What's so Baffling About Negative Numbers? – a Cross-Cultural Comparison

David Mumford

I was flabbergasted when I first read Augustus De Morgan's writings about negative numbers[1]. For example, in the *Penny Cyclopedia* of 1843, to which he contributed many articles, he wrote in the article *Negative and Impossible Quantities*:

> *It is not our intention to follow the earlier algebraists through their different uses of negative numbers. These creations of algebra retained their existence, in the face of the obvious deficiency of rational explanation which characterized every attempt at their theory.*

In fact, he spent much of his life, first showing how equations with these meaningless negative numbers could be reworked so as to assert honest facts involving only positive numbers and, later, working slowly towards a definition of abstract rings and fields, the ideas which he felt were the only way to build a fully satisfactory theory of negative numbers.

On the other hand, every school child today is taught in fourth and fifth grade about negative numbers and how to do arithmetic with them. Somehow, the aversion to these 'irrational creations' has evaporated. Today they are an indispensable part of our education and technology. Is this an example of our civilization advancing since 1843, our standing today on the shoulders of giants and incorporating their insights? Is it reasonable, for example, that calculus was being developed and the foundations of physics being laid — before negative numbers became part of our numerical language!?

The purpose of this article is not to criticize specific mathematicians but first to examine from a cross cultural perspective whether this same order of discovery, the late incorporation of negatives into the number system, was followed in non-Western cultures. Then secondly, I want to look at some of the main figures in

[1] De Morgan's attitudes are, of course, well known to historians of Mathematics. But my naïve idea as a research mathematician had been that *at least* from the time of Newton and the Enlightenment an essentially modern idea of real numbers was accepted by all research mathematicians.

Figure 1. Augustus De Morgan

Western mathematics from the late Middle Ages to the Enlightenment and examine to what extent they engaged with negative numbers. De Morgan was not an isolated figure but represents only the last in a long line a great mathematicians in the West who, from a modern perspective, shunned negatives. Thirdly, I want to offer some explanation of why such an air of mystery continued, at least in some quarters, to shroud negative numbers until the mid 19th century. There are several surveys of similar material[2] but, other than describing well this evolution, these authors seem to accept it as inevitable. On the contrary, I would like to propose that the late acceptance of negative numbers in the West was a strange corollary of two facts which were special to the Western context which I will describe in the last section. I am basically a Platonist in believing that there is a single book of mathematical truths that various cultures discover as time goes on. But rather than viewing the History of Mathematics as the unrolling of one God-given linear scroll of mathematical results, it seems to me this book of mathematics can be read in many orders. In the long process of reading, accidents particular to different cultures can result in gaps, areas of math that remain unexplored until well past the time when they would have

[2] Three references are (i) Jacques Sesiano, *The Appearance of Negative Solutions in Medieval Mathematics*, Archive for History of the Exact Sciences, vol. 32, pp. 105-150; (ii) Helena Pycior, *Symbols, Impossible Numbers and Geometric Entanglements,* Cambridge Univ. Press, 1997; (iii) Gert Schubring, *Conflicts between Generalization, Rigor and Intuition,* Springer 2005.

been first relevant. I would suggest that the story of negative numbers is a prime example of this effect.[3]

This paper started from work at a seminar at Brown University but was developed extensively at the seminar on the History of Mathematics at the Chennai Mathematical Institute whose papers appear in this volume. I want to thank Professors P. P. Divakaran, K. Ramasubramanian, C. S. Seshadri, R. Sridharan and M. D. Srinivas for valuable conversations and tireless efforts in putting this seminar together. On the US side, I especially want to thank Professor Kim Plofker for a great deal of help in penetrating the Indian material, Professor Jayant Shah for his help with both translations and understanding of the Indian astronomy and Professor Barry Mazur for discussions of Cardano and the discovery of complex numbers. I will begin with a discussion of the different perspectives from which negative numbers and their arithmetic can be understood. Such an analysis is essential if we are to look critically at what early authors said about them and did with them.

