Mathematical Explanation
Author(s): Mark Steiner
Source: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, Vol. 34, No. 2 (Aug., 1978), pp. 135-151
Published by: Springer
Stable URL: http://www.jstor.org/stable/4319237
Accessed: 27/01/2010 16:03

MARK STEINER

# MATHEMATICAL EXPLANATION*

(Received 23 August, 1977)

Philosophers have long pondered explanation in the natural sciences. If they have ignored it in the mathematical sciences, blame lies perhaps with a lingering distinction between 'matters of fact' and 'relations among ideas', the corollary being that mathematics (belonging to the latter class) has nothing to explain. Platonism, no less than empiricism, has also traditionally stressed the differences between natural science and mathematics. The growing acceptance, however, of continuity between the natural and mathematical sciences – urged by Quine, Putnam, and the present author[1] – has prepared the way for what follows here.

Mathematical explanation exists. Mathematicians routinely distinguish proofs that merely demonstrate from proofs which explain. Solomon Feferman puts it this way:

> Abstraction and generalization are constantly pursued as the means to reach really satisfactory explanations which account for scattered individual results. In particular, extensive developments in algebra and analysis seem necessary to give us real insight into the behavior of the natural numbers.[2]

Chang and Keisler, to cite two more logicians, propose to 'explain' preservation phenomena – i.e. certain theories are such that submodels of their models of their models are again models, or that unions of chains of their models are models, or that homomorphisms of their models are models – to 'explain' these phenomena "just by the syntactical form of the axioms".[3] For example, they prove that a theory is preserved under submodels iff it has a purely universal axiomatization; under unions of chains iff it has a universal-existential axiomatization; and so forth. Let us explore what is common to mathematical explanations such as these; we can always invoke 'family resemblances' later, if we fail.

An obvious suggestion is to identify explanation with generality or abstraction, as Feferman thinks. There *is* something general and abstract about complex analysis, which at present provides the greatest insight into the

*Philosophical Studies* 34 (1978) 135–151. *All Rights Reserved*

prime numbers, especially their aggregate behavior. We can divide Feferman's thesis into at least three:

(a) A proof is explanatory per se if it is abstract (or general) in some absolute sense, yet to be specified.

(b) A proof is explanatory per se if it is *more* abstract (or general) than what it proves.[4]

(c) Of two proofs of the same theorem, the *more* explanatory is the more abstract (or general).

Though (a) is probably not what Feferman has in mind, Feferman's example of analytic number theory illustrates either (b) or (c). Kreisel explicitly adopts (c) – in a private communication – writing that "familiar axiomatic analysis in terms of the greater generality of (the theorems occurring in) one proof than (in) the other" is 'sufficient' to distinguish between proofs' explanatory value.

Naturally Feferman, or Kreisel, need to clarify 'generality' or 'abstraction'. As for the latter, abstraction surely increases in the ascent from first to higher 'order' arithmetic (second-order arithmetic invokes sets of numbers; third-order, sets of sets of numbers; and so forth). And Feferman could claim the following simple theorem for support:

$$S(n) = 1 + 2 + 3 + \cdots + n = n(n+1)/2.$$

We can prove this by induction by remarking that

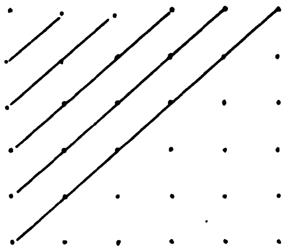
$$S(n+1) = S(n) + (n+1) = n(n+1)/2 + 2(n+1)/2 = \\ (n+1)(n+2)/2.$$

But a more illuminating proof is given by

$$\begin{array}{ccccccccc} 1 & + & 2 & + & 3 & + \cdots + & n & = S \\ n & + & (n-1) & + & (n-2) & + \cdots + & 1 & = S' = S \\ \hline (n+1) & + & (n+1) & + & (n+1) & + \cdots + & (n+1) & = n(n+1) \end{array}$$

and this proof, when formalized, quantifies over *sequences* of natural numbers, using techniques expounded in Quine's *Set Theory and Its Logic*, whereas the first proof makes do with the numbers themselves and is thus less abstract.

