, 24 (1967).
4. S. Weinberg, *Phys. Rev. Lett.* **37**, 657 (1976).
5. For a review and references to relevant data, see M. C. Gonzalez-Garcia, Y. Nir, http://www.arXiv.org/abs/hep-ph/0202058.
6. For a review, see A. G. Cohen, D. B. Kaplan, A. E. Nelson, *Annu. Rev. Nucl. Part. Sci.* **43**, 27 (1993).
7. For a summary, see A. Raspereza, http://www.arXiv.org/abs/hep-ex/0209021.
8. For a review, see W. Buchmüller, in *Proc. 2001 European School of High-Energy Physics, Beatenberg, Switzerland (26 Aug–8 Sept 2001)*, N. Ellis, J. March-Russell, eds., CERN, Geneva, Switzerland (2002). Also available online at http://www.arXiv.org/abs/hep-ph/0204288.
9. For a review with references to new results, see Y. Nir, http://www.arXiv.org/abs/hep-ph/0208080.
10. F. J. Gilman, K. Kleinknecht, B. Renk, in K. Hagiwara et al., *Phys. Rev. D* **66**, 010001 (2002), available online at http://pdg.lbl.gov/2002/kmmixrpp.pdf. ■