IS THERE BEAUTY IN MATHEMATICAL THEORIES?

ROBERT P. LANGLANDS

Contents

Introduction

1. Mathematics and beauty. I was pleased to receive an invitation to speak at a gathering largely of philosophers and accepted with alacrity but once having accepted, once having seen the program, I was beset by doubts. As I proceeded, it became ever clearer that the doubts were entirely appropriate. As an elderly mathematician, I am fixed more and more on those problems that have preoccupied me the most over the decades. I cannot, even with the best of intentions, escape them. That will be clear to every reader of this essay and of a second related, but yet unwritten, essay with which I hope to accompany it later.

I left the title suggested to me by the organizer of the symposium, Professor Hösle, pretty much as it was, *Is there beauty in purely intellectual entities such as mathematical theories?* I abbreviated it, for as a title it was too long, but did not otherwise change it. I do not, however, address directly the question asked. The word *beauty* is too difficult for me, as is the word *aesthetics*. I hope, however, that the attentive reader of this essay will, having read it, be able, on his own, to assess soberly the uses and abuses of the first of these words in connection with mathematics.

I appreciate, as do many, that there is bad architecture, good architecture and great architecture just as there is bad, good, and great music or bad, good and great literature but neither my education, nor my experience nor, above all, my innate abilities allow me to distinguish with any certainty one from the other. Besides the boundaries are fluid and uncertain. With mathematics, my topic in this lecture, the world at large is less aware of these distinctions and, even among mathematicians, there are widely different perceptions of the merits of this or that achievement, this or that contribution. On the other hand, I myself have a confidence in my own views with regard to mathematics that does not spring, at least not in the first instance, from external encouragement or from any community of perception, but from decades of experience and a spontaneous, unmediated response to the material itself. My views as a whole do not seem to be universally shared by my fellow mathematicians, but they are, I believe, shared to a greater or lesser extent by many. They are at worst only a little eccentric and certainly not those of a renegade. I leapt at the unexpected possibility to reflect on them, in what I hope is an adequately critical manner, and to attempt to present them in a way persuasive to mathematicians and to others.

Date: University of Notre Dame, January 2010.

I confess, however, that although aware that I agreed—in good faith—to speak not on mathematics as such, but on beauty and mathematics, I soon came to appreciate that it is the possibility of greatness, or perhaps better majesty and endurance, in mathematics more than the possibility of beauty that needs to be stressed, although greatness is unthinkable without that quality that practitioners refer to as beauty, both among themselves and when advertising the subject, although not infrequently with a sincerity that, if not suspect, is often unreflected. This beauty, in so far as it is not just the delight in simple, elegant solutions, whether elementary or not, although this is a genuine pleasure certainly not to be scorned, is, like precious metals, often surrounded by a gangue that is, if not worthless, neither elegant nor intrinsically appealing. Initially perhaps there is no gangue, not even problems, perhaps just a natural, evolutionary conditioned delight in elementary arithmetic—the manipulation of numbers, even of small numbers—or in basic geometric shapes—triangles, rectangles, regular polygons. To what extent it is possible to become a serious mathematician without, at least initially, an appreciation of these simple joys is not clear to me. They can of course be forgotten and replaced by higher concerns. That is often necessary but not always wise. Simple joys will not be the issue in this essay, although I myself am still attached to them. As an aside, I add that, as with other simple human capacities, such as locomotion, the replacement of innate or acquired mathematical skills by machines is not an unmixed blessing.

There is also a question to what extent it is possible to appreciate serious mathematics without understanding much of mathematics itself. Although there are no prerequisites for reading this note, it is not my intention to avoid genuine mathematics, but I try to offer it only in homeopathic doses.

Two striking qualities of mathematical concepts regarded as central are that they are simultaneously pregnant with possibilities for their own development and, so far as we can judge from a history of two and a half millennia, of permanent validity. In comparison with biology, above all with the theory of evolution, a fusion of biology and history, or with physics and its two enigmas, quantum theory and relativity theory, mathematics contributes only modestly to the intellectual architecture of mankind, but its central contributions have been lasting, one does not supersede another, it enlarges it. What I want to do is to examine, with this in mind, the history of mathematics—or at least of two domains, taken in a broad sense, with which I have struggled—as a chapter in the history of ideas, observing the rise of specific concepts in the classical period, taking what I need from standard texts, Plato, Euclid, Archimedes, and Apollonius, then passing to the early modern period, Descartes and Fermat, Newton and Leibniz, and continuing into the nineteenth, twentieth, and even twenty-first centuries.[1] It is appropriate, even necessary, to recount the history of the two domains separately. One can conveniently be referred to as algebra and number theory; the other as analysis and probability. Their histories are not disjoint, but their current situations are quite different, as are my relations to the two. The first is, I fear, of more limited general interest although central to modern pure mathematics. Whether it will remain so or whether a good deal of modern pure mathematics, perhaps the deepest part, has become too abstruse and difficult of access to endure is a question that will be difficult to ignore as we continue these reflections. Even if mankind survives, there may be limits to our ability or, more likely, to our desire to pursue the mathematical reflections of the past two or three thousand years.

I have spent more time, more successfully, with the first domain but am nevertheless even there if not a renegade an interloper and regarded as such. In the second, I am not regarded as an interloper, simply as inconsequential. Nevertheless, if for no other reason than for the

sake of my credibility as an impartial commentator on mathematics and its nature, I shall try to express clearly and concretely my conviction that the problems of both are equally difficult and my hope that their solutions equally rich in consequences. Unfortunately, for reasons of time and space the description of my concerns in the second domain must be postponed to another occasion.

2. Theorems and Theories. The difference between theorems and theories is more apparent in the first domain than in the second. Any mathematical theory will contain valid general statements that can be established rigorously accordingly to the commonly accepted logical principles, thus proved, and without which the theory would lack its structural coherence. They are referred to as theorems, but there is another sense to this word that refers to those statements that, independently of their necessity to the theory are provable within it, and, above all, of an importance that transcends their structural value within the theory. Fermat's theorem, to which we shall return later, is a striking instance and belongs to the first domain and to pure mathematics. Without theorems of this sort, a theory in pure mathematics would be an insipid intellectual exercise. On the other hand, theorems of this sort are seldom, perhaps never, established without the concepts that are only available within a theory. It is the latter, as should appear from the following observations, that give pure mathematics its honorable place in intellectual history. In analysis, familiar in the form of differential and integral calculus, the test of a theory is less this type of theorem, than its pertinence outside of mathematics itself.

The more I reflected on this lecture, the more I was aware that in presuming to speak, above all to a lay audience, on the nature of mathematics and its appeal, whether to mathematicians themselves or to others, I was overreaching myself. There were too many questions to which I had no answer. Although I hope, ultimately, to offer appreciations of two quite different mathematical goals, I do not share the assurance of the great nineteenth-century mathematician C. G. J. Jacobi of their equal claims. In a letter to A.-M. Legendre of 1830, which I came across while preparing this lecture, Jacobi famously wrote,

> *Il est vrai que M. Fourier avait l'opinion que le but principal des mathématiques était l'utilité publique et l'explication des phénomènes naturels; mais un philosophe comme lui aurait dû savoir que le but unique de la science, c'est l'honneur de l'esprit humain, et que sous ce titre, une question des nombres vaut autant qu'une question du système du monde.*

I am not sure it is so easy. I have given a great deal of my life to matters closely related to the theory of numbers, but the honor of the human spirit is, perhaps, too doubtful and too suspect a notion to serve as vindication. The mathematics that Jacobi undoubtedly had in mind when writing the letter, the division of elliptic integrals, remains, nevertheless, to this day unsurpassed in its intrinsic beauty and in its intellectual influence, although in its details unfamiliar to a large proportion of mathematicians. Moreover, the appeal to the common welfare as a goal of mathematics is, if not then at least now, often abusive. So it is not easy to find an apology for a life in mathematics.

In Jules Romains's extremely long novel of some 4750 pages *Les hommes de bonne volonté* a cardinal speaks of the Church as *"une œuvre divine faite par des hommes"* and continues *"La continuité vient de ce qu'elle est une œuvre divine. Les vicissitudes viennent de ce qu'elle est faite par les hommes."* There are excellent mathematicians who are persuaded that mathematics too is divine, in the sense that its beauties are the work of God, although

they can only be discovered by men. That also seems to me too easy, but I do not have an alternative view to offer. Certainly it is the work of men, so that it has many flaws and many deficiencies.[2]

Humans are of course only animals, with an animal's failings, but more dangerous to themselves and the world. Nevertheless they have also created—and destroyed—a great deal of beauty, some small, some large, some immediate, some of enormous complexity and fully accessible to no one. It, even in the form of pure mathematics, partakes a little of our very essence, namely its existence is, like ours, like that of the universe, in the end inexplicable. It is also very difficult even to understand what, in its higher reaches, mathematics is, and even more difficult to communicate this understanding, in part because it often comes in the form of intimations, a word that suggests that mathematics, and not only its basic concepts, exists independently of us. This is a notion that is hard to credit, but hard for a professional mathematician to do without. Divine or human, mathematics is not complete and, whether for lack of time or because it is, by its very nature, inexhaustible, may never be.

I, myself, came to mathematics, neither very late nor very early. I am not an autodidact, nor did I benefit from a particularly good education in mathematics. I learned, as I became a mathematician, too many of the wrong things and too few of the right things. Only slowly and inadequately, over the years, have I understood, in any meaningful sense, what the penetrating insights of the past were. Even less frequently have I discovered anything serious on my own. Although I certainly have reflected often, and with all the resources at my disposal, on the possibilities for the future, I am still full of uncertainties. These confessions made to the innocent reader, I am almost ready to begin. Before doing so, I would like to describe a theorem with a sketch of its proof, because it may be the most striking illustration available of the relation between theorems and theories that I have in mind.

3. Fermat's theorem. It is easy to solve the equation $a^2 + b^2 = c^2$ in integers, none of which is 0. One writes it as $a^2 = c^2 - b^2 = (c + b)(c - b)$ and takes $c + b = r^2$, $c - b = s^2$, $a = rs$. On the other hand, for $n > 2$ there is no solution of $a^n + b^n = c^n$ with none of a, b, c equal to 0. This is the famous theorem of Fermat. It is, of course, not my intention to provide a proof here, nor even to provide a history of the proof, of which there are a number of accounts, many reliable. I just want to point out certain elements. If there is a solution for n equal to a prime p, and this is the decisive case, then one considers, after Gerhard Frey and Yves Hellegouarch, the curve in the (x, y) plane defined by $y^2 = x(x - a^p)(x + b^p)$. The form of this curve is special: quadratic in y and cubic in x, what is called an elliptic curve. Moreover the coefficients are rational numbers. There are many conjectures of greater or less precision—rather greater than less and undoubtedly correct, although we are a long way from constructing the *theory* they entail and that entails them—establishing a correspondence between equations with rational coefficients and what are called *automorphic forms*, a notion to which we shall return. For elliptic curves the correspondence takes a particularly simple form whose meaning is clear to a sizeable fraction of mathematicians, and was formulated somewhat earlier than the general form, although its proof demands, among other things, an appeal to consequences of the little of the general theory that is available. In general and in this case as well, they affirm that associated to any equation with rational coefficients are one or several automorphic forms. It is surprising because these are apparently much more sophisticated objects. By and large automorphic forms are none the less easier to enumerate than curves. In particular, those automorphic forms with small *ramification*, a basic technical concept, that might correspond to elliptic curves can be counted and there are

not enough of them to account for the Frey-Hellegouarch curve. So it cannot exist and there is no solution of Fermat's equation for $n > 2$. Fermat's theorem is, and probably will remain, the quintessential example of an apparently elementary theorem that is a consequence of a very sophisticated, very demanding theory.

I have included, as an aid in the explanation of the historical evolution of the context in which this proof is effected, two diagrams. In the first, a large number of mathematical theories and concepts are given, within rectangular frames connected by lines in an attempt to express the relations between them. The historical development proceeds from top to bottom. In the hope that it will provide some temporal orientation to the reader unfamiliar with the concepts themselves, there is a second diagram with the names of some of the better-known creators of the concepts. It is not otherwise to be taken seriously. Not all names will be equally familiar; they are not equally important.

The labels in the elliptical frames stand apart (Diagram A). There are four of them, in two columns and two rows. The columns refer to the theories at their head: contemporary algebraic geometry and diophantine equations on the left and automorphic forms on the right. The first row corresponds to central concepts in these two theories, motives on the left, functoriality on the right. Neither of these words will evoke anything mathematical at first. Both are, in the sense in which they are used in the diagram, problematical. It is clear, at least to a few people, what is sought with these two concepts. It is also clear that at the moment they are largely conjectural notions. Both express something essential about the structure of the theories to which they are attached. This structure is expressed not so much in terms of what mathematicians call groups but in terms of their representations, which form what mathematicians call a Tannakian structure. That is a somewhat elusive concept and I will make no attempt to formulate it here. Its mystery is heightened as much as explained by an appeal to a laconic observation of Hermann Weyl in the preface to his celebrated exposition *Gruppentheorie und Quantenmechanik* published in 1928. "Alle Quantenzahlen sind Kennzeichen von Gruppendarstellungen".

The variety of wavelengths emitted by various atoms and molecules was in the early years of the last century completely inexplicable and a central puzzle for the physicists of the time. It was explained only in the context of the new quantum theory and, ultimately, with the notion of group representations and of the product of two representations, thus pretty much in what mathematicians refer to as a Tannakian context. This was a magnificent achievement and Weyl is referring to it. Although automorphic forms or motives are mathematical notions, not physical phenomena, and thus, especially automorphic forms, much closer, even in their definition, to groups than the spectra of atoms and molecules are to the rotation group and its representations, the clarifying connection between the ellipses on the left and those on the right is every bit as difficult to discover: between diophantine geometry or motives and automorphic forms or functoriality; and between ℓ-adic representations and Hecke operators. We are hoping not merely to elucidate the two concepts of motives and functoriality in the two theories to which they are here attached, diophantine equations and automorphic forms, but to create them. Although there is little doubt, at least for me, that the proposed conjectures are valid, it is certain that a great deal of effort, imagination, and courage will be needed to establish them.