But I doubt whether the mere abstractness of the latter proof carries the day; rather, its pictorial aspect, the abstract sequences being necessary to formalize the picture. Indeed, what is perhaps an even more explanatory proof than the latter is wholly geometric:



By dividing a square of dots, $n$ to a side, along its diagonal, we get an isosceles right triangle containing

$$S(n) = 1 + 2 + 3 + \cdots + n$$

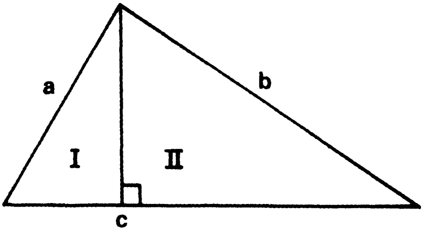
dots. The square of $n^2$ dots is composed of two such triangles – though if we put the triangles together we count the diagonal (containing $n$ dots) twice. Thus we have

$$S(n) + S(n) = n^2 + n, \quad \text{q.e.d.}$$

What then of *generality*? Here the problem of definition is acute. Kitcher is right (in his recent article on Bolzano[5]) in denying that *cardinality* has anything to do with generality (though he, as I mentioned in the previous note, deals primarily with thesis (b)). Nor is it easy to see why the geometric method of summing the first $n$ integers is more general than the inductive proof.

A partial criterion of generality emerges from considering that some proofs prove more than others. The so-called 'elementary' proof of the Prime Number Theorem proves that the ratio of $\Pi(x)$ – the number of primes less than or equal to $x$ – to $x/\log(x)$ approaches 1 as $x$ goes to infinity. But the analytic proofs which Feferman mentions given a much better estimate for given $x$ how much $\Pi(x)$ deviates from $x/\log(x)$. This is why some mathematicians, and perhaps Feferman himself, regard the analytic proofs as more explanatory. Or consider the Pythagorean proof that the square root of 2 is not rational: if $a^2 = 2b^2$, with $a/b$ reduced to lowest terms, then $a^2$ and thus $a$ itself have to be even; thus $a^2$ must be a multiple of 4, and $b^2$ – and thus $b$ – multiples of 2. Since therefore $a^2 = 2b^2$ implies that both $a$ and $b$ must be even, contradicting our (allowable) stipulation that $a/b$ be reduced to lowest terms, it can never be true, q.e.d. The key point here is the proposition that if $a^2$ is even so is $a$. This can be verified by squaring an arbitrary odd number $2q + 1$ and showing that the result must be odd. Indeed for each prime $p$, one can separately verify that if $p$ divides $a^2$ it must divide $a$ also, though the proofs become more and more complex (where $p = 5$, for example, one must square $5q + 1$, $5q + 2$, $5q + 3$, and $5q + 4$ and show that in no case is the result divisible by 5).[6] But by using the Fundamental Theorem of Arithmetic – that each number has a unique prime power expansion (e.g. 756 is uniquely $2^2 \times 3^3 \times 7^1$) – we can argue for the irrationality of the square root of two swiftly and decisively. For in the prime power expansion of $a^2$ the prime 2 will necessarily appear with an *even* exponent (double the exponent it has in the expansion of $a$), while in $2b^2$ its exponent must needs be *odd*. So $a^2$ never equals $2b^2$, q.e.d. Generally, the same proof shows that $a^2$ can never equal $nb^2$, unless $n$ is a perfect square (so that all the exponents in its prime power expansion will be even).[7]

A final example of an explanatory proof which seems to highlight the role of generality in explanation we glean from Polya's book, *Induction and Analogy in Mathematics.*[8] (Though Polya's book is a gold mine of examples of mathematical explanation, Polya himself does not discuss the notion.) The most explanatory proof of the Pythagorean Theorem – the proof Polya explains – is also the most general, i.e. proves the most.



First, the areas of similar plane figures are to each other as the squares of their corresponding sides. In particular, any three similar figures constructed on the above right triangle have areas which can be represented as $ka^2$, $kb^2$, and $kc^2$. Now if we could find any threesome of similar figures constructed on the sides of the triangle in which the sum of the figures on sides $a$ and $b$ were equal to the area of the figure on side $c$, we would be able to write

$$ka^2 + kb^2 = kc^2,$$

from which the Pythagorean Theorem follows immediately. Thus the Pythagorean Theorem is equivalent to a generalization; and the generalization, to any of its instances. But triangles I and II are obviously similar to each other and to the whole triangle, and the whole triangle may be regarded as being constructed on its own hypotenuse, $c$! Clearly, triangle I plus triangle II equals the whole triangle, so we have our similar threesome, q.e.d.

In sum, we have given a partial interpretation to thesis (c) above: of two proofs, the more explanatory is the more general. To deduce a theorem as an instance of a generalization, or as a corollary of a stronger theorem, is more explanatory than to deduce it directly.

But even this criterion fails. Let's look at the square root of two again. Hardy proves the general result that

$$a^2 = nb^2$$

implies that $n$ is a perfect square by a method unlike the one above. Assume that $a/b$ is reduced to lowest terms. If a prime $p$ divides $b$, and thus $b^2$, it must divide $a^2$ and thus $a$, contradiction. So no prime divides $b$, and $b$ must be the number 1; and $n$, of course, is a perfect square. Specializing to the case where $n = 2$, we get Pythagoras' result – that the square root of 2 is irrational – but it would be hard to claim that our present proof, though more general, is more explanatory of the specific result than Pythagoras' argument (we must here distinguish 'more explanatory' from 'explains more'). We reluctantly conclude that the proof invoking the Fundamental Theorem of Arithmetic is not more explanatory than Pythagoras' because more general, but for some other reason.