1. The Basis of Negative Numbers and Their Arithmetic

It is hard, after a contemporary education, to go back in time to your childhood and realize why negative numbers were a difficult concept to learn. This makes it doubly hard to read historical documents and see why very intelligent people in the past had such trouble dealing with negative numbers. Here is a short preview to try to clarify some of the foundational issues.

Quantities in nature, things we can measure, come in two varieties: those which, by their nature, are always positive and those which can be zero or negative as well as positive, which therefore come in two forms, one canceling the other. When one reads in mathematical works of the past that the writer discards a negative solution, one should bear in mind that this may simply reflect that for the type of variable in that specific problem, negatives make no sense and not conclude that that author believed all negative numbers were meaningless[4]. Below is a table. The first five are ingredients of Euclidean mathematics and the sixth occurs in Euclid (the unsigned case) and Ptolemy (the signed case, labeled as north and south) respectively.

What arithmetic operations can you perform on these quantities? If they are unsigned, then, as in Euclid, we get the usual four operations:

1. $a + b$ OK
2. $a - b$ $\underline{\text{but only if } a > b}$ (as De Morgan insisted so strenuously)

[3] I believe the discovery of Calculus and, especially, simple harmonic motion, the differential equations of sine and cosine, in India and the West provide a second example.

[4] For example, Bhaskara II has a problem in which you must solve for the number of monkeys in some situation, and obviously this cannot be negative.

TABLE I

Modern units	*Naturally Positive Quantities*	*Signed Quantities*
positive integer	# of people/monkeys/ apples	
positive real	proportion of 2 lengths (Euclid, Bk V)	
meters	length of movable rigid bar/stick	
meters^2	area of movable rigid flat object	
meters^3	volume of movable rigid object or incompressible fluid	
degrees (of angle)	Measure of a plane angle	distance N/S of equator
dollars		fortune/debt; profit/loss; asset/liability
meters		(a) distance on line/road, rel. to fixed pt, the ‘number line’ (b) also, height above/below the surface of earth.
seconds		time before or after the present or relative to a fixed event
meters per second		velocity on a line, forwards or backwards
degrees (of temperature)	Kelvin temperature	Fahrenheit or Celsius temperature
grams	Mass or weight of an object	
gram-meters/sec.^{2}		your weight on a scale $=$ force of gravity on your body (a vector)

3. $a * b$ OK but units of the result are different from those of the arguments, e.g. length × length $=$ area, length × length × length $=$ volume

4. a/b OK but again units are different, e.g. length / length $=$ pure number, area / length $=$ length

If they are signed quantities, addition and subtraction are relatively easy – but modern notation obscures how tricky it is to define the actual operation in all cases!

TABLE II

First summand	*Second summand*	*Sum*	*Difference*
a	b	usual $a + b$	$a - b$ if $a > b$ (neg)$(b - a)$ if $b > a$
(neg)a	(neg)b	(neg)$(a + b)$	$b - a$ if $b > a$ (neg)$(a - b)$ if $a > b$
a	(neg)b	$a - b$ if $a > b$ (neg)$(b - a)$ if $b > a$	$a + b$
(neg)a	b	$b - a$ if $b > a$ (neg)$(a - b)$ if $a > b$	(neg)$(a + b)$

We write the simple expression $a - b$, and consider it obviously the same as any of these:

$$a + (-b) = a - (+b) = a + (-1) \cdot b$$

but each is, in fact, a different expression with a different meaning. Given an ordinary positive number a, $-a$ is naturally defined as the result of subtracting a from 0. For a minute, to fix ideas, *don't* write $-a$, but use the notation (neg)a for $0 - a$. Then note how complicated it is to define $a + b$ for all signs of a and b. Starting with a and b positive, Table II gives the sums and differences of a and (neg)a with b and (neg)b,

Understanding this table for the case of addition seems to be the first step in understanding and formalizing negatives. The second step is to extend subtraction to negatives so as to get the last column. This is contained in the rule:

$$a - (-b) = a + b, \text{ for all positive numbers } a, b.$$

The basic reason for this is that we want the identity $a - x + x = a$ to hold for all x, positive or negative or, in other words, subtraction should always cancel out addition. If we take x equal to $-b$, then replacing $a - (-b)$ by $a + b$ makes this identity hold. The argument one finds in some historical writings may be paraphrased as "taking away a debt of size x is the same as acquiring a new asset of size x", a fact obvious to any merchant. In any case, understanding of negatives up to this point seems to be a natural stage that one encounters in various historical documents. In modern terminology, while acknowledging that our modern words distort historical truth, one would paraphrase this stage by saying that it incorporates the idea that the integers, positive and negative are an abelian group under addition.