Nevertheless some of them, the very easiest ones, have been proved. These were adequate to the proof of Fermat's theorem. Rather, some few examples of functoriality, which had implications not only for the column on the right but also for the correspondence between

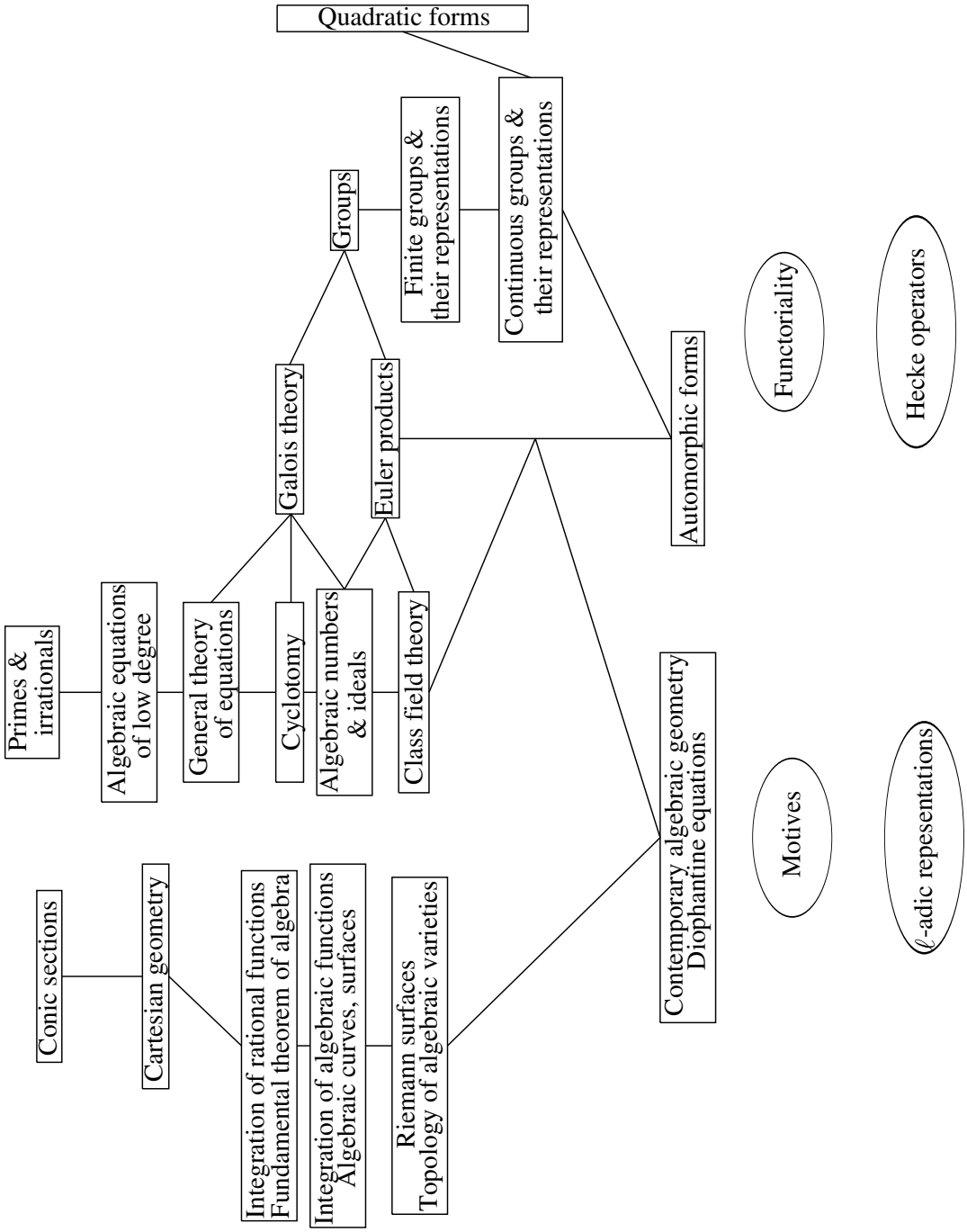
Quadratic forms
Continuous groups & their representations
Finite groups & their representations
Groups
Galois theory
Euler products
Automorphic forms
Functoriality
Hecke operators
Primes & irrationals
Algebraic equations of low degree
General theory of equations
Cyclotomy
Algebraic numbers & ideals
Class field theory
Conic sections
Cartesian geometry
Integration of rational functions Fundamental theorem of algebra
Integration of algebraic functions Algebraic curves, surfaces
Riemann surfaces Topology of algebraic varieties
Contemporary algebraic geometry Diophantine equations
Motives
ℓ-adic repesentations

Diagram A

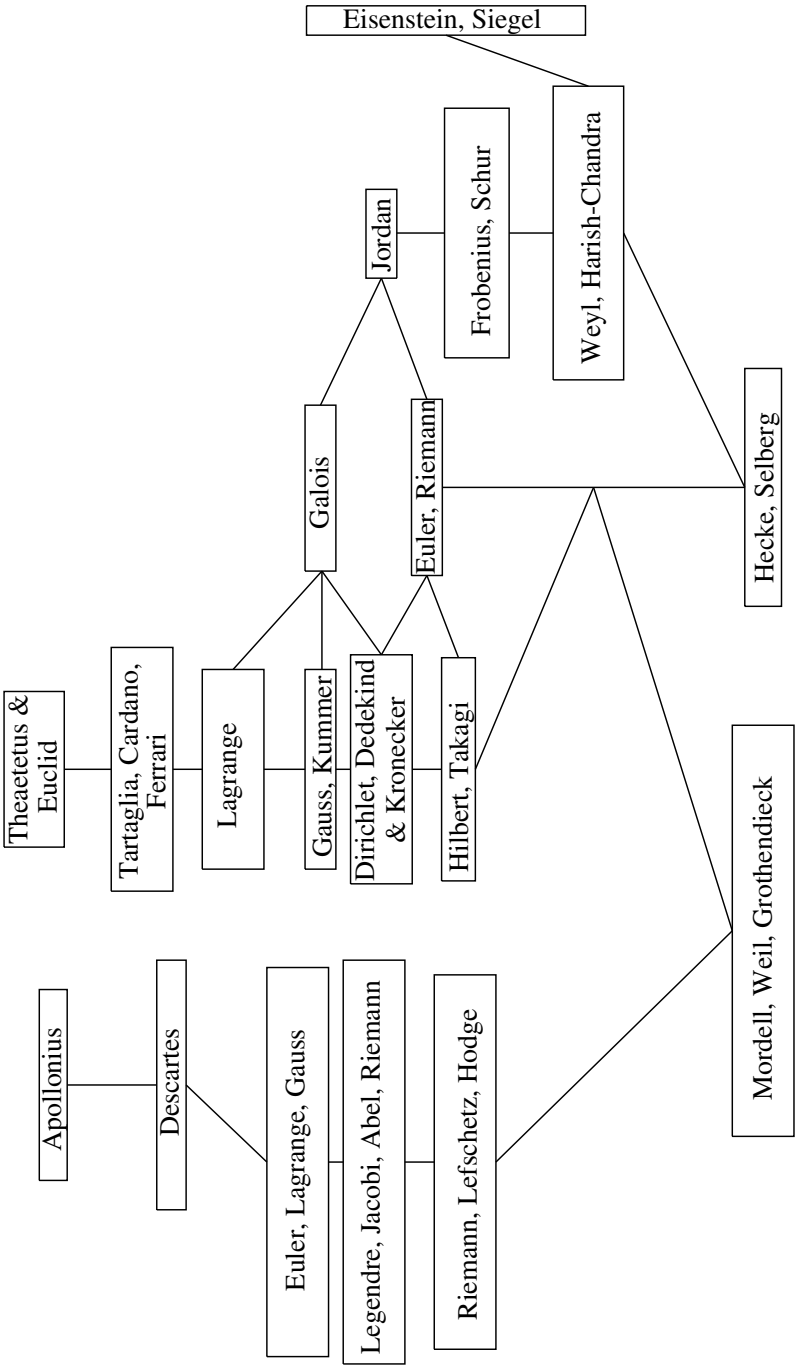
Eisenstein, Siegel
Jordan
Frobenius, Schur
Weyl, Harish-Chandra
Galois
Euler, Riemann
Hecke, Selberg
Theaetetus & Euclid
Tartaglia, Cardano, Ferrari
Lagrange
Gauss, Kummer
Dirichlet, Dedekind & Kronecker
Hilbert, Takagi
Apollonius
Descartes
Euler, Lagrange, Gauss
Legendre, Jacobi, Abel, Riemann
Riemann, Lefschetz, Hodge
Mordell, Weil, Grothendieck

Diagram B

the two columns, were available to Andrew Wiles, who recognized that with them another much more difficult conjecture, the Shimura-Taniyama conjecture, again a part of the correspondence but more difficult, was within his reach. This great achievement represents a second stage, not simply the creation of the two theories in the two columns, but the proof, or at least a partial proof, that the one on the left is reflected at the level of algebraic curves in the one on the right. This is the essence of the proof of Fermat's theorem, contemporary in its spirit. If the theorem were false, it would imply the existence of an elliptic curve, a motive, thus a diophantine object attached to the left-hand column. Since the diophantine theory is reflected in the theory of automorphic forms, there would be in the right-hand column an object attached to it, a corresponding automorphic form that, thanks to the nature of the correspondence would have some very specific properties, properties that permit us to conclude that it could not possibly exist. So Wiles's glory is perhaps not so much to have established Fermat's theorem as to have established the Shimura-Taniyama conjecture.

Indeed, if it had never been conjectured that Fermat's equations had no nontrivial integral solutions and if mathematicians had arrived by some other route at the Shimura-Taniyama conjecture and its proof—not in my view entirely out of the question although some essential steps in Diagram A were clearly inspired by the search for a proof of Fermat's theorem—and if then some modest mathematician had noticed that, as a consequence, the equation $a^n + b^n = c^n$, $n > 2$ had no integral solutions except the obvious ones with $abc = 0$, little attention would have been paid to him or to the equation. It might be inferred from this observation that the relation between the intrinsic value of a mathematical notion and the depth of its historical roots is subtler than I suggested when beginning this essay. I find none the less that is of great value as a mathematician to keep the historical development in mind.

The last row offers for completeness the parameters that are presently most commonly used for comparing the two columns, or rather for deciding whether any proposed correspondence of the objects in the two columns is correct. The existence of the ℓ-adic representations is an acquisition of the last few decades and, like the fundamental theorem of algebra, which will be formulated below, expresses something fundamental about the relation between geometry and irrationality.

When struggling with the difficulties, and they are many, that must be overcome in any attempt to establish the correspondence here adumbrated, I am troubled by a question of a quite different nature. Certainly, for historical reasons if for no other, any mathematical theory that yields Fermat's theorem has to be taken seriously. Fermat's theorem now established, what is the value of developing the theory further? What will the correspondence give? This is by no means clear. My guess is that it will be difficult to create the theory entailing the correspondence of the two columns without establishing a good part of what is known as the Hodge conjecture, perhaps the outstanding unsolved problem in higher-dimensional coordinate geometry. It has remained inaccessible since its first—slightly incorrect, but modified later—formulation in 1950. I see, however, no reason that the theory could not be established without coming any closer to the Riemann hypothesis, on which a great deal of effort has been expended over the past 150 years, but the theory does suggest a much more general form of the hypothesis and puts it in a much larger context. A different and very hard question is what serious, concrete number-theoretical results does the theory suggest or, if not suggest, at least entail. A few mathematicians have begun, implicitly or explicitly, to reflect on this. I have not.

A. Pure Mathematics

A.1. From then to now. The proof of Fermat's theorem is a cogent illustration of the concrete consequences of the correspondence symbolized by the first pair of ellipses in Diagram A. The diagram itself and the explanations, partial and unsystematic, in the following paragraphs are an attempt, more for my own sake than for that of the reader, to integrate this correspondence—the central theme of this essay, devoted to pure mathematics and, God willing, the first of two essays—with various insights or contributions over the centuries that would be regarded by all mathematicians as decisive and to demonstrate that it is the legitimate issue of these contributions. It would take more than a lifetime to appreciate them all, but each and every one offers, each time that it is taken up and its content and implications just a little better understood, a great deal of immediate pleasure and satisfaction. In my life as a mathematician I have spent very little, too little time reflecting on the past of mathematics, and this lecture is an occasion to remedy the neglect and redeem myself.

There are two major themes arising in antiquity and accumulating mass and structure over the centuries that contribute to the correspondence symbolized by the two ellipses: geometry, of which Apollonius is a major classical representative, for our purposes more telling than Euclid; and arithmetic—a word often used among mathematicians as a substitute for the phrase "theory of numbers", which is more embracing—for which I have chosen in the restricted space of the diagram Theaetetus and Euclid as representatives, rather than Pythagoras. Although Book X of the Elements is readily available in well-edited English translations and contains much of which, so far as I know, Theaetetus was unaware, his name is likely to have more sentimental appeal to philosophers, because of his modest but poignant appearance in the Platonic dialogue that bears his name, first briefly and off-stage, as a soldier wounded and ill with dysentery, and then, in the recollected dialogue, as a very young man and a foil for the questions of Socrates, although he is allowed some mathematical interjections. The work of Apollonius on conic sections, in which the equations of ellipses, hyperbolas, and parabolas familiar to most of us appear, although without symbols and without Cartesian coordinates, are, with the editorial help needed for penetrating the unfamiliar style of exposition, a pleasure to read even in translation. They have been published in several languages. In particular, they were made available in Latin in 1566. Descartes had, apparently read and understood the first parts, much to his profit and much to ours.

The first column in both diagrams refers to the development of Cartesian geometry, based on Descartes interpretation of Apollonius, from its introduction in *Sur la géométrie* to the present. The central feature of Cartesian geometry is the use of coordinates and it is preferable, for many reasons, to refer to it as coordinate geometry, but there are also good reasons for recalling the name of Descartes, whose contribution to mathematics was of enduring value, although as a mathematician he was brash and not always frank in acknowledgement of his debt to Apollonius. The use of coordinates ultimately allowed the introduction of what are called the projective line, the projective plane, or higher-dimensional projective spaces, whose purpose is, one might say, largely to avoid the constant introduction of special cases, when some of the solutions of algebraic equations might otherwise slip off to infinity. Even these solutions are captured in projective geometry.

This enables us to exploit systematically the consequences of the fundamental theorem of algebra. This theorem asserts in its contemporary form that every equation of the form

$$X^n + a_{n-1}X^{n-1} + \cdots + a_1X + a_0 = 0. \tag{1.1}$$

with complex numbers $a_{n-1}, \dots, a_1, a_0$ as coefficients has exactly n roots, or, better expressed, because some of the roots may be multiple, that the left side factors as

$$X^n + a_{n-1}X^{n-1} + \cdots + a_1X + a_0 = (X - \alpha_1)(X - \alpha_2)\cdots(X - \alpha_n). \tag{1.2}$$

The most familiar example is

$$X^2 + bX + a = 0,$$

for which $\alpha_1 = (-b + \sqrt{b^2 - 4a})/2$, $\alpha_2 = (-b - \sqrt{b^2 - 4a})/2$. The adjective "complex", whose meaning and consequences may perhaps not be familiar to every reader, allows $b^2 - 4a$ to be negative. The modern facility with the notion of a complex number was only slowly acquired, over the course of the eighteenth and early nineteenth centuries.