An even more striking refutation of the generality criterion is furnished by the Eulerian identity

$$(1+x)(1+x^3)(1+x^5)\cdots = 1 + x^2/(1-x^2) + \\ x^4/(1-x^2)(1-x^4) + x^9/(1-x^2)(1-x^4)(1-x^6) + \cdots$$

of which I will present two proofs which, though involved, are well worth the reader's patience. The first, by Euler,[9] uses the device of introducing a second parameter, $a$:

$$\begin{aligned} \text{Let } F(a) = F(a,x) &= (1+ax)(1+ax^3)(1+ax^5)\cdots \\ &= 1 + c_1 a + c_2 a^2 + c_3 a^3 + \cdots \end{aligned}$$

So each of the $c$'s depends on $x$ but not on $a$. $F(a)$ is nothing but $(1+ax)F(ax^2)$, so we have

$$1+c_1a+c_2a^2+\cdots=(1+ax)(1+c_1ax^2+c_2a^2x^4+\cdots).$$

Equating the coefficients of $c$, we obtain

$$\begin{aligned} c_1 &= x+c_1x^2 \\ c_2 &= c_1x^3+c_2x^4 \\ &\vdots \\ c_m &= c_{m-1}x^{2m-1}+c_mx^{2m}. \end{aligned}$$

So

$$c_m = \frac{x^{2m-1}}{1-x^{2m}\;c_{m-1}} = \frac{x^{1+3+5+\cdots+(2m-1)}}{(1-x^2)(1-x^4)\cdots(1-x^{2m})} = \frac{x^{m^2}}{(1-x^2)(1-x^4)\cdots(1-x^{2m})}.$$

Summing over $m$

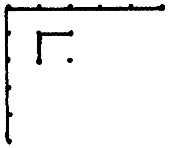
$$\begin{aligned} F(a) &= (1+ax)(1+ax^3)(1+ax^5)\cdots = \\ &\quad 1+ax/(1-x^2)+a^2x^4/(1-x^2)(1-x^4)+\cdots \end{aligned}$$

from which Euler's theorem follows by letting $a=1$. This proof, despite its brillance and generality does not really explain the identity. Note, in particular, that this proof could have been used, not only to verify the theorem, but to discover it. Thus, in passing, we have refuted a plausible suggestion even before we have had a chance to discuss it: that the explanatory proof is the one which could be used to determine, not merely verify, the result. Though our previous example – the proof, using the Fundamental Theorem of Arithmetic, that $\sqrt{2}$ is irrational – lends credence to this suggestion, the present example shows that 'discoverability' is at best a symptom of explanation in mathematics, not a criterion.

If Euler's own proof does not provide an explanation of his identity, how do we explain it? When we multiply the infinite product

$$(1+x)(1+x^3)(1+x^5)\cdots$$

we see that the coefficient of $x^n$ will be $(1+1+1+\cdots)$, as many units as the number $n$ can be partitioned into odd and unequal parts. Now, consider for example the partition of 15 into 11 + 3 + 1, representing it graphically as in the figure:



Clearly, any such partition – into odd and unequal parts – can be seen as a 'self-adjoint' partition, that is, a partition which, like our figure, is symmetrical about a diagonal running NW to SE. Now, since the infinite product 'enumerates' the partitions of $n$ into odd and unequal parts, to prove Euler's identity, it would be sufficient to prove that the right side of the identity,

$$1 + x^2/(1 - x^2) + x^4/(1 - x^2)(1 - x^4) +$$
$$x^9/(1 - x^2)(1 - x^4)(1 - x^6) + \cdots$$

'enumerates' the *self-adjoint* partitions of $n$. (This means, remember, that when this infinite sum is resolved into a simple power series of $x$, the coefficient of $x^n$ will be always the number of such self-adjoint partitions of $n$.)