But multiplication of negatives is a subtler operation, the third and final step in the arithmetic of negatives. Modern notation again obscures the subtlety. When you write the simple identity $-a = (-1) \cdot a$, you are making a big step. Perhaps this is a contemporary mathematician splitting hairs because historically this seems to have been assumed as completely natural by nearly every mathematician once they knew the rules for subtracting negative numbers (with the exception perhaps of Cardano and Harriot, see below). One difficulty in arguing for this rule is that there are not many simple cases of quantities in the world where the units of the two multiplicands allow us to infer the multiplication rule using our physical intuition about the world. Here are a number of ways of arguing that the identity $(-1)\cdot(-1) = +1$ must hold.

Method I: Use the basic, intuitively obvious, identity:

$$\textit{distance} = \textit{velocity} \times \textit{time}$$

and argue that if you substitute:

(a) *velocity* = movement of one meter *backwards* per second, a negative number,

(b) *time* = second in the *past*, also negative,

(c) then one second ago, you were 1 meter ahead, i.e. *distance* = +1 meter.

This ‘proves’ $(-1)\cdot(-1) = +1$.

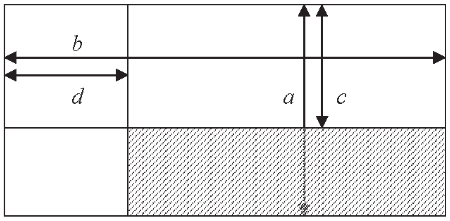
Method I′: I know of only one other real world situation where the rule is intuitively obvious. This variant of the previous argument concerns money and time. We use the simple equation obvious to any merchant describing the linear growth of a business’s assets:

$$\textit{assets at time } t = (\textit{rate of change of assets}) \times (\textit{elapsed time } t) + (\textit{assets at present})$$

Now suppose a business is *losing* \$10,000 a year and is going bankrupt right now. How much money did it have a year ago? Substitute $t = -1$, *rate* $= -10000$, *present assets* $= 0$ and the obvious fact that *assets a year ago* $= +10000$ to conclude that $(-1)\cdot(-10000) = +10000$.

Method II: (as in Euclid’s geometric algebra)

In Euclid, multiplication occurs typically when the area of a rectangle is the product of the lengths of its two sides. Consider the diagram below:



The big rectangle has area $a \cdot b$ but the shaded rectangle has area $(a-c)\cdot(b-d)$. Since the area of the shaded rectangle equals the area of the big rectangle minus the areas of the top rectangle and the left rectangle *plus* the area of the small top-left rectangle (which has been subtracted twice), we get the identity

$$(a-c) \times (b-d = ab - bc - ad + cd, \text{ if } a, b, c, d > 0, a > c, b > d$$

Now we use the idea that identities should always be extended to more general situations so long as no contradiction arises. If we extend this principle to arbitrary a, b, c, d, (which will bring in negative lengths and areas), we get for $a = b = 0$:

$$(-c)(-d) = +cd$$

This approach is probably the most common way to derive the multiplication rule. It can be phrased purely algebraically if you extend the distributive law to all numbers and argue like this (using also $0 \cdot x = 0$ and $1 \cdot x = x$):

$$1 = 1+(-1)\cdot 0 = 1+(-1)\cdot(1+(-1)) = 1+(-1)\cdot 1+(-1)\cdot(-1) = (-1)\cdot(-1).$$