The origins of the theorem lie, however, not in algebra as such or in geometry, but in the integral calculus and in formulas that will be familiar to a good many readers, those who have encountered the formulas,

$$\begin{aligned}
&\int \frac{1}{x-a}\,dx = \ln(x-a),\\
&\int \frac{1}{(x-a)^n}\,dx = -\frac{1}{n-1}\frac{1}{(x-a)^{n-1}}, \qquad n > 1,\\
&\int \frac{1}{x^2-a}\,dx = \frac{1}{2\sqrt{a}}\ln\left(\frac{x-\sqrt{a}}{x+\sqrt{a}}\right),\\
&\int \frac{1}{\sqrt{x^2-a}}\,dx = \ln(x+\sqrt{x^2-a}) = \operatorname{arccosh}\left(\frac{x}{\sqrt{a}}\right) + \text{const.},\\
&\int \frac{1}{\sqrt{x^2+a}}\,dx = \ln(x+\sqrt{x^2+a}) = \arccos\left(\frac{x}{\sqrt{-a}}\right) + \text{const.},
\end{aligned} \tag{1.3}$$

or similar formulas. The third is a formal consequence of the first. The third is an immediate consequence of the first two. The first equality of the fourth line or of the fifth can be deduced from the parametrizations $x + y = z$, $x - y = a/z$ or, rather,

$$x = x(z) = \frac{1}{2}\left(z + \frac{a}{z}\right), \qquad y = y(z) = \frac{1}{2}\left(z - \frac{a}{z}\right) \tag{1.4}$$

of the plane curve $x^2 - y^2 = a$. The second equality of these two lines is a consequence of the relations between the logarithm and the trigonometric functions or the hyperbolic functions, for example,

$$\begin{aligned}
\theta &= \frac{1}{2i}\ln\left(\frac{1 + i\tan\theta}{1 - i\tan\theta}\right),\\
a &= \frac{1}{2}\ln\left(\frac{1+\tanh a}{1-\tanh a}\right).
\end{aligned}$$

These equalities, a reflection of the intimate relation between the trigonometric functions and the hyperbolic functions and between them both and the exponential functions were

not immediately evident in the eighteenth century when the integrals first began to be investigated.

The passage from the third to the fourth and fifth lines requires understanding that the line with the parameter z is for many purposes almost the same as the curve $x^2 - y^2 = a$, for as z runs over the line, the point $\big(x(z), y(z)\big)$, runs over this curve. So the calculation of integrals brings with it a freer use of geometric abstractions that carries us forward in the left-hand columns of the two diagrams.

It is, however, more important to understand that the fundamental theorem of algebra arose in large part as an essential element of the urge to calculate integrals of the form

$$\int \frac{X^m + b_{m-1}X^{m-1} + \cdots + b_1X + b_0}{X^n + a_{n-1}X^{n-1} + \cdots + a_1X + a_0}\, dx, \tag{1.5}$$

but these integrals are seldom if ever discussed. First of all, the general formulas are less important than the special formulas; secondly they require a calculation of the roots of the denominator and that is not an easy matter. It can in most cases only be effected approximately. These are the integrals that can, in principle, be treated in introductory courses. Even though a good deal of experience is implicit in the treatment, they are definitely at a different level than many other integrals that, at first glance, appear to be of the same level of difficulty. A typical example is

$$\int \frac{1}{\sqrt{x^3 + 1}}. \tag{1.6}$$

These are the integrals with which Jacobi was familiar and that inspired his eloquence. They are astonishingly rich in geometric implications.

In order to treat (1.5) even on a formal level, the denominator has to be factored as in (1.2) or, if one is uneasy about complex numbers, as a product in which quadratic factors also appear

$$X^n + a_{n-1}X^{n-1} + \cdots + a_1X + a_0 = (X - \alpha_1) \cdots (X^2 + \beta_1 X + \beta_2) \cdots .$$

For example, the factorization

$$X^4 - 1 = (X - 1)(X + 1)(X^2 + 1),$$

is then preferable to

$$X^4 - 1 = (X - 1)(X + 1)(X + \sqrt{-1})(X - \sqrt{-1}).$$

Although such uneasiness was overcome two hundred years ago, traces of the quadratic factors will certainly remain in introductory courses in the integral calculus. For the purposes of integration, the factorization (1.2) has to be found explicitly and that entails supplementary skills.

So the difficulties raised by the formal calculation of the integrals (1.5) raised two difficulties, one formal and logical, the other mathematical, both formidable although in different senses. The first requires the acceptance of complex numbers as genuine mathematical objects; the second a proof of the possibility of the factorization (1.2) in the domain of complex numbers, no matter what real or complex coefficients are taken on the left-hand side.

Once mathematicians were easy with the factorization (1.2) and certain that it was always possible, there were two directions to pursue, or rather many, but two that are especially pertinent to Diagram A and Diagram B. One was the study of the collection of all solutions in complex numbers of one or several algebraic equations. This is the first column, most of

which has to be reckoned geometry. The other is the study of all solutions of one algebraic equation in one unknown, thus the equation (1.1), but with coefficients from a restricted domain, above all with ordinary fractions or ordinary integers as coefficients. Restrictions on the coefficients entail restrictions on the roots and the nature of these restrictions is to be revealed.

The presence of the two different long columns on the left of Diagram A, in which the one on the far left refers largely to the first direction and the one in the middle refers to the second, is to a considerable extent arbitrary, to a considerable extent a consequence of my own much more extended experience with the topics leading to automorphic forms than with those leading to diophantine equations, but not entirely artificial. The development of algebra in the late middle ages and early renaissance was every bit as much a prerequisite for Descartes as for the analysis of the roots of equations of degree three and four and the beginnings in the writings of Lagrange and others of a general analysis of the roots of the equation (1.1). Nevertheless the two columns, one carried by the study of integrals into the study of plane curves in complex coordinate geometry, really surfaces, namely Riemann surfaces, because to fix a complex number $a + b\sqrt{-1}$ requires two real numbers a and b, the other carried by the explicit solutions of equations of degree three and four discovered by Tartaglia, Ferrari and others, through the reflections of Lagrange into the study of groups, above all Galois groups, are useful as a symbol of the accumulation and the fusion of concepts that had to occur for modern pure mathematics to arise.

The configuration on the right of Diagram A is more unfamiliar. The history and the role of the theory of quadratic forms is not well understood even by mathematicians, but would be technical and too lengthy to undertake here. The triangular patch with Galois theory, groups, and automorphic forms at its vertices refers to some extent to little known, or at least little understood, overlapping of mathematics and physics on which I will not be able to resist dwelling. The triangle, class field theory, Galois theory, automorphic forms, is more arcane, but it is perhaps here where the pivotal difficulties lie. I shall attempt to adumbrate the pertinent issues in the final section.

A.2. The origins of the theory of numbers. One of the most appealing presentations of basic elements of number theory appears in Plato's dialogue *Theaetetus*, another in Book VII of Euclid. Both deal, the first only incidentally, with primes; the first also deals with the notion of irrational number, although in the guise of incommensurability.[3] These two notions of prime number and irrational number are linked to this day and lie at the foundation of number theory. Both are familiar to many laymen; both can be readily explained to an attentive, reasonably intelligent auditor.

The modern theory of numbers is very abstruse. Even among its practitioners and in spite of the occasionally elementary nature of its conclusions, it is fragmented. Because of the often fortuitous composition of the faculty of the more popular graduate schools, some extremely technical aspects are familiar to many people, others known to almost none. This is inevitable. One can already see in the Xth book of Euclid how quickly the examination of the simplest irrationalities leads to what appear to be uselessly arcane reflections.

These reflections are closely related to the considerations of the modern *Galois theory* created by the very young mathematician Évariste Galois early in the nineteenth century.[4] This theory, which analyses in a brief and incisive way, with the aid of the notion of a group of symmetries, the essential nature of algebraic irrationalities, is a part of the current undergraduate curriculum and modern pure mathematics would be unthinkable without it

or without the notion of a group, which, as we shall see, penetrates mathematics far more broadly, far more deeply than Galois theory itself.

It is useful, as an introduction to Galois theory and to symmetries, but also to acquire some insight into the genesis of mathematical theories, to recall briefly the lengthy, but little read, Book X. It begins by a careful analysis of incommensurability and thus of the irrational, a word that it is best to use here in its modern mathematical sense and not in that of Euclid, or rather that of the English translation. Rational numbers are common fractions. Irrational numbers are the remaining numbers. Euclid and Galois are concerned by and large with irrational numbers that are also algebraic, thus they satisfy an algebraic equation of the form (1.1) in which all the coefficients $a_{n-1}, \dots, a_0$ are ordinary fractions, thus rational numbers. We have already emphasized that it is central to all of algebra and to a good deal of geometry that, whether or not the coefficients are rational, the left-hand side of this equation may be written in exactly one way as

$$X^n + a_{n-1}1X^{n-1} + a_{n-2}X^{n-2} + \cdots + a_1X + a_0 = (X - \alpha_1)(X - \alpha_2)\cdots(X - \alpha_n). \tag{2.1}$$

As we observed this statement, known as the fundamental theorem of algebra, is basic to the integral calculus. It also seems, in some obscure way, to be the source of Descartes's prophetic insistence in *Sur la géométrie*, 1637, on the notion of the degree of a curve but was proved only much later, as a consequence of profound reflections on the nature of what are known today as *real* numbers. This geometrical or topological aspect of the fundamental theorem is quite unrelated to the distinction (2.1) allows between algebraic numbers and the remaining numbers in our number system.

The integers $1, 2, 3, \dots$ form a familiar collection of numbers. The division of a unit into a number of parts of equal length leads quickly to rational fractions, 1/2 or 1/3, 2/3, and so on, but the transition to algebraic numbers, for example, $\sqrt{2}$ as a root of $X^2 - 2 = 0$, was a major conceptual step beyond them. There are important numbers, like π, one-half the circumference of a circle of radius 1, that are not algebraic, but they are not our concern here. It is not immediately clear, but the sum of two algebraic numbers is algebraic as is their product. So is their quotient if the denominator is not 0. A remarkable and fundamental tneorem, although from the correct optic not difficult to prove, is that if we admit as coefficients a_i of the equation (1.1) algebraic numbers the roots are still algebraic.

It is not easy for a modern reader to understand the later books, for example Book X, a major contribution to the study of algebraic numbers in antiquity, of Euclid without a commentator, although I would guess that it is quite likely that much of the commentary places the mathematics in a false modern context. It already requires a great deal of imagination to penetrate the conceptual predilections of mathematicians of the previous century. For those of two thousand or more years ago an almost impossible effort is required to which, perhaps, the rewards are, for most of us, not commensurable.

As the English translator and editor Thomas Heath explains in the notes to his famous English translation, we are dealing in Book X with roots of the equations

$$X^2 \pm 2\alpha X \pm \beta = 0 \tag{2.2}$$

and

$$X^4 \pm 2\alpha X^2 \pm \beta = 0. \tag{2.3}$$

The roots of the second equation are the square roots of those of the first, which are $\pm\alpha \pm \sqrt{\alpha^2 \mp \beta}$, only those for which $\alpha^2 \mp \beta \geqslant 0$ having a meaning for Euclid. The coefficients

α and β are allowed by Euclid to be rational numbers m/n, m and n positive integers, or simple surds $\sqrt{m/n}$. Such equations arose in various geometric problems of Euclid's time and later. Book X is to some extent a classification, very elaborate, of the solutions and a discussion of their nature, rational or irrational. A quite sophisticated possibility is, expressed in modern terms, thus algebraically, is

$$\frac{\lambda^{1/4}}{\sqrt{2}}\sqrt{1+\frac{k}{\sqrt{1+k^2}}} \pm \frac{\lambda^{1/4}}{\sqrt{2}}\sqrt{1+\frac{k}{\sqrt{1+k^2}}}, \tag{2.4}$$

in which λ and k are simple fractions. The nature of this number depends on the nature of λ and k.

This number is the solution to Proposition 78 of Book X, given in Heath's translation as the following problem:

> *If from a straight line there be subtracted a straight line which is incommensurable in square with the whole and which with the whole makes the sum of the squares on them medial, twice the rectangle contained by them medial, and further the squares on them incommensurable with twice the rectangle contained by them, the remainder is irrational; ...*

I have omitted a brief phrase at the end because it is irrelevant. This is a complicated statement that needs explanation. Even after its meaning is clear, one is at first astonished that any rational individual could find the statement of interest. This was also the response of the eminent sixteenth century Flemish mathematician Simon Stevin, reported by Heath as writing, "La difficulté du dixiesme Livre d'Euclide est à plusieurs devenu en horreur, matière trop dure à digérer, et en la quelle n'aperçoivent aucune utilité." In fact, again according to Heath's commentary, a great deal of attention was given to the book by the early algebraists, who studied in the later Middle Ages and the early Renaissance the nature of the solutions of equations of the third and fourth degree, thus equation (1) with $n = 3$, 4. The theory of Galois, which cannot be considered a complete theory, appeared only after several more centuries of reflection on equations of higher degree by major mathematicians. It is probably the principal source of the notion of a group, an indispensable ingredient, for reasons not always transparent, perhaps not even yet understood, not only of many branches of mathematics but also of much of physics, both relativity and in quantum theory. So, in spite of the explanation to follow, there is no question of dismissing Euclid's Proposition 78 or any of the large number of similar propositions in Book X as frivolous distractions of misguided savants. There is no need to examine the explanation in detail. Enough can be learned for the present purposes simply by glancing at the formulas.

Two lines of length τ and τ' are incommensurable in square if their quotient squared is not a rational number. Two areas are commensurable if their quotient is rational, otherwise incommensurable. A line (read length of a line) is medial if its fourth power is rational but its square is not, a definition arising from geometrical considerations, and the area of a square is medial if its side is medial. So the proposition begins with two numbers τ, the whole, and σ, the part to be subtracted, such that σ^2/τ^2, $\tau^2+\sigma^2$, $2\tau\sigma$ are not rational numbers but $(\tau^2+\sigma^2)^2$ and $4\tau^2\sigma^2$ are. The final hypothesis is that the quotient

$$\frac{\tau^2+\sigma^2}{2\tau\sigma} \tag{2.5}$$

is irrational. The conclusion of the proposition is, in spite of the language, that $(\tau-\sigma)^2$ is irrational because the word *irrational* as used in Heath's translation means not irrational in the current sense but with an irrational, again in the current sense, square. Since

$$(\tau-\sigma)^2=\left(\frac{\tau^2+\sigma^2}{2\tau\sigma}-1\right)2\tau\sigma, \qquad (\tau-\sigma)^4=(\tau^2+\sigma^2)^2-4(\tau^2+\sigma^2)\tau\sigma+4\tau^2\sigma^2,$$

the first and the last terms on the right of the second equality are rational and the middle term is the product of $(\tau^2+\sigma^2)/2\tau\sigma$ and $8\tau^2\sigma^2$. By hypothesis the first of these numbers is irrational and the second rational. So the sum equal to $(\tau-\sigma)^4$ is irrational and, as a consequence, so is $(\tau-\sigma)^2$.

We are struck here, on the one hand, by the importance the ancient Greeks attached to the distinction between rational and irrational, on the other, by the simplicity, or rather triviality, of the argument, but that is not evidence for our superior intelligence, but for the advantages of the algebraic formalism. For a modern student at least, the geometric arguments of Euclid are far from transparent. A basic construct in Book X is the *apotome* ("a cutting off"), expressed algebraically, basically a difference $\alpha-\beta$ where α is rational and β is a proper surd $\beta=\sqrt{\gamma}$, γ rational, β not. An example would be $1-\sqrt{2}$. It is proved in Book X as Proposition 73 that an apotome $\delta=\alpha-\beta$ is necessarily irrational. This seems clear to us because $\beta=\alpha-\delta$ and if α and δ were both rational, β would be too. The argument of Euclid is, however, quite long. Namely, as a consequence of the assumptions and of some basic, but for us formal, propositions $\alpha^2+\beta^2$ is commensurable with α^2, $2\alpha\beta$ is commensurable with $\alpha\beta$ and as a consequence of the assumption α^2 and $\alpha\beta$ not commensurable. Therefore $(\alpha-\beta)^2=(\alpha^2+\beta^2)-2\alpha\beta$ is the difference of two incommensurable quantities, one of which is irrational, and therefore irrational. Proposition 78 is proved as a consequence of Proposition 73 and, like it, is for us astonishingly elaborate.