But, again referring to our diagram, any self-adjoint partition of $n$ can be seen as a square of size $m^2$ dots, in this case $m = 3$, and two 'tails' each representing the same partition of the number $1/2(n - m^2)$ into parts not exceeding $m$. In this case each tail represents a partition of 3 into 1 + 1 + 1. Keeping this comment in mind, we turn now to the general term of our infinite sum:

$$x^{m^2}/(1 - x^2)(1 - x^4)(1 - x^6) \cdots (1 - x^{2m}).$$

The fraction

$$1/(1 - x^2)(1 - x^4)(1 - x^6) \cdots (1 - x^{2m})$$

gives us the product of $m$ different geometric series:

$$1 + x^2 + x^{2+2} + x^{2+2+2} + \cdots$$
$$1 + x^4 + x^{4+4} + x^{4+4+4} + \cdots$$
$$1 + x^6 + x^{6+6} + x^{6+6+6} + \cdots$$
$$\vdots$$
$$1 + x^{2m} + x^{2m+2m} + x^{2m+2m+2m} + \cdots.$$

Multiplying, we get a power series where the coefficient of $x^n$ is $(1 + 1 + \cdots + 1)$, the number of units equalling the number of partitions of $n$ into even parts not exceeding $2m$. (The coefficient for odd $n$ is, accordingly, 0.) In other words, the fraction $1/(1 - x^2)(1 - x^4)(1 - x^6) \cdots (1 - x^{2m})$ enumerates the partitions of $n$ into even parts not exceeding $2m$. Obviously, then, the general term, which is this fraction multiplied by $x^{m^2}$, enumerates the partitions of $n - m^2$ into the same even parts not exceeding $2m$. But each such partition of $n - m^2$ is obviously correlated with a partition of $1/2(n - m^2)$ into even *and odd* parts not exceeding $m$. We thus have that the general term

$$x^{m^2}/(1 - x^2)(1 - x^4) \cdots (1 - x^{2m})$$

enumerates the partition of $1/2(n - m^2)$ into (odd and even) parts not exceeding $m$. But recall that this is also the number of partitions of $n$ into a square of size $m^2$ with two tails, i.e. the number of *self-adjoint* partitions of $n$ which are based on a square of side $m$ dots! Clearly, if we sum up the general term over $m$, we will get an enumeration of *all* self-adjoint partitions of $n$, q.e.d. This proof, though it proves only the result in question, is far more explanatory than the original Eulerian proof using an introduced parameter, even though the Eulerian proof is an instrument of discovery besides being more general!

As for the explanatory proof of the Pythagorean Theorem, which was more general than the usual proofs, remember that *any* proof of the Pythagorean Theorem is tantamount to a proof of its generalization, once we notice that the areas of similar figures are to each other as the squares of their corresponding sides. Note, too, that one need not in our proof detour through figures in general in order to prove the Pythagorean Theorem for squares. Once we realize that any right triangle can be decomposed into two similar right triangles similar to the original triangle by dropping an altitude to the hypotenuse, we can then immediately infer, from the decomposition of the whole triangle into triangle I + triangle II, that $a^2 + b^2 = c^2$ *without* first concluding that the theorem holds for all similar figures. We need not even know that areas of similar figures are proportional to the squares of the corresponding sides, for which we need the general definition of the area of a plane figure as the limit of the areas of little squares which can be fit into the figure. For we have already the area of a triangle as $1/2bh$ independently.

Perhaps the explanatory value even of our proof of the Pythagorean Theorem lies not in its generality.

So far we have rejected generality or abstractness as criteria for explanation in mathematics – in particular we have rejected the view that a proof is more explanatory than another because more general. Another suggestion, invoking mathematical discovery, was brusquely dismissed. A third criterion links explanation with ability to visualize a proof – and many explanatory proofs have this character. Aside from possible counter-examples, however, this criterion is too subjective to excite. Yet a satisfactory theory of mathematical explanation must show why all these suggestions are plausible.

My view exploits the idea that to explain the behavior of an entity, one deduces the behavior from the essence or nature of the entity. Now the controversial concept of an essential property of $x$ (a property $x$ enjoys in all possible worlds) is of no use in mathematics, given the usual assumption that all truths of mathematics are necessary. Instead of 'essence', I shall speak of 'characterizing properties', by which I mean a property unique to a given entity or structure within a *family* or domain of such entities or structures. (I take the notion of a family as undefined in this paper; examples will follow shortly.) We have thus a relative notion, since a given entity can be part of a number of differing domains or families. Even in a single domain, entities may be characterized multiply. Thus, one way of epitomizing the number 18 is that it is the successor of 17. But often it is more illuminating to regard 18 in light of its prime power expansion, $2 \times 3^2$.

My proposal is that an explanatory proof makes reference to a characterizing property of an entity or structure mentioned in the theorem, such that from the proof it is evident that the result depends on the property. It must be evident, that is, that if we substitute in the proof a different object of the same domain, the theorem collapses; more, we should be able to see as we vary the object how the theorem changes in response. In effect, then, explanation is not simply a relation between a proof and a theorem; rather, a relation between an array of proofs and an array of theorems, where the proofs are obtained from one another by the 'deformation' prescribed above.[10] (But we can say that each of the proofs in the array 'explains' its individual theorem.) Note that this proposal is an attempt at explicating mathematical explanation, not *relative* explanatory value, as the previous criteria.