Method III: Start with the multiplication

$$(\text{positive integer } n) \times (\text{any quantity } a) = (\text{more of this quantity } na)$$

(e.g. 4 × (quart of milk) = a gallon of milk), then by *subdividing* quantities as well as replicating them, you can define multiplication

$$(\text{positive rational}) \times (\text{quantity } a)$$

and by continuity (as in Eudoxus), define

$$(\text{positive real}) \times (\text{quantity } a)$$

What we are doing is interpreting multiplication of any quantity by a positive dimensionless real number as *scaling* it, making bigger or smaller as the case may be. Now if the quantity involved is signed you find it very natural to interpret reversing its sign as scaling by -1, i.e. to make the further definition:

$$(-1) \times (\text{quantity } a) = (\text{quantity } -a)$$

Now you have multiplication by any real number, positive or negative. In other words, the negative version of scaling is taking quantities to their opposites.

The core of this argument is the algebraic fact that the *endomorphisms of an abelian group form a ring* and we are constructing multiplication out of addition as *composition of endomorphisms*. This makes the third approach arguably the most natural to a contemporary mathematician trained in the Bourbaki style.

2. Negatives in Chinese and Indian Mathematics

We will discuss China first. The classic of Chinese mathematics is the *Jiuzhang Suanshu* (*Nine Chapters on the Mathematical Art*). Like Euclid, this is a compendium of the mathematical concepts and techniques which had been developed slowly from perhaps the Zhou (or Chou) dynasty (begins c.1000 BCE) through the Western Han dynasty (ending 9 CE). Unlike Euclid, it is a list of practical real world problems and algorithms for their solution, without any indication of proofs. Since then, the *Nine Chapters* had a long history of ups and downs, sometimes being required in civil service exams and sometimes being burned and nearly lost. Each time it was republished though, new commentaries were added, starting with those of the great mathematician Liu Hui in 263 CE and continuing through those in the English translation by Shen, Crossley and Lun[5]. Page numbers in our quotes are from this last edition.

Starting some time in the first millennium BCE, arithmetic in China began to be carried out using counting rods, which were arranged in rows using a decimal place notation. When doing calculations, different numbers were laid out by rods in a series of rows, forming a grid: a Japanese illustration of how they were used is shown in the figure below.

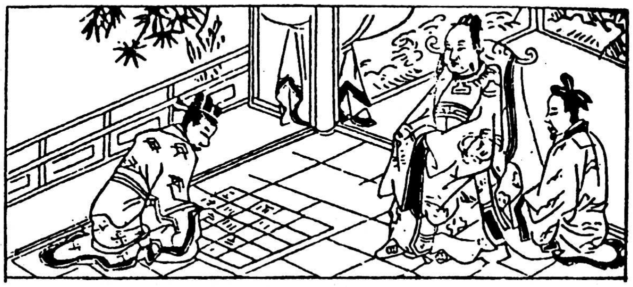


Figure 2. A Japanese illustration of calculation with counting rods

The section of the *Nine Chapters* in which negative numbers are introduced and used extensively is Chapter 8, *Rectangular Arrays*. This Chapter deals with the solutions of systems of linear equations and expounds what is, to all intents and purposes, the method of Gaussian Elimination. In fact, it is indistinguishable from the modern form. The coefficients are written out in a rectangular array of rod numerals and one adds and subtracts multiples of one equation from another equation until the system has triangular form. Examples as large as five equations in

[5] *The Nine Chapters on the Mathematical Art: Companion and Commentary*, Shen Kangshen, John N. Crossley, and Anthony W. -C. Lun, Oxford University Press, 1999.

five unknowns are worked. Naturally negative numbers appear all the time in such an algorithm.