I find the commentary of Heath somewhat imprecise, but the algebra is clear. Let $\lambda=(\tau^2+\sigma^2)^2$. It is rational. We suppose with Euclid that it is not a square. It is easy to see that for an a, $0<a<1$,

$$\tau=\frac{\lambda^{1/4}}{\sqrt{2}}\sqrt{1+a}, \qquad \sigma=\frac{\lambda^{1/4}}{\sqrt{2}}\sqrt{1-a}.$$

since

$$4\tau^2\sigma^2=\lambda(1-a^2), \qquad \frac{\sigma^2}{\tau^2}=\frac{1+a}{1-a},$$

a is not rational but a^2 is. Since $0<a<1$, we may write

$$a^2=\frac{l}{1+l}$$

with l rational. Let $k^2=l$, so that

$$\tau=\frac{\lambda^{1/4}}{\sqrt{2}}\sqrt{1+\frac{k}{\sqrt{1+k^2}}}, \qquad \sigma=\frac{\lambda^{1/4}}{\sqrt{2}}\sqrt{1-\frac{k}{\sqrt{1+k^2}}}.$$

It is the difference

$$(\tau-\sigma)^2 = \frac{\lambda^{1/2}}{2}\left(1+\frac{k}{\sqrt{1+k^2}} - 2\sqrt{1-\frac{k^2}{1+k^2}} + 1 - \frac{k}{\sqrt{1+k^2}}\right)$$
$$= \lambda^{1/2}\left(1-\frac{1}{\sqrt{1+k^2}}\right)$$

that Euclid asserts is irrational. This is so, but Heath's commentary suggests that k will be rational and I see no reason that this is necessarily the case. If it is, then $\tau-\sigma$ is a number of the form (2.4).

From the point of view of Galois theory, what is being examined is less the particular number (2.4) but rather a collection of numbers, called a field, formed by taking a succession of square roots and then taking all possible sums of all possible quotients of all possible products of these roots, and possibly repeating this process. In equation (2.4), we have taken the square root of λ to obtain $\sqrt{\lambda}$, then the square root of this number divided by 2 to obtain $\lambda^{1/4}/\sqrt{2}$. But we have also taken the square root of $1+k^2$, then formed the number $1+k/\sqrt{1+k^2}$, taken its square root, and then taken a sum—as well as a difference—of products of the numbers obtained.

There is another theme in Book X, implicitly emphasized by the equations (2.2) and (2.3) of Heath's commentary, namely, when for two integers α and β is the sum $\alpha^2+\beta^2$ or difference $\alpha^2-\beta^2$ also a square? This question is closely related to equation $\alpha^n+\beta^n=\gamma^n$ appearing in the famous Theorem of Fermat, in fact a problem, proposed in the seventeenth century and solved in the twentieth. It is answered in Book X, where its presence suggests a different optic than that I have stressed: not a structural analysis of the irrationalities generated by geometrical problems as in Galois theory, but simply an analysis of the dichotomy, rational-irrational, for the solutions.

The structural analysis, of the thirteenth to sixteenth century for equations of degrees three and four, of the seventeenth and eighteenth century, with the first attempts by Lagrange and others to reflect on the relations between the roots of any given equation and to stress the importance of these relations and with the sudden, largely unexpected—there were antecedents, but so far as I have been able to understand, there is no reason to think that any of his predecessors were even on the verge of his discoveries—success of a very young Gauss with the equation $X^n-1=0$, whose roots are $e^{2\pi ij/n}$, $j=0,1,2,\dots,n-1$, was perhaps not in any strong sense implicit in Euclid, nor was the final comprehensive theory of Galois. The problems and the notions of Euclid's time were modified over the years but the seeds were sown then.

In the theory of Galois, the central issue is the collection of relations between the n roots of the equation (1.1) with coefficients taken from some prescribed collection. If, for example, all the coefficients are rational, the interest is in relations with rational coefficients. Consider for example the two equations,

$$X^2-2=0, \qquad X^2-4=0.$$

The first has two irrational solutions $X_1=\sqrt{2}$, $X_2=-\sqrt{2}$. The second has two rational solutions $X_1=2$, $X_2=-2$. A relation in both cases is an equality with 0 of a sum of products of powers of X_1 and X_2 with rational numbers, for example

$$X_1^2-2=0, \qquad X_2^2-2=0, \qquad X_1X_2+2=0,$$

are three relations for the first equation, while

$$X_1^2 - 4 = 0, \qquad X_2^2 - 4 = 0, \qquad X_1X_2 + 4 = 0.$$

are relations for the second. In addition, for the second equation, we have

$$X_1 - 2 = 0, \qquad X_2 + 2 = 0.$$

It is clear that there are no such simple relations for the first equation, for otherwise its roots would be rational. Another way of expressing this is that, for the first equation, we may replace in any valid relation—with rational coefficients!—X_1 by X_2 and X_2 by X_1 without invalidating it. This is clearly not so for the second. Thus there is a symmetry for the first equation missing for the second. From the point of view of an observer familiar only with rational numbers, and in some sense that is all of us, the two roots of the first equation may be different but they are also indistinguishable, while the roots of the second equation are different and distinguishable.[5] One is 2; one is -2. They add nothing new to the collection of rational numbers. For the first equation, however, the numbers $aX_1 + b = a\sqrt{2} + b$ are all different and, if $a \neq 0$, irrational. The number X_2 appears among them as $-X_1$. This means that the collection of all numbers generated by X_1 and X_2, thus the numbers of the form $a + bX_1$ with a and b rational, for X_2 can be replaced in any expression, no matter how complex, by $-X_1$ and X_1^2 by 2, is a larger domain (field is the better word here) than just that of the rational numbers. To have, at least conceptually, a square root of $\sqrt{2}$ we have to imagine, introduce, or construct a larger domain or field of numbers and in it, at least if the old norms and customs are maintained, there are necessarily two different roots. In the domain, and again from the point of view of the norms imposed by the customs or formalities of addition and subtraction there are two different but indistinguishable square roots of 2. More generally $a - bX_1 = a + bX_2$ is different from $a + bX_1$ but indistinguishable from it. So there is a *symmetry* of the domain $X_1 \to X_2 = -X_1$ which entails $a + bX_1 \to a - bX_1$.

Another example would be.

$$X^4 - 5X^2 + 6 = (X^2 - 2)(X^2 - 3) = 0.$$

The roots are $X_1 = \sqrt{2}$, $X_2 = -\sqrt{2}$, $X_3 = \sqrt{3}$, $X_4 = -\sqrt{3}$. We have already observed that there is a simple relation between X_1 and X_2. We might ask whether in the domain generated by X_1, thus in the collection of numbers $aX_1 + b$, there is already a square root of 3. Thus can

$$3 = (a + bX_1)^2 = a^2 + 2abX_1 + b^2X_1^2.$$

If $a = 0$ or $b = 0$ such an equation is impossible because of the theorem of Theaetetus. If $b = 0$ it entails $3 = a^2$; if $a = 0$ it entails $6 = 4b^2$. Both equations are impossible. If $ab \neq 0$, then it implies

$$X_1 = \frac{3 - 2a^2 - b^2}{2ab},$$

but that too is impossible for X_1 cannot be rational. Thus the introduction of X_3 entails the existence of a genuinely new domain larger than the collection of numbers $a + bX_1$. A little reflection establishes that this number is the domain $a + bX_1 + cX_3 + dX_1X_3$, which contains the square roots of 6 as $\pm X_1X_3$. It has three symmetries, even four,

$$X_1 \to \alpha X_1; \qquad \beta X_3; \qquad X_1X_3 \to \alpha\beta X_1X_3,$$

in which α and β can be independently taken to be ± 1. When they are both taken to be 1, the symmetry is not of great interest: objects are indistinguishable from themselves. It is useful nevertheless to introduce it for formal reasons.

These symmetries entail a permutation of the roots of the equation. Indeed they are in effect defined by a permutation of the roots, a kind of musical chairs in which the number of chairs is equal to the number of players, and in which the nature of the mixing of players and chairs at each round is circumscribed by perhaps complex rules that may none the less allow for madder scrambling than that customarily permitted. Permutations can be composed to form what mathematicians call a group: the result of two different rounds could clearly be the result of a single round. For many centuries equations of the form (1.1) were a puzzle. Could their roots be obtained by successive extractions of surds, thus with the help of numbers of the form $\sqrt[n]{a}$, with a rational or at least in the field generated by the coefficients and with n perhaps greater than 2, but nevertheless with a result along the lines of equation (2.4), typical for the problems investigated in Book X of Euclid? Galois discovered that the decisive factor in the answer was the nature of the group of symmetries, thus of the rules governing their composition. Have mathematicians abused his discovery? Since one notion appearing in Galois's brief writings is that of a group, which is of universal importance, not alone for the theory of equations but for mathematics in general, this question may appear frivolous. To understand its sense demands some background.

A.3. Algebraic numbers at the beginning of the nineteenth century. Carl Friedrich Gauss, born in 1777, died in 1855, was a mathematician on whose contributions I have never dared to reflect seriously. He was a prodigy, but in his case the troubling connotations, almost universal among prodigious mathematicians, of early, forced efflorescence, followed by a premature withering were absent. The child of parents in modest circumstances, his talent was early remarked and he was fortunate enough to receive the support of the Duke of Brunswick.[6] Thanks to the excellent education this permitted and to Gauss's innate talent, he was at the age of eighteen, when he published in the *Allgemeine Literaturzeitung*, a journal founded by, among others, Goethe and Schiller and not so dissimilar to the contemporary New York Review of Books, his discovery of the possibility of the construction with only ruler and compass of a regular heptadecagon (a polygon with 17 sides of equal length and with all adjacent sides meeting at the same angle), already a mature mathematician. Whether such a construction was possible was a problem that had remained unsolved for two thousand years, longer than the impossibility of duplicating the cube or trisecting a general angle, both difficult but easier problems to which it is none the less similar, and almost as long as the question of the squaring of the circle, which is analyzed by quite different considerations.

Gauss's career began on the threshold of the nineteenth century, and it is instructive to begin a discussion of it with a brief description of two key contributions that influenced the mathematics of that century in very different ways. I have found over the course of the years, without making any particular effort, that the more one penetrates Gauss's writings the closer one comes to mathematics. His early writings manifest youthful force and the pleasure in detail it allows and at the same time a mature, critical understanding of the contributions of his predecessors and the problems they faced. Whether he was always just to them, I cannot say, but his primary concern seems to have been not to seize fame but to grasp the essence of a question.

I observe in passing that attention to detail is not a luxury even when occupied with the most abstruse questions, but contemporary mathematicians, including those of my generation,

usually start late and have a great deal to master. So willy-nilly we remain superficial. I see no obvious remedy except too sharp a focus and that is worse.

Gauss's *Disquisitiones Arithmeticae*, published in 1801 when he was 23, was the foundation and the inspiration of algebraic number theory, one of the great nineteenth-century achievements of mathematics and was the source not only of Galois's inspirations but of the quite different contributions of a number of the century's great number theorists. The seventh and last chapter was the beginning of the theory of *cyclotomy*, thus of the division of the circumference of a circle into n equal parts and, thus, the construction of regular polygons with n sides. It is a question whose relation to beauty is immediate and, if one were so inclined and had the necessary competence, it would be appropriate at this conference to devote a full lecture to it, but I introduce it only as one aspect of a much larger question.

The geometric problem, which is to construct these figures with a ruler and compass, had already been considered by the Greeks and is the principal topic of Euclid's Book IV, where regular polygons of three, four, five, six, and fifteen sides are treated. The geometric problem is, at its core, algebraic. If we allow ourselves the freedom of using complex numbers, thus numbers $x + y\sqrt{-1} = x + yi$ represented by points (x, y) in the plane, then the vertices of a regular polygon of n sides inscribed in a circle of radius r with center at the origin of coordinates are represented by the complex numbers $\cos(2\pi/k) + \sin(2\pi/k)i$, $k = 1, 2, \ldots, n-1$. Taking the radius as the unit of length, these are for a triangle $1 + 0 \cdot i = 1$, $\cos(2\pi/3) + \sin(2\pi/3)i = -1/2 + \sqrt{3}i/2$, $\cos(2\pi/3) + \sin(2\pi/3)i = -1/2 - \sqrt{3}i/2$, because $\cos(2\pi/3) = -1/2$ and $\sin(2\pi/3) = \sqrt{3}/2$. As points in the plane these are $(1, 0)$, $(-1/2, \sqrt{3}/2)$, $(-1/2, -\sqrt{3}/2)$. For a pentagon, we take again one of the vertices to be $(1, 0)$, then the others are

$$
\begin{aligned}
\big(\cos(2\pi/5), \sin(2\pi/5)\big) &= \left(\frac{-1+\sqrt{5}}{4}, \sqrt{\frac{5}{8} + \frac{\sqrt{5}}{8}}\right) \\
\big(\cos(4\pi/5), \sin(4\pi/5)\big) &= \left(\frac{-1-\sqrt{5}}{4}, \sqrt{\frac{5}{8} - \frac{\sqrt{5}}{8}}\right) \\
\big(\cos(6\pi/5), \sin(6\pi/5)\big) &= \left(\frac{-1-\sqrt{5}}{4}, -\sqrt{\frac{5}{8} - \frac{\sqrt{5}}{8}}\right) \\
\big(\cos(8\pi/5), \sin(8\pi/5)\big) &= \left(\frac{-1+\sqrt{5}}{4}, -\sqrt{\frac{5}{8} + \frac{\sqrt{5}}{8}}\right).
\end{aligned}
\tag{3.1}
$$

These two regular polygons appear in Figures I and II. We see in (3.1) a phenomenon with which we are familiar: with the expressions in which there is a square root appear simultaneously the expressions in which the sign of this square root is reversed. This is a manifestation of the Galois symmetry among the roots of the algebraic equation satisfied by the complex numbers yielding the vertices of the pentagon, namely

$$(X-1)(X^4 + X^3 + X^2 + X + 1) = X^5 - 1 = 0.$$

The same phenomenon for the triangle is less striking because the equation is

$$(X-1)(X^2+X+1) = X^3 - 1 = 0,$$

and we have perhaps seen too many quadratic equations.