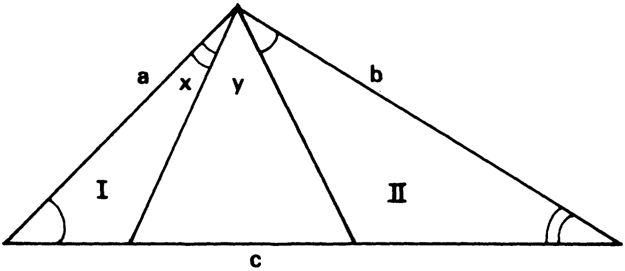
Our examples of explanation in mathematics are all analyzable this way.

Our proof that $a^2 = 2b^2$, which uses the prime power expansions of $a$ and $b$ (and 2), conforms to our description, since the prime power expansion of a number is a characterizing property. It's easy to see what happens, moreover, when 2 becomes 4 or any other square; the prime power expansion of 4, unlike that of 2, contains 2 raised to an even power, allowing

$$a^2 = 2b^2 .$$

In the same way we get a general theorem: the square root of $n$ is either an integer or irrational. Generalizing further, almost the same reasoning gives us the same for the $p$th root of $n$. It is not, then, the general proof which explains; it is the *generalizable* proof.

Or take our 'good' proof of the Pythagorean Theorem. It characterizes the right triangle as the only one decomposable in this way, into two triangles similar to each other and to the whole.



If we let the vertex of the right triangle vary (calling the largest side of the triangle, $c$, the hypotenuse), and try to decompose the triangle as before, by drawing lines $x$ and $y$ from the vertex to $c$, such that triangles I and II remain similar to each other and to the whole, we find that triangles I and II fail to exhaust the whole when the vertex varies between 90° and 180°; overlap when the vertex diminishes from 90° to 60°; and at 60°, coincide. We can even calculate the (positive or negative) difference, then, between the sum of squares constructed on the sides of the triangle, and the square on the hypotenuse for any triangle – calculate the error. The characterizing property for the right triangle, then, is simply the coincidence of lines $x$ and $y$. (It is interesting to note that at the 'extremes', where the vertex is 60° or 180°, the sum of the squares on the sides is twice and one-half the hypotenuse square, respectively.)

Both explanatory proofs that the sum of the first $n$ integers equals $n(n+1)/2$

proceed from characterizing properties: the one, by characterizing the symmetry properties of the sum $1+2+\cdots+n$; the other its geometrical properties. By varying the symmetry or the geometry we obtain new results, conforming to our scheme. The proof by induction does not characterize anything mentioned in the theorem. Induction, it is true, characterizes the *set* of all natural numbers; but this *set* is not mentioned in the theorem.[11]

Turning finally to Euler's identity, the explanatory proof is now seen to be so because it links the theorem to a characterizing property of the infinite sum and the infinite product – the property of enumerating certain partitions of $n$. Changing the geometry of the situation, we can suggest and prove new theorems, for example:

$$(1+x)(1+x^2)(1+x^3)\cdots = \\ 1+x/(1-x)+x^3/(1-x)(1-x^2)+x^6/(1-x)(1-x^2)(1-x^3) \\ +\cdots$$

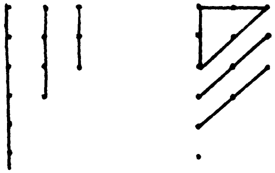
where the general term on the right is

$$x^{1+2+\cdots+m}/(1-x)(1-x^2)\cdots(1-x^m),$$

or

$$x^{m(m+1)/2}/(1-x)(1-x^2)(1-x^3)\cdots(1-x^m).$$

Now the left side clearly enumerates all partitions of $n$ into unequal parts.



$13=6+4+3$

$7=3+3+1$

And a partition of $n$ into unequal parts can be seen as a decomposition of $n$ into an isosceles right triangle of side $m$ – whose total dots are $1+2+\cdots+m = m(m+1)/2$ – *plus* a partition of the remainder $n-m(m+1)/2$ into equal and unequal parts no greater than $m$. (In the above diagrams, a partition of 13 is redescribed as an isosceles right triangle of side 3 dots (total: 6) and a partition of the remainder (7) into parts not exceeding 3.) But the general term of the right hand side obviously does enumerate all partitions of

$$n-(1+2+\cdots+m)$$

into (equal and unequal) parts no greater than $m$, corresponding to the partitions of $n$ into *unequal* parts based on a triangle of side $m$. Summing the right hand side thus enumerates *all* partitions of $n$ into unequal parts, the same as the left hand side, q.e.d. Thus the isosceles right triangle has taken the place of the square.[12]

Admittedly, Euler's own proof of the original identity also yields this last identity. For Euler, you remember, proved

$$(1+ax)(1+ax^3)(1+ax^5)\cdots = 1+ax/(1-x^2)+ a^2x^4/(1-x^2)(1-x^4)+\cdots$$

and merely setting $a=x$ and replacing $x^2$ by $x$ immediately gives us our result. The new result is contained within the old. The point is, however, that generalizability through varying a characterizing property is what makes a proof explanatory, not simple generality.

Our account of mathematical explanation suggests why the other criteria are plausible. First, generality is often necessary for capturing the essence of a particular, and the same goes for abstraction. To characterize the primes may take the full resources of complex analysis. Another suggestion was that the explanatory proof could have been used to discover the result, and it is often the case that a characterizing proof is intuitive enough to serve as an instrument of discovery. Finally, a characterizing property is likely to be visualizable (as is certainly the case with a geometrical property).

The present theory illuminates, aside from explanation, the notion of relevance in mathematics. For example: Euclid demonstrated the impossibility of a sixth regular solid by remarking that the sum of the angles around any vertex must be less than 360°, that it takes three polygons to form a vertex, and that regular polygons of six or more sides require angles of 120° or more. But actually this theorem has nothing to do with regular solids, for it follows from Euler's discovery that in any polyhedron

$$V-E+F=2$$

(where $V$ is the number of vertices; $E$, the number of edges; $F$, the number of faces), using only the facts that in a regular solid all the faces are bounded by the same number of edges, and that the number of edges meeting in each of the vertices is the same. Euclid's theorem then follows from the topological regularity of a regular solid, without the assumption of metrical regularity (i.e. the assumption that each face is a regular polygon). Euler's discovery cited above does not, actually, merely characterize polyhedrons, but all their continuous deformations. In fact the number 2 in Euler's formula really characterizes the sphere and surfaces 'homeomorphic' to it. (In fact, there are polyhedrons – those, for example, with 'tunnels' – which are not homeomorphic to the sphere, and in fact Euler's discovery does not apply to them.) This suggests asking how many topologically regular polyhedrons there are homeomorphic to some other topological surface, such as the torus, in which

$$V - E + F = 0$$

Euclid's theorem proceeds from a characterizing property of an entity (the sphere) not mentioned overtly in the theorem itself. The properties responsible for there being no more than five regular solids do not uniquely characterize regular solids, since 'deforming' the solids does not 'deform' the theorem. Rather they characterize a topological space, and for this reason we say that Euclid's theorem is 'really' topological in character, and that any geometrical proof is 'irrelevant'.

We have, then, this result: an *explanatory proof* depends on a characterizing property of something mentioned in the theorem: if we 'deform' the proof, substituting the characterizing property of a related entity, we get a related theorem. A *characterizing property* picks out one from a *family* ('family' in the essay undefined); an object might be characterized variously if it belongs to distinct families. 'Deformation' is similarly undefined – it implies not just mechanical substitution, but reworking the proof, holding constant the proof-idea.

We have not analyzed mathematical explanation, but explanation *by proof*; there are other kinds of mathematical explanation. Feferman (in a communcation) mentions the example of $\Pi_1^1$ sets of natural numbers, which behave somewhat like recursively enumerable sets, and that of $\Delta_1^1$ sets which behave somewhat like recursive sets. Many results from recursion theory about recursive and recursively enumerable sets carry over, but the analogy breaks down. Kreisel and Sacks explain both the analogy and the breakdowns by (a) generalizing the notion of recursion to recursion on transfinite ordinals; (b) showing that *generalized* recursive sets bounded by an ordinal less than $\omega_1$, the first nonrecursive ordinal, resemble *finite* sets from ordinary recursion theory; (c) showing that the generalized recursively enumerable sets bounded by $\omega$ are exactly the $\Pi_1^1$ sets, and that the generalized recursive sets bounded by $\omega$ are exactly the $\Delta_1^1$ sets. This would explain, for example, why Craig's theorem – providing a recursive axiom set for recursively enumerable theories – fails to provide a $\Delta_1^1$ axiomatization for a $\Pi_1^1$ theory: Craig's theorem does not guarantee a *finite* axiomatization. We explain here not a single theorem, but the 'behavior' of a mathematical object. Other examples of this include Hardy's explanation of the lawless behavior of a certain numerical function by regarding it, a la Ramanujan, as a 'snapshot' at each $n$ of the resultant of infinitely many sine waves of incompatible periods and decreasing amplitudes. It is significant, however, that both explanations involve characterizing properties: in the case of the $\Delta_1^1$ and the $\Pi_1^1$ sets, their being generalized recursive (or recursively enumerable) sets bounded by $\omega$; in the latter case, it is the Ramanujan characterization of the function. Perhaps all mathematical explanations, then, may be treated similarly.