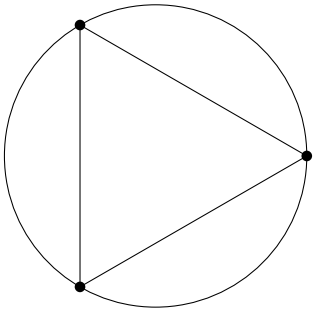

Figure I

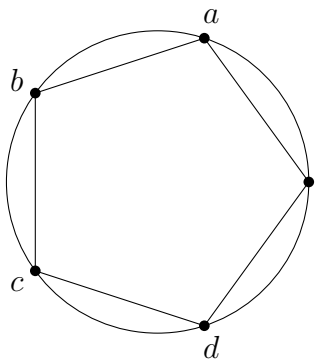

Figure II

Nevertheless it is best when beginning to reflect on the symmetries of the two equations

$$X^4 + X^3 + X^2 + X + 1 = 0, \qquad X^2 + X + 1 = 0,$$

to consider them simultaneously. The algebraic symmetries are not the geometric symmetries, thus not the obvious rotational symmetries of the triangle or the pentagon. For the triangle, the algebraic symmetry is the absolute indistinguishability from an algebraic standpoint of the two roots $-1/2 \pm \sqrt{3}i/2$. The symbol i is just that, a symbol, thus a convention that we have learned to accept as having a real meaning, as it may have, but in its intrinsic mathematical properties it is indistinguishable from $-i$. For the quadratic equation $X^2 + X + 1 = 0$ one might dismiss this distinction as pedantic logic-chopping, but for the first equation, that of degree four, the significance is not to be denied. There are five rotational symmetries for the pentagon, those through angles $2\pi/5$, $4\pi/5$, $6\pi/5$, $8\pi/5$ together with the trivial rotation, which is always present. There are also five reflections that preserve the pentaton. There are only four algebraic symmetries and they cannot easily be represented geometrically. The four roots of $X^4 + X^3 + X^2 + X + 1 = 0$ are listed in (3.1): $a = \cos(2\pi/5) + i\sin(2\pi/5)$; $b = \cos(4\pi/5) + i\sin(4\pi/5)$; $c = \cos(6\pi/5) + i\sin(6\pi/5)$; $d = \cos(8\pi/5) + i\sin(8\pi/5)$. These four roots determine a *field* of numbers, those numbers that can be formed by adding rational multiples of these four,

$$\alpha a + \beta b + \gamma c + \delta d,$$

where the coeffecients α, β, γ, δ are rational numbers. As a simple example we could take $\alpha = \beta = \gamma = \delta = 1$, but that yields $a + b + c + d$ which is easily verified to be -1. Different coefficients lead to different numbers and it is difficult to gain any geometric insight into their totality. It is however easily verified that the sum or the difference of two such numbers is again such a number, as is the product or a quotient. It is this possibility of forming sums, differences, products and quotients within a given collection of numbers that mathematicians express by referring to the collection as a field, a word that we have used more than once already but without making its meaning precise. The Galois symmetries are symmetries of the field. For example, if a is replaced by $a_1 = b$, then b is replaced by $b_1 = d$, c by $c_1 = a$, and d by $d_1 = c$. All algebraic relations are preserved, $a^2 = b$ entails $a_1^2 = b_1$ or $ab = c$ entails $a_1 b_1 = c_1$. The consequences of this are not immediately apparent.

We can guess, if we have any familiarity with complex numbers, that for other possible integers n, the equation

$$(X-1)(X^{n-1}+\cdots+X+1)=X^n-1=0$$

is pertinent to the construction of a regular polygon with n sides. It is best when first examining the equation to treat prime numbers n, thus $n=3,5,7,11,13,17,\dots$. There are again n solutions, of which the first, $1+0\cdot i$, or, in the plane, the point $(1,0)$ corresponds to the arbitrarily assigned first vertex. The problem is to construct the others once it is given, thus to construct geometrically the angle $2\pi/n$. That depends on the symmetries, or, as we now would say, the Galois symmetries of the $n-1$ roots, $\theta_i=\cos(2\pi k/n)+\sin(2\pi k/n)i$ of the equation

$$X^{n-1}+\cdots+X+1=0. \tag{3.2}$$

At this point, I have to rely on you to recall from basic trigonometry two identities that are at the core of any treatment of equation (3.2),

$$\begin{aligned}\cos(2\pi k/n)\cos(2\pi l/n)-\sin(2\pi k/n)\sin(2\pi l/n)&=\cos\big(2\pi(k+l)/n\big),\\ \cos(2\pi k/n)\sin(2\pi l/n)+\sin(2\pi k/n)\cos(2\pi l/n)&=\sin\big(2\pi(k+l)/n\big).\end{aligned} \tag{3.3}$$

As a consequence $\theta_k\theta_l=\theta_{k+l}$ if $k+l<n$. It is 1 if $k+l=n$ and θ_{k+l-n} if $k+l>n$. It has to be proven, but from an algebraic point of view all the numbers θ_k are indistinguishable. In so far as only equations with rational coefficients are admitted, they satisfy exactly the same equations. Moreover all of them can be represented as algebraic expressions of any given one, for example, as a consequence of the identities (3.3), $\theta_j=\theta_1^j$, or if $l-kj$ is a multiple of n, $\theta_l=\theta_k^j$. So, in principle, we understand the Galois theory of the equation (3.2), as Gauss did before Galois was born. The symmetries, in other words the permutations of the roots that preserve all relations between them, are determined by $\theta_1\to\theta_k$. Thus θ_1 is replaced by θ_k, where k can be any integer from 1 to $n-1$, and then θ_j is necessarily replaced by θ_1^l, where $l-kj$ is divisible by n.

We seem to be losing our way. The initial question, a question that presumably confronted the Hellenistic mathematicians but was abandoned by them as too difficult, was whether for other integers than those n appearing in Book IV of Euclid, it is possible to construct with ruler and compass, thus by the successive extraction of square roots, regular polygons with n sides. For which n can $\theta_1=\cos(2\pi/n)+\sin(2\pi/n)i$ be represented as an expression similar to those in (3.1). This imposes some conditions on the possible algebraic symmetries of the equation, which form what came to be called the Galois group. Each time we add a square root, thus each time we perform a construction with ruler and compass, we enlarge the field generated by the roots, and as a result of the general theory there will be additional symmetries, those already there together with one that replaces the newly adjoined square root by its negative. Thus when concretely expressed the elements of the Galois group will have to appear simply as sign changes. This means, to make a long story short, that the total number of permutations will be a power of 2. We have seen that for the field of numbers of concern to us the number of permutations is $n-1$. It does not happen very often for n a prime that this number is a power of 2. It is true for $n=3$, $n-1=2$, and, as the equations (3.1) make explicitly clear, also for $n=5$, $n-1=4=2^2$. The next possibility is $n=17$, $n-1=16=2^4$, and the consequences were understood only after 2000 years. As one infers from its publication in the *Allgemeine Literaturzeitung* the possibility of constructing a regular heptadecagon was expected to astound not only mathematicians but the entire

educated world. It astonishes mathematicians to this day. Much as I admire the achievement and the style of Galois, and troubled as I am by any exaggerated admiration of Gauss, I find that any judgement of the nature of Galois's achievement has to be preceded by a sober reflection on what Gauss had already accomplished.

A complete analysis of the Galois theory of the equation

$$X^{16}+X^{15}+X^{14}+X^{13}+X^{12}+X^{11}+X^{10}+X^9+X^8+X^6+X^5+X^4+X^3+X^2+X+1=0,$$

important because

$$(X-1)(X^{16}+X^{15}+X^{14}+X^{13}+X^{12}+X^{11}+X^{10}+X^9+X^8+X^6+X^5+X^4+X^3+X^2+X+1)$$

is equal to $X^{17}-1$, would be out of place. A basic root is $\theta = \cos(2\pi/17) + \sin(2\pi/17)i$. The symmetries are $\theta \to \theta^k$, $k = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16$. There are sixteen of them. The first, for $k = 1$, is simple, $\theta \to \theta$. The second

$$\theta \to \theta^2, \quad \theta^2 \to \theta^4 \quad \theta^4 \to \theta^8, \quad \theta^8 \to \theta^{16},$$

and continuing, because, for example $\theta^{16} = \theta^{-1}$ and the square of θ^{-1} is $\theta^{-2} = \theta^{15}$,

$$\theta^{16} \to \theta^{15}, \quad \theta^{15} \to \theta^{13} \quad \theta^{13} \to \theta^9, \quad \theta^9 \to \theta,$$

which gives only eight in all because we return to the initial point. If we had started with $\theta \to \theta^3$, which is missing from this list of eight, we would have

$$\begin{aligned} \theta \to \theta^3 \to &\theta^9 \to \theta^{10} \to \theta^{13} \to \theta^5 \to \theta^{15} \to \theta^{11} \to \theta^{16} \\ \to &\theta^{14} \to \theta^8 \to \theta^7 \to \theta^4 \to \theta^{12} \to \theta^2 \to \theta^6 \to \theta \end{aligned} \tag{3.4}$$

We return at the end to the initial point, generating on the way all sixteen possible roots. The group of symmetries has 16 elements and is generated by a single one, that which replaces θ by θ^3.

We can construct from these diagrams, step by step, combinations of the roots that are less and less symmetric. A very symmetric combination is obtained by adding together all the roots. It is equal to -1. If we just add together every second root in (3.4), we obtain

$$\theta + \theta^9 + \theta^{13} + \theta^{15} + \theta^{16} + \theta^8 + \theta^4 + \theta^2$$

or

$$\theta^3 + \theta^{10} + \theta^5 + \theta^{11} + \theta^{14} + \theta^7 + \theta^{12} + \theta^6.$$

These numbers look very complicated but a patient calculation and appropriate use of the equation

$$X^{16} + X^{15} + \cdots + X^2 + X + 1 = 0$$

satisfied by θ establish that the first is equal to $(-1+\sqrt{17})/2$ and the second to $(-1+\sqrt{17})/2$.

We can continue with the combinations

$$\theta + \theta^{13} + \theta^{16} + \theta^4, \qquad \theta + \theta^{16}, \qquad \theta,$$

each in some sense one-half as symmetric as the previous one, with respect to which it satisfies a quadratic equation, arriving finally at the expression

$$-\frac{1}{16} + \frac{1}{16}\sqrt{17} + \frac{1}{16}\sqrt{34-2\sqrt{17}} + \frac{1}{8}\sqrt{17+3\sqrt{17}-\sqrt{34-2\sqrt{17}}-2\sqrt{34+2\sqrt{17}}},$$

for $\cos(2\pi/17)$. It is complicated with three telescoping square roots. The number

$$\sin(2\pi/17) = \sqrt{1-\cos^2(2\pi/17)}$$

requires one more square root. Thus we need four in all, together with that for $i = \sqrt{-1}$, to express θ. It is worthwhile to ask oneself how an eighteen-year old might have felt on writing down these equations, the solution of a problem more than two-thousand years old.

For reasons of convenience, Gauss wrote a thesis that contained neither this theorem nor any of the other major discoveries from which the modern theory of numbers developed. The central feature of the thesis is a proof of the fundamental theorem of algebra, which we have already encountered. I repeat, with a different emphasis, its enunciation, both in a modern form and in the form for which Gauss offered at the age of 20 a proof, sometimes considered as the first proof, although I think it is fairer to say that Gauss was, even though the initiator, just one of the architects of the proof or proofs, which ultimately rely on a very sophisticated understanding of the nature of a real number that was unavailable at the end of the eighteenth century. Gauss himself published over the course of several years four different proofs.[7]

The notion of a real number has an immediate intuitive basis in our notion of distance. So I take it as undefined, but in fact its formal mathematical definition demands considerable preliminary reflection. The real numbers available, a complex number is taken to be a formal object of the form $a + b\sqrt{-1}$, where a and b are arbitrary real numbers. Two equivalent statements of the fundamental theorem of algebra are that any polynomial with complex coefficients, the first being 1, factors as a product of linear polynomials

$$X^n + aX^{n-1} + \cdots + fX + g = (X - \alpha_1)(X - \alpha_2)\cdots(X - \alpha_n)$$

with complex numbers α_i or that any polynomial with real coefficients factors as a product of linear factors and quadratic linear factors, all coefficients being real,

$$X^n + aX^{n-1} + \cdots + fX + g = (X - \alpha_1)\cdots(X - \alpha_k)(X^2 - \beta_1 X + \gamma_1)\cdots(X^2 - \beta_l X + \gamma_l),$$

$k + 2l = n$. It is the second form that Gauss proved.

His first proof is of interest even today, simply as a mathematical proof taken by itself, regardless of the historical interest or scientific implications of the theorem. Moreover it is preceded by a review of the proofs, or attempts at proofs, by earlier authors. He points out their errors systematically. One observation I find curious. Apparently earlier authors, who are not named, were willing to accept the possibility that a polynomial like

$$X^n + aX^{n-1} + \cdots + fX + g = 0$$

had fewer than n complex roots, even if they were counted with the correct multiplicities. They were unwilling to abandon the missing roots and proposed to introduce them—somewhat along the lines, I suppose, of $i = \sqrt{-1}$—as *impossible* roots. Gauss found, for reasons I cannot understand, such a step inadmissible. We know today that if the fundamental theorem of algebra were false it would be possible, and certainly legitimate, to include within the formalism of mathematics a larger collection of "numbers" that permitted the desired factorization of polynomials. The fundamental theorem is true however. So this larger collection is unnecessary. No new numbers are necessary to account for the *impossible* roots.

What if the fundamental theorem had been false? Although he was, almost two hundred years before Gauss, in no position to express himself clearly, for Descartes the fundamental theorem of algebra is a fundamental theorem of geometry. The number of points at which two curves each of some fixed algebraic form, for example two conics, meet depends only on these forms. This is a consequence of the fundamental theorem of algebra. If the fundamental theorem of algebra were false, there would be the "true" Cartesian geometry, in which there

were only complex numbers, but in which Descartes's principle was false, and a formal, algebraic one, in which it was valid. The true one would be the one deduced from geometry and from our geometric notions based on the relation between number and distance; the formal one would be one in which the impossible numbers were granted legitimacy. However, Descartes's principle,[8] thus the fundamental theorem of algebra as the fundamental theorem of geometry, is intimately linked to the possibility of introducing topological invariants into geometry. We now know, thanks to an insight of André Weil in 1949 and to the very elaborate theory of *algebraic* geometry created by many, many practitioners over the nineteenth and twentieth centuries, that their introduction would have also been possible, at least to some extent, in the formal, algebraic theory alone, without reference to the concrete geometry informed by our intuitive notions of distance and number, but this would have been a difficult route to take. Fortunately it was not necessary. Pure mathematicians were permitted for a very long time to rely on their geometric intuition and still are. Some preparation is necessary before we can begin to appreciate the—strange and still mysterious—modern relation between algebra and geometry, much more abstruse than in Euclid.