Consider a final objection (also suggested by Feferman) to the views expressed here. Galois theory explains the prior results that the general polynomial equation is solvable in radicals if and only if the equation is of degree less than five (special cases of this had been proved by Cardan and Abel). The explanation assigns a group of automorphisms to each equation, and studies the groups instead of the equations. Let $E$ be an equation with coefficients in a field $F$. We can demonstrate the existence of a smallest field $K$ containing $F$ in which $E$ can be factored and thus solved (not necessarily in radicals) – call this the *splitting field* of $E$. The group $G$ of automorphisms of $K$ which leave $F$ alone is called the *Galois group* of $E$. Now an automorphism of $K$ (i.e. any member of $G$) is determined by its action on the roots of $E$ (for $K$ is the smallest field containing $F$ and the $n$ roots). Also, it is obvious that any member of $G$ maps a root of $E$ onto (the same or another) root of $E$. Thus the Galois group of $E$ is a certain subset of the permutations of the roots of $E$. A final definition from group theory: $G$ is *solvable* if it is the culmination of a finite chain of groups

$$0 \subset G_1 \subset G_2 \subset \cdots \subset G_n = G,$$

in which each group is a *normal* subgroup of the next ($H$ is a *normal* subgroup of $G$ if $GHG^{-1} = H$). The explanation of Cardan's and Abel's results consists of the following triad:

(a) An equation is solvable in radicals if and only if its Galois group is solvable.

(b) The Galois group of the general polynomial equation of degree $n$

consists of every possible permutation on $n$ different objects. This group is called the *symmetric group on n letters*.

(c) The symmetric group on $n$ letters is solvable if and only if $n$ is less than five (this is the whole reason for this detour into group theory; groups are easier to study than equations).

But this example, in fact, supports our view. If we take the family $E(n)$ of equations of degree $n$, then the Galois group of $E(n)$ – the symmetric group on $n$ letters – characterizes each equation as required.[13] Nevertheless, Feferman's example suggests another which does compel further development: if $E$ is an irreducible (unfactorable), solvable (in radicals) equation of prime degree $q$, then its Galois group turns out to be a group of *linear* permutations of the roots, all (therefore) with the property of fixing at most *one* of the $q$ roots of $E$, except the identity map (also in the Galois group, of course) which fixes all of them. This striking fact yields information about $E$. Suppose $E$ has rational coefficients; the splitting field, therefore, a subfield of the complex numbers. The map taking each complex number into its complex conjugate is certainly an automorphism of the complex numbers (the conjugate of the sum/product is the sum/product of the conjugates). So this map is in the Galois group of $E$. But since each real number is its own complex conjugate, the map fixes each real number. Having already obtained the result that each pemutation in the Galois group (if not the identity map) fixes at most one root, we get the following provocative theorem:[14]

If $E$ is a solvable, irreducible equation of prime degree $q$ with real coefficients, then it either has exactly one real root, or all its roots are real. (For example, consider

$$x^5 - 4x + 2 = 0.$$

Sketching $x^5 - 4x + 2$, we can see that it crosses the $x$-axis three times. Thus the equation cannot be solved.)

This proof seems explanatory – the nub of the explanation is that the Galois group here is a group of linear permutations. Yet here we don't need to know the exact Galois group, as we did before, Indeed, an arbitrary equation with rational coefficients has not a unique Galois group, in the sense that no other equation has it (though it is, of course, the only Galois group of the equation). And there is no obvious way to find an equation with rational coefficients which has a given group as its Galois group – an unsolved Hilbert problem. This is no isolated example; the contemporary style is to study domain $X$ by assigning a counterpart in domain $Y$ to each object in $X$. The object in $Y$ need not uniquely characterize anything in $X$; examples are Galois theory and algebraic topology.

For exaplanations in Galois theory (and others), the concept of 'characterization' will have to be weakened to allow for partial characterization. The Galois group of $E$ characterizes it in that the properties of the Galois group tell us much about $E$. 'Deforming' a proof in Galois theory produces results linking a new Galois group to any equation with that group – but we still must look for equations *having* the group (an unsolved problem in general). Thus our analysis (suitably developed in the direction sketched here) should account for explanatory proofs in contemporary, as well as classical, mathematics; but the detailed demonstration of this, I shall leave for another occasion.

*Hebrew University of Jerusalem*

## NOTES

* I would like to thank Sidney Morgenbesser for many hours of discussion. Isaac Levi, Jim Higginbotham, Charles Parsons, Richard Friedberg, and Howard Stein also provided helpful critique. A version of this paper was read at the University of Wisconsin-Milwaukee, and at a philosophical conference organized by Parviz Morwedge.