A.4. Algebraic numbers during the nineteenth century. Thanks to the influence not only of his study of cyclotomy but also of the many other varied, but all basic, early contributions of Gauss to the solution of outstanding unsolved problems of late eighteenth-century number theory, a general theory of algebraic numbers was created in the course of the nineteenth century, initially in order to search, in part successfully, for a proof of the famous theorem of Fermat, to the effect that for $n > 2$, there are no integral solutions to the equation

$$X^n + Y^n = Z^n, \qquad XYZ \neq 0. \tag{4.1}$$

It was observed, in particular by the German mathematician E. E. Kummer, whose central concern it became, that the problem is related to a notion already appearing in connection with Theaetetus, that of prime and that of the unique decomposition of an arbitrary number into primes, thus $30 = 2 \times 3 \times 5$. For example, when n is a prime the possibility of unique factorization into primes in the domain of numbers

$$a_0 + a_1\theta + \cdots + a_{n-2}\theta^{n-2} + a_{n-1}\theta^{n-1}, \quad \theta = \cos\left(\frac{2\pi}{n}\right) + i\sin\left(\frac{2\pi}{n}\right), \tag{4.2}$$

where the coefficients $a_0, a_1, \dots$ are taken to be rational, is the key to the impossibility of finding a solution to (4.1). It turns out, however, that for some n this is possible, and for some n it is not. This entailed complications and decades of effort by some very strong minds.[9]

Rather than continuing with cyclotomic domains, I describe a simple example taken from a presentation of the basic ideas of the theory of algebraic numbers by one of its creators, Richard Dedekind. To study the solution in integers or in rational numbers of an equation like (4.1), referred to for historical reasons as diophantine equations, it is usually imperative to enlarge the possibilities and to study it not only in domains given by algebraic numbers, like that given by the numbers (4.2), or, as a simpler example, by the numbers $a + b\sqrt{-5} = a + b\theta$, a, b rational, $\theta^2 = -5$. For this one needs to be able to factor them into prime factors, more precisely, just as it is ordinary integers and not fractions that we customarily factor, it is best to consider the factorization of elements in the new domain that can be regarded as analogues of ordinary integers. The appropriate definition is subtle but, in the present case,

the result is simple. A number of the form $\alpha = a + b\theta$, a, b rational, is to be regarded as integral if a and b are integers. These factors have now often to be *ideal* prime factors, but the factorization has still to be unique. Let $\mathfrak{o}$ be the collection of integral elements $a + b\theta$.

Dedekind considers the question of whether there are primes in this domain and whether every number can be factored in an essentially unique way into a product of primes. As we shall see it cannot. To deal with the consequent difficulties, in particular, to undertake the investigation of Fermat's theorem along the lines proposed by Kummer, it is necessary to introduce the notion of *ideal* primes and, to some extent as a terminological convenience, ideal numbers. In spite of the name, usually abbreviated to *ideal* in the singular and *ideals* in the plural, these are objects that exist within the accepted mathematical formalism and that are of great use not only in number theory but in other domains as well, in particular in geometry. They are not, however, numbers! The mathematical concept of ideal is now commonplace, but it was difficult to arrive at.

Notice that $(a + b\theta)(c + d\theta) = (ac - 5bd) + (bc + ad)\theta$ is a number of the same form. Thus we can, indeed, take products of these numbers. We introduce, as an aid to our reflections the norm of these numbers, $\mathrm{N}\,\alpha = a^2 + 5b^2$, $\alpha = a + b\theta$, noting that $\mathrm{N}\,\gamma = \mathrm{N}\,\alpha\,\mathrm{N}\,\beta$ when $\gamma = \alpha\beta$ and that $\mathrm{N}\,\alpha = 1$ only for $\alpha = \pm 1$. Just as for ordinary integers, an element α of $\mathfrak{o}$ may or may not be decomposable as a product $\alpha = \beta\gamma$. Such a decomposition is interesting only if neither β nor γ is ± 1. At first glance it is tempting to take a number to be prime if it is not decomposable.

Dedekind considers, with this in mind, the following numbers $a = 2$, $b = 3$, $c = 7$, and

$$\begin{array}{llll} b_1 = -2 + \theta, & b_2 = -2 - \theta; & c_1 = 2 + 3\theta, & c_2 = 2 - 3\theta; \\ d_1 = 1 + \theta, & d_2 = 1 - \theta; & e_1 = 3 + \theta, & e_1 = 3 - \theta; \\ f_1 = -1 + 2\theta, & f_2 = -1 - 2\theta; & g_1 = 4 + \theta, & g_2 = 4 - \theta. \end{array}$$

For these numbers we have $\mathrm{N}\,a = 4$, $\mathrm{N}\,b = 9$, $\mathrm{N}\,c = 49$, and for the norms of the others

$$\begin{array}{llll} 9, & 9, & 49, & 49; \\ 6, & 6, & 14, & 14; \\ 21, & 21, & 21, & 21. \end{array}$$

If any of these numbers factored as a product $\omega\omega'$, with neither ω nor ω' equal to ± 1, then necessarily $\mathrm{N}\,\omega > 1$ and $\mathrm{N}\,\omega' > 1$, and thus necessarily both of these numbers are equal to one of 2, 3, 7. These three numbers are clearly not representable as $x^2 + 5y^2$, with x and y integral. So we have several, namely 15, indecomposable numbers available, and as the following table indicates, we can find products among them that are equal. So unique factorization is out of the question.

$$\begin{array}{lll} ab = d_1 d_2, & b^2 = b_1 b_2, & ab_1 = d_1^2, \\ ac = e_1 e_2, & c^2 = c_1 c_2, & ac_1 = e_1^2, \\ bc = f_1 f_2 = g_1 g_2, & af_1 = d_1 e_1, & ag_1 = d_1 e_2. \end{array}$$

Dedekind observes, indeed it is evident, that all these equalities are compatible with the existence of factorizations,

$$a = \alpha^2, \qquad b = \beta_1\beta_2, \qquad c = \gamma_1\gamma_2;$$

$$b_1 = \beta_1^2, \qquad b_2 = \beta_2^2; \qquad c_1 = \gamma_1^2, \qquad c_2 = \gamma_2^2;$$
$$d_1 = \alpha\beta_1, \qquad d_2 = \alpha\beta_2; \qquad e_1 = \alpha\gamma_1, \qquad c_2 = \alpha\gamma_2;$$
$$f_1 = \beta_1\gamma_1, \qquad f_2 = \beta_2\gamma_2; \qquad g_1 = \beta_1\gamma_2, \qquad c_2 = \beta_2\gamma_1;$$

and then continues to show the existence of ideals (ideal numbers) that yield these identities, but the search for the correct concepts preoccupied Dedekind—and Kronecker—for a lifetime. Although over time, various mathematicians have expressed a preference for Kronecker's treatment, it appears to me that Dedekind's resolution of the problem—in which an ideal number is a set in the modern mathematical sense—has prevailed. The basic theory of algebraic numbers established in the last decades of the nineteenth century, entirely new problems were posed and a new theory proposed, now referred to as class-field theory, with decisive suggestions from Kronecker, Weber, and Hilbert, and then solved in the early decades of the twentieth by Fürtwangler, Takagi and Artin. These were all outstanding mathematicians, but it is probably the name of David Hilbert and his intimate involvement with the theory that offers to non-mathematicians or even to mathematicians with no special knowledge of algebraic numbers the most credible guarantee of its interest. The subject was, nevertheless, largely neglected in the early post-war years, partly because it had been sustained principally by German mathematicians and the mathematical strength of Germany did not survive the war, partly because of internal exhaustion. It is instructive to recall the words of Emil Artin, who had emigrated to the USA and who was not only a major contributor to the subject in the twenties but also, I should think, the mathematician principally responsible for introducing the subject to the new world, to the Princeton Bicentennial Conference on *Problems of Mathematics*, a gathering of very many, very distinguished mathematicians in 1946. After recalling a theorem of Richard Brauer and stating that it represents a decisive step in the "generalization of class-field theory to the non-Abelian case, which is commonly regarded as one of the most difficult and important problems in modern algebra", the report of the conference cites Artin, one of the participants, as observing, "My own belief is that we know it already, though no one will believe me—that whatever can be said about non-Abelian class field theory follows from what we know now, since it depends on the behaviour of the broad field over the intermediate fields—and there are sufficiently many Abelian cases."

In the next few lines, I shall describe the Abelian theory but only very briefly before turning to other domains, less abstruse and known to broader classes of mathematicians and scientists, where the influence of the theory of algebraic numbers has been decisive, but often forgotten, and which have had an essential influence on the beginnings of theories that are expected to contain the sought for non-Abelian theory. It is still premature, however, to attempt to assess the prophetic significance—or lack of it—in Artin's words. The immediate lesson is the importance or, better, the value in the search for theorems of some profundity or for the theories in which they are implicit of escaping the confines of any particular specialty.

At this point, the reader has perhaps enough experience to appreciate more precisely of what Galois's ideas consisted. We have introduced domains of numbers, technically referred to as fields, that are obtained by taking some fixed algebraic number θ, a root of an equation of the form (1.1) with rational coefficients $a_0, a_1, \dots$. It will then satisfy a unique such equation of lowest degree, say n and the fields are then the collection of numbers of the form $\alpha_0 + \alpha_1\theta + \alpha_2\theta^2 + \cdots + \alpha_{n-1}\theta^{n-1}$. As we have observed, the sum or difference of two numbers with such a representation again has one, as does the product, and the quotient provided the denominator is not 0. It is very advantageous to assign this domain, which is a collection

or set of numbers a label, say k or K. The first is contained in the second, $k \subset K$, if every element of the first is also an element of the second. A simple example of this is to take k to be the collection of numbers $a + bi$, $i^2 = -1$, a and b rational and K to be the numbers $a + b\theta + c\theta^2 + d\theta^3$, where

$$\theta = \cos \pi/4 + i \sin \pi/4 = 1/\sqrt{2} + i/\sqrt{2}.$$

The object of Galois theory is such pairs K/k. There is one kind of pair that is especially important in the context of that theory. K is called a Galois extension of k if when K contains η and the element μ is such that every equation of the form (1.1) with coefficients $a_0, a_1, \ldots$ from k that has η as a root also has μ as a root, then K also contains μ. For example, if k is the field of rational numbers and K is the field of numbers $a_0 + a_1\theta + \cdots + a_{16}\theta^{16}$, $\theta = \cos 2\pi/17 + \sqrt{-1} \sin 2\pi/17$ that plays a central role in the study of the heptadecagon then K/k is a Galois extension. K is also a Galois extension of the field k' formed from the numbers $a_0 + a_1\sqrt{17}$, a_0 and a_1 rational. This we discovered implicitly when examining Gauss's construction of the heptadecagon. The number μ is in general called a conjugate of η over k.

The basic affirmations of Galois theory as created by Galois are that if K/k is a Galois extension then we can find an element η of K and an integer n such that, first of all, every element of K can be written uniquely as a sum $\alpha_0 + \alpha_1\eta + \alpha_2\eta^2 + \cdots + \alpha_{n-1}\eta^{n-1}$, secondly that η has exactly n conjugates μ , including itself; thirdly that for each of these conjugates μ the correspondence

$$\phi : a_0 + a_1\eta + a_2\eta^2 + \cdots + a_{n-1}\eta^{n-1} \to a_0 + a_1\mu + a_2\mu^2 + \cdots + a_{n-1}\mu^{n-1}$$

preserves all algebraic relations. These correspondences can be composed and form what is called a group, the Galois group. What cannot be sufficiently emphasized in a conference on aesthetics and in a lecture on mathematics and beauty is that whatever beauty the symmetries expressed by these correspondences have, it is not visual. The examples described in the context of cyclotomy will have revealed this.

Galois theory became, together with the theory of ideals, a central guiding feature of the study of algebraic numbers in the nineteenth century and continues to inform a great deal of number theory, especially the theory of diophantine equations.

The Galois groups encountered in the theory of cyclotomy are, as we have seen in the examination of specific cases although without explicitly drawing attention to it, rather special. They are abelian, in the sense that when we compose any two of the correspondences to obtain a third, the order in which the correspondences are applied does not matter. A decisive observation of Leopold Kronecker was that, when k is the field of rational numbers itself, abelian Galois groups appear only for cyclotomic fields K and for fields K lying within them, thus only for those fields that arise in the study of regular polygons. The problem then was posed in various forms, which became increasingly precise, whether it would be possible over any field k of algebraic numbers to describe all the *abelian* Galois extensions. Although I have no intention of presenting it in detail here this description was achieved and in connection with an apparently unrelated phenomenon, the decomposition of primes in k in a larger field K, but only for K/k Galois with an abelian Galois group. This is a problem to which over the course of the nineteenth century Gauss, Dirichlet, Dedekind and Frobenius had contributed.

We can already see the phenomenon in Dedekind's example, for which the field k is the field of rational numbers and K the field of numbers $a + b\sqrt{-5}$ with a and b rational, a field that

arises in connection with the construction of a regular polygon with 20 sides and that is closely related to the field containing the numbers (3.1). Continuing the calculations of Dedekind, we could ask ourselves what ideals in k, thus what prime numbers $p = 2, 3, 5, 7, 11, 13, 17, 19, \dots$ are the products of two prime ideals in K in the sense for which this is true for 3 and 7 in Dedekind's explanatory calculations and which, like 11, 13, 17, 19, remain prime, in the sense that such factorizations are not possible. It turns out that, apart from 2 and 5 which are exceptional, the distinction is determined by the remainder left on division by 20. If it is 1, 3, 7 or 9 such a factorization is possible; if it is 11, 13, 17 or 19 it is not. The core of class field theory and the key to Kronecker's discovery, striking in its time and striking today, is that these two phenomena are intimately related.

Can the revelation of these arcane properties of numbers, their factorizations, in whatever sense, and their relation to the structure—abelian or non-abelian—of a group with no apparent visual appeal in any sense be considered beautiful? Of what value is it? It certainly lacks the immediate appeal of elementary geometry and arithmetic or indeed of simple algebra, which can fascinate even a child; it is also a far cry from the sharpening of our sensuous pleasure in fluid flow or other kinds of movement offered by its description in terms of differential equations and their solutions. On the other hand, these matters are less recondite than they at first seem and remain reasonably close to their origins in the reflections of eighteenth century mathematicians like Euler, Legendre and Gauss on more immediate properties of the integers, and are certainly essential for the understanding of Fermat's theorem, for a long time the favorite puzzle of amateur mathematicians and of many professionals. Class field theory and related matters have become the stock in trade of a very large number of professional mathematicians, so that the question of its value is an existential question, best avoided.

A.5. From algebraic number theory to spectroscopy. Up to this point, we have examined the development suggested by the second column of the diagram, although in a very superficial manner. It is a long development, reaching from time of Plato, or even Pythagoras to the early twentieth century. Now I want to move to the development suggested by the collection of topics on the right-hand side. It took place in a much shorter period from the middle of the nineteenth century to the end of the twentieth, and has not yet, I think, been digested by mathematicians, and for very reasons that are not difficult to grasp. The ideas of Galois, close as they were in some sense to ideas in Lagrange and Gauss, took most mathematicians, the majority of whom conform, then and now, to the classic German observation, *"Was der Bauer nicht kennt, das frißt er auch nicht,"* by surprise. They had, none the less, a tremendous influence on both Dedekind and on F. G. Frobenius, the two of whom, first of all, used them in conjunction with the theory of ideals created by Dedekind and Kronecker and with the theory of functions of a complex-variable to develop the notions ultimately employed in the creation of class field theory and, secondly, introduced the notions of group determinant and group representation, closely related to each other. For these two mathematicians, the group was finite, as in the theory of Galois, the group of permutations of the roots of an equation that preserved all algebraic relations between them. In the hands of others, notably Issai Schur, Élie Cartan, and Hermann Weyl, the notion of representation was extended to continuous groups, the most important example being the group of rotations about a given point in three-dimensional space. This group has dimension three. For example, if we are thinking of rotations of the earth about its center, we have first to decide to what point we shall rotate the north pole. To describe this point takes two coordinates. We have

then to add a third, because we are still free to rotate around the new pole through an angle, whose size can be between 0° and 360°. It is the representations of this group to which the oracular pronouncement of Hermann Weyl about quantum numbers and group representations applies.

The understanding of the fundamental role of this group and its representations in atomic spectroscopy and, ultimately, not only of representations of this group but also of many other groups as well in various domains of quantum physics led to a vigorous interest of the representation theory of finite and of continuous groups on the part of both physicists and mathematicians. The mathematical development remained for some time largely independent of the questions raised by spectroscopy and the later, more sophisticated theories of fundamental particles. There is not even, so far as I can see, a close relation between Weyl's reflections on group theory and quantum mechanics and his fundamental contributions to the mathematics of group representations. Before describing how the concepts introduced by the physicists found their way back into mathematics, I recall briefly the issues that faced spectroscopists in the period after 1885, when the structure of the frequencies or wavelengths, thus the color in so far as the light is visible and not in the infrared or ultraviolet regime, of the light emitted by hydrogen atoms was examined by Balmer, and in the period after 1913, when N. Bohr introduced his first quantum-mechanical explanations.

There are at least two physical issues to keep in mind when reconstructing the path that leads from the ideas of Dedekind, Kronecker and Frobenius through spectroscopy to the introduction of continuous groups and their representations into the theory of automorphic forms. The first has nothing to do with groups. The frequencies emitted by, say, the hydrogen atom are not arbitrary. There will be intervals of frequencies, in principle infinite in length, that can all be emitted, but these continuous intervals do not concern us at the moment. What concerns us are the frequencies that appear in isolation from the others. The Balmer formula recognizes, as do other formulas of the same nature, that it is best to express them as a difference. Bohr recognized that this difference should be taken seriously as a difference of energies, the energies of the atomic states that preceded and followed the emission of light, this being the energy of the emitted photon.

In the network of rectangles in Diagram A I have left no place for linear algebra, which provides in the contemporary world the material for a mathematical course perhaps even more basic than calculus. It is in the context of linear algebra that the possible energies are calculated, as proper values, characteristic values, or eigenvalues of matrices or linear operators. The notions of matrix or of linear operator are essentially the same. The other three terms, proper, characteristic, or eigen-values refer to exactly the same notion and are all translations of the same word *Eigenwerte*. It is the last translation, barbaric as it is, that comes most easily to my tongue and to the tongue of most mathematicians. I explain the notion with the help of the two simultaneous linear equations:

$$4x + 2y = \lambda x,$$
$$2x + y = \lambda y.$$

For most values of λ, this equation has exactly one solution in (x, y), namely $(0, 0)$, but for certain exceptional values of λ, at most two, it has an infinite number of solutions. These values of λ are called eigenvalues. They are calculated as the roots of

$$(\lambda - 4)(\lambda - 1) - 2 \cdot 2 = \lambda^2 - 5\lambda = 0.$$

Here the matrix of coefficients is given by

$$\begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}.$$

In general the four coefficients would be written as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

and the equation for λ would be

$$(\lambda - a)(\lambda - d) - bc = 0,$$

often written as

$$\begin{vmatrix} \lambda - a & -b \\ -c & \lambda - d \end{vmatrix} = 0. \tag{5.1}$$

So it has two solutions. With a small effort of imagination, one can pass to equations with three, four, or an arbitrary number of unknowns, and therefore with three, four, or more eigenvalues. With even more imagination, one can generate an infinite number of eigenvalues by passing to infinite matrices or to linear operators on a space of infinite dimensions. This is done in mathematics and in quantum mechanics, where the equation (5.1) is closely related to Schrödinger's equation and its eigenvalues are the energy levels. It usually appears in both mathematics and quantum mechanics as a linear differential equation. There is some legerdemain here, but we are on solid ground, and if we are to understand Diagram A from top to bottom, it is essential to accept that for the theory of algebraic numbers and for the theory of diophantine equations group representations, even in their infinite-dimensional guise, have important insights to offer. They are for algebraic number theory, in which they arose, and for its practitioners, like the prodigal son, who had taken his journey into a far country, was lost, and is now found.

Although the appearance of infinitely many dimensions is troubling to those who, like the elder brother in the parable, served without transgression for so many years, they are, I believe, an inescapable element of any adequate theory. The linear operators came later than Bohr's theory. Bohr's calculation of the energy levels was inspired by simpler physical notions, but from the very beginning, these levels had to be infinite in number because the light was emitted with infinitely many different frequencies, the most interesting being those that occurred discretely. The initial theory was also, so far as I know, unencumbered by group representations. As a consequence it could not fully explain the observed spectra.

The elements of the rotation group are given as acting on a three-dimensional space, the one with which we are familiar, it being clearly possible to compound one rotation r_1 with a second r_2 to obtain a third that is written as a product $r_2 r_1$, thus defining the group law. So each rotation maps the space onto itself, but without destroying the straight lines. Each straight line is transformed into a second straight line. Such transformations are called linear. Although it is a little harder to imagine, many other groups can be realized as groups of linear transformations, and the same group can be realized in many different ways and in many different dimensions.

The frequencies appearing in the various series of frequencies, Lyman, Balmer, Paschen, or principal, sharp, diffuse, and so on, discovered by the spectroscopists in the closing decades of the nineteenth century and the first decades of the twentieth century revealed themselves on closer and closer examination to have a structure whose complexities, among them the

splitting of lines in the presence of a magnetic field, were only unravelled with an ingenuity and inventiveness that for an outsider almost pass belief. These complexities are all related either to the angular momenta of the particles, nucleus or electrons, involved or to their composition, in which to complicate matters the exclusion principle of Pauli has major effects. The key to the analysis of the angular momenta is the rotation group and its different representations. The exclusion principle of Pauli is expressed by a much simpler group, one with only two elements, or, perhaps more precisely, by the group of all permutations. It is not our purpose here to attempt to follow the physicists in their ruminations except to conclude that the energy levels that are calculated mathematically have a structure that is very difficult to analyze, in part for dynamical reasons, thus in part expressed by a differential equation, and in part because of the mathematical complexities that arise when several representations of the group of rotations appear simultaneously. There are marvelous early texts on spectroscopy for the more adventurous to consult.

As the theory developed and as relativity theory made its appearance in quantum mechanics, so that the group of rotations was to some extent displaced by the Lorentz group, it was suggested by two physicists, separately and independently although they were brothers-in-law, Paul Dirac and Eugene Wigner, that the representations of the Lorentz group, of which the basic ones—the technical term is irreducible—those from which the others can be constructed with the help of familiar mathematical methods, would be the most important, could be of some physical interest. In contrast to the group of rotations, for which a fixed center is demanded, the possibility of motion is implicit in the Lorentz group, so that not only the complex structure entailed by the simultaneous appearance of several representations of the rotation group but also the mixture of discrete and continuous spectra met in spectroscopy and given through the Schrödinger equation or Heisenberg matrix appear and are to be analyzed together. This entails representations in an infinite number of dimensions. The physical interest never, so far as I know, materialized. Nevertheless a former student of Dirac, Harish-Chandra, whose thesis was inspired by the suggestion of Dirac and Wigner and who, having turned to mathematics, very quickly returned to the suggestion, his reflections now informed by a mathematical perspective, created over the years a magnificent and absolutely general mathematical theory. It is, to a considerable extent although there are other factors, this theory that enlarges the optic of class field theory and allows it to encompass, at least potentially, not merely the very limited, but none the less extremely difficult, class of equations in one variable with which it began but, as in the lowest rectangle on the left of Diagram A, all the diophantine equations of modern algebraic geometry.

What is fascinating here is not how much or how little the content of the rectangle *automorphic forms* is influenced by the theory of infinite-dimensional representations rather than by algebraic number theory or by, say, the theory of elliptic modular forms, but rather the curious journey of representation theory, arising in the elementary calculations of Dedekind of group determinants and the elegant and astonishing theory to which they led Frobenius, and then passing through the far countries of quantum mechanics and relativity theory. I would not suggest that it there wasted its substance—on the contrary. In spite of this, not all have rejoiced at its homecoming.

Scenic as the route from algebraic numbers and ideals to automorphic forms and functoriality over quantum theory and relativity is and essential as the concepts acquired on traversing it are, the core of the matter seems to me to lie rather in the little understood triangular region formed by Euler products, class field theory and automorphic forms. Frankness demands

that something be said about it, but the explanations will have no meaning for those readers with little mathematical training, and almost none for most mathematicians. It would, nevertheless, be irresponsible to omit them entirely and not to confess that something very important is lacking in my explanations. I postpone them to the final section.

A.6. Cartesian geometry and the mysteries of integration. We spent a good deal of time in §1 recalling, largely for the benefit of those who may not have seen them before, the basic formulas for the integration of rational functions and their relation to the fundamental theorem of algebra.

Although this theorem was certainly not known to Descartes, it is difficult not to recall it when reading the essay on geometry that supplements his *Discours de la méthode*. Much of his essay is given over to matters every bit as technical as those to which we have introduced the reader in the preceding passages, but in the initial pages it is his enthusiasm about some simple structural elements that appear when coordinates are introduced into planar geometry, especially about the degree of a curve and the influence of degree on the properties of intersection, that is most in evidence. The influence of Apollonius is transparent.

Consider in the plane a curve of the form

$$y = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0, \tag{6.1}$$

supposing at first that $a_n \neq 0$. The number of intersections of this curve with the line $y = b$ is the number of roots of

$$X^n + \frac{a_{n-1}}{a_n}X^{n-1} + \cdots + \frac{a_1}{a_n}X + \frac{a_0 - b}{a_n} = 0,$$

a root θ leading to the intersection (θ, b). If we, to take care of accidental degeneracy, count a multiple root as a multiple intersection and if we, in addition, include complex roots of the equation as complex intersections, then the line and the curve always intersect in exactly n points. Thus the geometric form of the solutions of the simultaneous equations,

$$y = b, \qquad y = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

is determined by the algebraic form. An intersection of a curve of degree one, the line $y = b$, and a curve of degree n, the curve (6.1), is a collection with $1 \times n$ points. Its geometric form is determined by the algebraic form of the simultaneous equations. This is a simple example of the principle of Descartes.

There is another modern feature that is not explicit in the essay. A line is of degree one, so that two lines have $1 \times 1 = 1$ points of intersection. This is, of course, not so if there the lines are parallel. As a remedy, one adds to the plane a line at infinity and obtains the *projective* plane in which a curve of degree m and a curve of degree n have exactly mn intersections, counted of course with multiplicities to take points at which the curves become tangent into account. The most familiar examples arise in the theory of conic sections, in which many of the intersections are complex, so that they do not appear in the usual diagrams. An ellipse and a circle can have four points of intersection, but if the ellipse becomes a circle, these points all become complex. If the circles coincide, there is a degeneracy about which nothing can be done and there is no question of counting points.

That the algebra contains the geometry, expressed so simply and intuitively, by the assertion that the number of intersections of two curves of degrees m and n is mn, manifests itself in very sophisticated ways in modern geometry and number theory and is the theme of the column on the left in the diagrams. To see this, we recall the geometry of plane curves, but

for which we allow not just real but also complex points. Start with the line, thus a single coordinate z, but now complex so that $z = x + iy$, x and y real. Thus the line becomes a plane. To the *line* has to be added a point at infinity to form the projective line. It is a single point, but the effect is to take the infinite plane of possible z and to fold it into a pouch that is closed at the top to form a ball, at first shapeless but that it is best to think of as spherical. This is a step that, although elementary and intuitive, requires some experience, some explanations, and some diagrams to grasp. It has many consequences as do the other improvements of Descartes's intuitions.

Consider the curve $E : y^2 = x^3 + 1$, one of those that so fascinated Jacobi. If we plot it in the usual way, there appears to be nothing of great interest about it. It extends smoothly from the left to the right of the page, rising steadily, $y \geqslant 0$, or falling, $y \leqslant 0$. With the projective plane in mind, we observe that the tangents have a slope whose magnitude increases to ∞ as we move to the left or to the right. So they become parallel and their point of intersection moves off to infinity. Thus, in the context of projective geometry, there is at least one additional point to be added to the curve, the common point of intersection of two lines parallel to the vertical axis of ordinates. This is a point on the line at infinity in the projective plane. So the curve E, or at least the points on it with real coordinates, has to be completed and the result is a closed loop. The two infinite ends close on each other. This done, there are still the points with complex coordinates. It is more of an effort to envisage them. As x sweeps out over the imaginary plane it is completed by assigning two different values for y, namely $y = \pm\sqrt{x^3 + 1}$. This yields a surface which it is an art, one introduced, I believe, by Riemann, to visualize. The surface is that of a ring or, if one likes, a doughnut, on which the curve of real points is a closed curve, say a meridian. There is on such a surface many, many other closed curves, of which the latitudinal curves are the first to suggest themselves.

Notice that if we take a doughnut-shaped surface, say an inner-tube, and cut it first along a meridian and then along a latitudinal curve, the result can be unrolled as a rectangular sheet, so that these two curves, or at least the number two, is a significant element of the geometry or, better, topology of the surface. Can this topology be seen directly in its algebraic form? It can, and not just for plane curves but for curves, surfaces, and higher-dimensional varieties in the plane, in three-dimensional projective coordinate space, or in coordinate space of any dimension. So the fundamental theorem of algebra has unsuspected implications. Basically all of the geometry of the variety—the mathematical term for curve, surface, three-fold, and so on—of complex points defined by algebraic equations can be recovered in a prescriptive manner from the equations themselves. The simplest cases are, as already observed, the number of points on the locus in the projective plane defined by two equations or in three-dimensional projective space by three.

For curves in the plane, this is best expressed by a theorem attributed in its general form to Niels Abel, a Norwegian mathematician who died young in 1829. It is a theorem about integrals, but as one learns in introductory calculus and as is clear from §1, integration is for many purposes an entirely algebraic matter. We continue to examine the curve E. We consider integrals of the form

$$\int F(x, y)\, dx,$$

in which it is understood that $y = \sqrt{x^3 + 1}$ depends on x and that F is a quotient of two polynomials in x and y. A typical example is (1.6), which is

$$\int \frac{dx}{y}.$$

Many of the integrals will contain a logarithmic term as do some of the integrals in (1.1). It is easy to see how. The logarithm is not an algebraic function and it is best to exclude these integrals from the statement of Abel's theorem. They would only add to its complexity.

If we add to F a function that is a derivative, thus a function of the form dG/dx, where G is also a quotient of two polynomials in x and y, and if $F_1 = F + dG/dx$, then

$$\int F_1(x, y)\, dx = \int F(x, y)\, dx + G(x, y),$$

so that from the point of view of integration theory one of these functions is tantamount to the other, or at least just as easily treated. So we need to study the integrals only for representatives of the possible functions F up to such modifications.

Up to such modifications the pertinent F can all be written in the form

$$\frac{A}{y} + \frac{Bx}{y}, \tag{6.2}$$

where A and B are two constants. To show this requires a little calculation but is not difficult. It shows that the number two significant for the topology also appears in the algebra. Of course, one example is not convincing, but, according to Abel's theorem, the same principles govern all curves, and according to subsequent developments algebraic varieties of any dimension. For example, the surface given by $y^2 = x^6 - 1$ is a double doughnut, thus a little like a pretzel, the number of curves along which it must be cut before it can be developed onto the plane is four, and (6.2) is replaced by

$$\frac{A}{y} + \frac{Bx}{y} + \frac{Cx^2}{y} + \frac{Dx^3}{y}. \tag{6.3}$$

By the way, for the curve given by the complex projective line itself, thus with a single parameter x, the associated surface is a sphere. So there are no closed curves on it that cannot be deformed continuously to a point and the two basic integrands of (6.2) or the four of (6.3) are not necessary at all. Every integral is given by a rational function. This corresponds to the second equation of (1.3) and, one would wish, to the experience of every student beginning the study of the calculus.