I am grateful to the following mathematicians who showed me examples of mathematical explanation, and their attitudes toward it: Sylvain Cappell, Paul Rosenbloom, Leon Ehrenpreis, Leslie Tharp, Harold Shapiro, Martin Davis, Georg Kreisel, Solomon Feferman, and Ellis Kolchin. Some of what I learned from them will appear in subsequent work. I acknowledge a general debt to the writings of Hao Wang.

After completing this paper (in an earlier version), I received helpful comments from Feferman, Tharp, Mike Resnik, Dale Gottlieb, Philip Kitcher, and Mic Detlefsen. I have responded to some of their detailed criticism; objections of a methodological nature I have left to the future.

Finally, thanks to the National Endowment for the Humanities, the Columbia Council for Research in the Humanities, and the National Science Foundation, for grants which enabled me to devote time to mathematics and philosophy.

[1] W. V. Quine, *Philosophy of Logic* (Prentice Hall, Englewood Cliffs, N.J., 1970); Hilary Putnam, *Philosophical Papers*, Vol. I (Cambridge University Press); Mark Steiner, *Mathematical Knowledge* (Cornell University Press, Ithaca, 1975).

[2] Solomon Feferman, 'Systems of Predicative Analysis', in Jaakko Hintikka (ed.), *The Philosophy of Mathematics* (Oxford University Press, 1969), p. 98. Feferman has subsequently argued that the passage is literally true, since it does not state that the method of abstraction and generalization succeeds in providing mathematical explanations, only that these are means pursued for the latter purpose; and, in any event, that the quoted remarks do not represent a considered opinion. Kreisel has also asserted subsequently that his words should not be taken literally as an endorsement of my (c).

[3] C. C. Chang and H. J. Keisler, *Model Theory* (North-Holland, Amsterdam, 1973), pp. 123–124.

[4] This position is discussed in Philip Kitcher, 'Bolzano's Ideal of Algebraic Analysis'., *Studies in the History and Philosophy of Science* VI (1975), 229–269. The tripartite distinction itself is due to Sidney Morgenbesser.

[5] *Ibid.*

[6] See G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 2nd ed. (Clarendon Press, Oxford, 1945), pp. 39–43, for a thorough mathematical and historical discussion of these points.

[7] This proof idea is due to Georg Kreisel.

[8] G. Polya, *Induction and Analogy in Mathematics* (Princeton University Press, 1954), pp. 16–17.

[9] Both proofs are given in Hardy and Wright, pp. 275–278, from which the present account is adapted.

[10] I am indebted to Morgenbesser for the formulation.

[11] Dale Gottlieb has objected that the inductive proof does make use of the definitions

$$S(0) = 0,$$
$$S(n+1) = S(n) + (n+1),$$

which characterize $S$. But a characterizing property is not enough to make an explanatory proof. One must be able to generate new, related proofs by varying the property and reasoning again. Inductive proofs usually do not allow deformation, since before one reasons one must have laready conjectured the theorem. Changing the equations for $S$ will not immediately reevaluate $S(n)$ – it must be conjectured anew.

[12] Adapted from Hardy and Wright, pp. 277–278.

[13] Feferman has also suggested the following counter-example: If $A = aa_1 + bb_1 + cc_1 + dd_1$, $B = ab_1 - ba_1 + cd_1 - dc_1$, etc., then – Euler discovered –

$$(a^2 + b^2 + c^2 + d^2)(a_1^2 + b_1^2 + c_1^2 + d_1^2) = A^2 + B^2 + C^2 + D^2 .$$

(This formula is used to show that every number is a sum of four squares once one knows it for every prime.) Feferman continues:

> Euler's formula can be verified by direct calculation. However, we feel we have explained why this is so when we look at the algebra of quaternions $\alpha = a + bi + cj + dk$; $\alpha_1$ may be represented as quaternion $A + Bi + Cj + Dk$ using the table of products of $i, j, k$. Taking $N(a + bi + cj + dk) = a^2 + b^2 + c^2 + d^2$, we can verify (just as for complex numbers on various algebraic number fields) that $N(\alpha\alpha_1) = N(\alpha)N(\alpha_1)$. Combining these gives Euler's formula. Now just what is the 'essence' of the entities involved in Euler's formula which is operative in this explanation?

The answer is that in Euler's formula each sum of four squares is the norm of a quaternion – substituting the norm of a complex number, we deform Euler's proof and conclude that the product of sums of two squares is the sum of two squares.

[14] The details of this proof are to be found in the supplement, by A. N. Milgram, to E. Artin, *Galois Theory*, 2nd ed. (University of Notre Dame Press, Notre Dame and London, 1971).