This intimate relation between the study of algebraic equations and topology has, as already mentioned, even more surprising expressions suggested—or conjectured—about sixty years ago by André Weil, realized by Alexander Grothendieck, and now central to the thinking of many specialists of the theory of numbers. They appear at the bottom of the column on the left of Diagram A: contemporary algebraic geometry, diophantine equations, motives, ℓ-adic representations. They are therefore central to the goals of the theories that contemporary pure mathematicians are trying to construct.

If the equations under consideration have, say, coefficients that are rational numbers, as does, say, the equation $y^2 = x^3 - 1$, then we can consider the solutions that are algebraic numbers. This brings Galois theory and Galois groups into play. Indeed the number of arbitrary coefficients that appear in (6.2) or (6.3) is, in fact, a dimension and this is so both from an algebraic standpoint and from a topological standpoint. The theory of Weil-Grothendieck—to

which of course others have made important contributions—reveals that these dimensions are also the dimensions of representations of the Galois group. The representations are not of the same nature as those appearing in spectroscopy or in the theory of automorphic forms as in §5, but very similar.

As an apparently simple, but still central and in the context of Diagram A extremely difficult, case, one can take the coefficients of equation (1.1) to be rational and to consider it as a diophantine equation. Suppose the roots are distinct. They are algebraic numbers $\theta_1, \dots, \theta_n$ and determine a field K, the collection of finite sums

$$\sum_{l \geqslant 0} \alpha_{k_1,\dots,k_l} \theta_1^{k_1} \theta_2^{k_2} \cdots \theta_n^{k_n}.$$

It is a Galois extension of the field k of rational numbers. If, for example, (1.1) is $X^2 + 5 = 0$, it is the field examined in §4. The introduction of sophisticated theories of integration is no longer necessary. The algebraic equations do not determine a curve, a surface of anything so difficult; they determine a finite collection, the points $\theta_1, \dots, \theta_n$, and a group, the Galois group of K/k whose elements permute these points among themselves. The associated representation is defined directly and simply by these permutations.

A.7. In recent decades. A modern goal, of which I have been a proponent for many years, is not merely to extend the abelian class field theory described succinctly in §4 to arbitrary Galois extensions K/k but to create a theory that contains it and relates more generally the representations produced by the Weil-Grothendieck principles to automorphic forms. The final result would be the theory envisaged in the final lines of Diagram A. After examining the material so far available, I persuaded myself—with reservations—that the theory would develop in two stages, some elements of both being already available, first a correspondence for the representations for the varieties of dimension zero defined by equations like (1.1), and then a second stage, the passage to varieties of any dimension. Although the second stage is implicit in the diagram, there are some aspects of it, probably the critical ones, with which I have not been particularly engaged and about which I have nothing to say here. I have as much familiarity as anyone with the first stage, but to describe it in any detail would be an imposition even on the most dedicated reader and somewhat presumptuous as well, because with it we enter a region in which the intimations that I described earlier play a large part. Even though there is a great deal left to be done, a few words may do no harm.

The arguments for the first stage can, in my view, which is not universally shared, be formulated entirely in the context of the theory of automorphic forms without any reference to geometry, as a part of what I call functoriality. This will not be an easy matter, but we have, I believe, begun. As was observed at the end of §4, class field theory in its classical form, rested on two considerations: a careful study of Galois extensions K/k of a very specific type with, in particular, abelian Galois groups; the study of the decomposition of prime ideals in k in the larger field K. Similar features may very well appear in the proof of functoriality. First of all, a similar study of Galois extensions but with arbitrary Galois groups, perhaps, as for the abelian case, with certain simplifying assumptions. No-one has yet undertaken this, perhaps simply because it offered until now no rewards.

The general form of the study of the decomposition of prime ideals is impossible to describe within a brief compass. The current theory of automorphic forms requires the introduction of groups that, like those appearing in the suggestion of Dirac-Wigner, admit infinite-dimensional representations, but at the same time are related not directly to the decomposition of prime

ideals as such but to the decomposition of matrices with elements from these prime ideals and their powers. The necessary theorems or propositions are much more complex and have a much larger component of techniques from the integral and differential calculus and its successor, modern analysis. A beginning on their development is made in the paper *La formule des traces et fonctorialité: le début d'un programme*, written in collaboration with Ngô Bảo Châu and Edward Frenkel.

I have suggested elsewhere that functoriality once acquired, the passage from motives to automorphic forms with the help of p-adic deformations, a standard technique, exploited in particular by Andrew Wiles, will become a much more effective and much less haphazard procedure than at present.

Notes

1. I am not concerned, however, with history in any systematic way. Just as the present is richer, perhaps even more tolerable and acceptable, when viewed with a knowledge and understanding of the past, even though the knowledge and the understanding are for most, maybe all, of us necessarily individual, provisional, accidental, and partial, so is the judgement of a mathematician likely to be juster and more valuable the more insight he has into the waxing and waning importance over time of various mathematical notions. I cannot pretend to have much insight and certainly no systematic knowledge.

2. It may be excessive to introduce the Devil as well, but there is an appealing fable that I learned from the mathematician Harish-Chandra, and that he claimed to have heard from the French mathematician Claude Chevalley. When God created the world, and therefore mathematics, he called upon the Devil for help. He instructed the Devil that there were certain principles, presumably simple, by which the Devil must abide in carrying out his task but that apart from them, he had free rein. Both Chevalley and Harish-Chandra were, I believe, persuaded that their vocation as mathematicians was to reveal those principles that God had declared inviolable, at least those of mathematics for they were the source of its beauty and its truths. They certainly strived to achieve this. If I had the courage to broach in this paper genuine aesthetic questions, I would try to address the implications of their standpoint. It is implicit in their conviction that the Devil, being both mischievous and extremely clever, was able, in spite of the constraining principles, to create a very great deal that was meant only to obscure God's truths, but that was frequently taken for the truths themselves. Certainly the work of Harish-Chandra, whom I knew well, was informed almost to the end by the effort to seize divine truths.

   Like the Church, but in contrast to the arts, mathematics is a joint effort. The joint effort may be, as with the influence of one mathematician on those who follow, realized over time and between different generations—and it is this that seems to me the more edifying—but it may also be simultaneous, a result, for better or worse, of competition or cooperation. Both are instinctive and not always pernicious but they are also given at present too much encouragement: cooperation by the nature of the current financial support; competition by prizes and other attempts of mathematicians to draw attention to themselves and to mathematics. Works of art or of literature, however they may have been created in the past or in other cultures, are presently largely individual efforts and are judged accordingly, although movements in styles and in materials are perhaps more a consequence of the human impulse to imitate than of spontaneous individual inspiration.

   Harish-Chandra and Chevalley were certainly not alone in perceiving their goal as the revelation of God's truths, which we might interpret as beauty, but mathematicians largely use a different criterion when evaluating the efforts of their colleagues. The degree of the difficulties to be overcome, thus of the effort and imagination necessary to the solution of a problem, is much more likely than aesthetic criteria to determine their esteem for the solution, and any theory that permits it. This is probably wise, since aesthetic criteria are seldom uniform and often difficult to apply. The search for beauty quickly lapses in less than stern hands into satisfaction with the meretricious. So the answer to Vittorio Hösle's question whether there is beauty in mathematical theories might be: there can be, but it is often ignored. This is safer and appears to demand less from the judges. On the other hand, when the theorem in which the solution is formulated is a result of cumulative efforts by several mathematicians over decades, even over centuries, and of the theories

they created, and when there may have been considerable effort—the more famous the problem the more intense—in the last stages, it is not easy, even for those with considerable understanding of the topic, to determine whose imagination and whose mathematical power were critical, thus where the novelty and insight of the solution lie, and whose contributions were presented with more aplomb and at a more auspicious moment.

3. The mathematician Theaetetus is famous for his understanding that the square roots of numbers like 2, 3 and 5 are irrational, but this, although briefly mentioned at the beginning of the lengthy epistemological dialogue, is not particularly relevant to its discussion of the nature of knowledge. Presumably Plato chose a well-known mathematician as Socrates's interlocutor and mentioned his principal accomplishment in part for a reason not so dissimilar to one of mine for introducing a Platonic dialogue: to provide a little color!

   These lines were in the text prepared in connection with the lecture at Notre Dame even before I came to the conference, although I did not reach them in the talk itself. At Notre Dame, I had a brief conversation with Kenneth Sayre, from whom I learned that the introduction of Theaetetus not only into the dialogue that bears his name but into the next dialogue, the Sophist, had more serious literary purposes than I, in my innocence and ignorance, had imagined. These are clarified in his book *Plato's Literary Garden*.

   It would be presumptuous for me to comment on the force of Plato's use of quadratic surds as epistemological metaphors. I do not have the experience that allows any confidence that I have any inkling how they might have been understood by Plato, nor do I have the courage to pursue the metaphor in a modern context, although that would not be ridiculous, for the study of quadratic and higher surds continues to be of fundamental importance for the theory, class field theory, that occupies a central position in Diagram A, and in pure mathematics. One of the first appearances of quadratic surds in a novel, modern way is in Gauss's theory of binary quadratic forms, which was followed a hundred years later by a similar theory for higher surds that, according to the once very familiar, but now little read although still important, 1925 report of Helmut Hasse *den Kernpunkt des ganzen Gedankenganges* (of class field theory) *bildet.* There are, I believe, in this connection still critical issues to be resolved. I shall return to them in §7.

4. Galois died in 1832 at the age of twenty-one in what is often considered to be a duel with a ruffian but looks very much to me as though it could have been an assassination by agents of Louis-Philippe. So far as I know the circumstances of his death have never been examined critically by any one familiar with the methods of the secret police in the July Monarchy or with the government's stand with regard to uncompromising revolutionary elements.

   These lines, too, were in the text before I came to the conference, although, again, I did not reach them in the talk itself. During the course of the conference, Mario Livio drew my attention to his book *The equation that couldn't be solved*, to the wealth of biographical material, new and old, on Galois that it contains, and to his own reflections on the material and its import. Even without the appeal of Galois's reputation as a mathematician, it would be instructive, I believe, to have the comments on his political activities and his death of someone thoroughly familiar with social conditions in France at the time of the July revolution and with a detailed understanding of internal security under Louis-Philippe.

5. Those familiar with the square root i of -1 and its role in complex analysis are perhaps accustomed to distinguishing between i and $-i$, because in the customary geometric representation, i is above the axis of real numbers and $-i$ below it. This is an illusory distinction, although mnemonically of great convenience in geometric arguments.

6. The Duchy was very small. It is fascinating, but no doubt idle, to reflect on the extent to which Gauss's native ability was favored by fortuitous circumstances. He was remarkably fortunate, for he drew himself in the Duchy's schools not only to the attention of his teacher, J. G. Büttner, but also to that of his assistant, Martin Bartels, only eight years older than Gauss himself. It is not clear how much he may have learned from Bartels, but in such circumstances a little goes a long way and the early acquaintance with Bartels, who went on to a distinguished career as professor of mathematics in Reichenau, Dorpat and Kazan, was likely to have been decisive in Gauss's intellectual development. Both G. W. Dunnington, the author of *Gauss, Titan of science* (1954) and W. K. Bühler, the author of *Gauss, A biographical study* (1981), perhaps the two best known Gauss biographies, seem to me singularly insensitive to the value to an adolescent of even a very brief introduction to intellectual possibilities, let alone that of sustained intercourse over a period of

years. I am also taken aback by Bühler's suggestion that someone of Gauss's mathematical distinction might have difficulty learning Latin or Greek. It is instructive to compare Gauss's early career with that of his near contemporary, the geographer Karl Friedrich von Klöden, born in 1786 in Berlin, who describes his own genuinely impoverished childhood and early career in his once fairly widely read *Jugenderinnerungen*, but this is a matter of social history.

In descriptions of Gauss's achievement, his excellent early education and his early introduction to serious mathematics by highly competent teachers is obscured, in hopes, perhaps, of dazzling innocent readers. Gauss himself, with the occasional convenient reticence, may—or may not—have been complicit in the deception, but not in any very troubling way. Unfortunately, an exaggerated emphasis on his prodigiousness rather than on his education, industry, and good fortune deprives his early biography of a number of its important lessons.

7. Three of the proofs were published in Latin, the fourth in German. All are available in German, on line at

   `https://www.archive.org/details/dieviergausssche00gausuoft`

   Thanks to the kindness of Ahmet Feyzioğlu, who presented me with a recently published collection, "Cebirin Temel Teoremi için dört İspat", of translations by Gülnihal Yücel, I had also the pleasure of reading the first proof, the one that concerns me here, as well as Gauss's comments on the attempts of his predecessors in an elegant contemporary Turkish. The refusal of the vernacular not only by mathematicians in countries to which modern mathematics came late but also by those in countries like France, Germany, Russia in which it was once vigorous and in which the older, vernacular mathematical literature is still extremely important has cost mathematics and mathematicians more than they have gained and more than they imagine—even if it is just the leisure to reflect undisturbed on our own ideas. So it is a joy to see even modest efforts to counter it.

8. I am very likely reading more into Descartes's explanations than is there. I have made no effort to discover what precisely he could have had in mind with the notion of degree of a curve when writing the supplement on geometry that is attached to the Discours. In some sense, but I suppose it has to be a very weak sense, he anticipates the theorem of Bézout, first formulated by the French mathematicien Étienne Bézout in the eighteenth century. It affirms that the intersection of two plane curves defined respectively by an equation of degree m and one of degree n has in general mn points. Bézout was the author of text-books on algebra used in French lycées of the period. He has also acquired an extremely modest extra-mathematical notoriety. The novelist Stendhal, who as Henri Beyle began his career as a student of mathematics, flaunts in his memoirs his knowledge—not substantial—of mathematics and, in particular, is given to mocking Bézout.

9. Whether the effort was worthwhile is a fair question, especially for those concerned not with mathematics as such but the beauty of mathematics. Does mathematical beauty or pleasure require such an accumulation of concepts and detail? Does music? Does architecture? Does literature? The answer is certainly "no" in all cases. On the other hand, the answer to the question whether mathematical beauty or pleasure admits such an accumulation and whether the beauty achieved is then of a different nature is, in my view, "yes". This response is open to dispute, as it would be for the other domains. My purpose here is to make the case for a mathematical beauty that transcends neither the simple mathematical pleasures of arithmetic and geometry nor the fascination of difficult problems but that integrates them with the quite different intellectual pleasure of creating order from seeming chaos, even in an inconsequential—but not for me—realm. One man's order is sometimes another's chaos. So I do not expect everyone to be persuaded. I myself am haunted by Rudyard Kipling's familiar conundrum,

   > And each man hears as the twilight nears, to the beat of his dying heart,
   > The Devil drum on the darkened pane: "You did it, but was it Art?"

Compiled on July 30, 2024.