As described in Liu's commentary, red rods or upright rods were used for positive numbers which he calls gains (*zheng*) and black rods or slanting rods for negative numbers which he calls losses (*fu*). He says "*red and black counting rods are used to cancel each other*". Curiously, his colors are the exact opposite of our Western accounting convention! Here is Problem 8 from this Chapter, p.409 in the Shen, Crossley and Lun edition:

> *Now sell 2 cows and 5 sheep, to buy 13 pigs. Surplus: 1000 cash. Sell 3 cows and 3 pigs to buy 9 sheep. There is exactly enough cash. Sell 6 sheep and 8 pigs. Then buy 5 cows. There is 600 coins deficit. Tell: what is the price of a cow, a sheep and a pig respectively?*

This means the three equations (all of which have negative coefficients as well as positive):

$$\begin{aligned} 2C + 5S - 13P &= 1000 \\ 3C - 9S + 3P &= 0 \\ -5C + 6S + 8P &= -600 \end{aligned}$$

The solution is found to be $C = 1200$, $S = 500$, $P = 300$. The *Nine Chapters* goes on rather mysteriously (p.404):

> *Method: Using rectangular arrays lay down counting rods for each entry to be added.*
>
> *The Sign Rule*
>
> *Like signs subtract; opposite signs add; positive without extra, make negative; negative without extra, make positive.*
> *Opposite signs subtract; same signs add; positive without extra, make positive; negative without extra, make negative.*

Liu's commentary explains: the first set of sign rules refers to subtraction of array entries, the second to addition. He goes on to clarify the meaning of the cryptic Sign Rule. In fact, the rule is precisely what we wrote out in Table II above for both addition and subtraction. What is clear is that negative numbers were analyzed and treated correctly as soon as the need arose, presumably for the first time anywhere in the world.

I cannot find in the Shen et al edition of the *Nine Chapters* any treatment of multiplication of negative numbers, although Martzloff[6] quotes the Chinese edition of Qian Baocong as saying: "*Rods of the same name multiplied by each other make positive. Rods of different names multiplied by each other make negative*". In any

[6] Jean-Claude Martzloff, *A History of Chinese Mathematics*, 2nd edition, Springer, 1997, page 203.

case, Liu's commentary, written in the 3rd century CE, makes the remark (p.405): "*Interchanging the red and black rods in any column is immaterial. So one can make the first entries of opposite sign.*" This is the correct rule for multiplication by -1.

Chinese algebra had a renaissance in the Song and Yuan (Mongol) dynasties. In particular, Zhu Shijie (c.1260–c.1320) extended the ideas of Gaussian elimination to the simultaneous solution of *polynomial* equations, inventing the equivalent of the resultant and using ever larger and more complex arrays of coefficients. At this stage, as one would expect, the full rules for negative arithmetic emerge quite explicitly as well those for the algebra of polynomials. Having a theory of negatives is the clear prerequisite for going further in the study of algebra. Zhu's algebra reached a stage not attained in Europe until the late 19th century.

I want to turn to India next. In every culture, one of the main reasons for the development of arithmetic – arguably the principle driving force – is the need of merchants to keep accounts. In fact, it is even hypothesized that arithmetic and writing itself emerged in the 3rd millennium BCE in Mesopotamia as a development of a crude system of tracking transactions of agricultural goods by means of small specially shaped and inscribed tokens[7]. By around the year 2000 BCE, one finds tablets from Ur[8] with a yearly summary accounting, showing budgeted and actual inputs (with value converted into a common unit of barley), budgeted and actual outputs, budgeted and actual labor and differences, shortfalls or profits! In India, very sophisticated principles of accounting were codified in Kautilya's comprehensive manual of statecraft, the Arthaśātra written in the 4th century BCE. The Arthaśātra covers in amazing detail every aspect of setting up and managing of a kingdom (including managing a special forest for elephants). In Book II, Chapter 6 and also in many later Chapters of Book II, Kautilya details how accounts are to be kept[9]. He describes a complete system of book keeping: he has a ledger for *income* with dates, times, payers, categories, etc. and a ledger for *expenditures* and finally a third ledger for *balances*. There are sections on auditing, insurance against theft, debtors, borrowings, mortgages, auditing, etc. and subtler accounting issues such as current vs. deferred receipts, how to account for price changes of items in inventory, fixed vs. variable costs. Although he does not use negative numbers explicitly, he is

[7] The pioneer here has been Denise Schmandt-Besserat, who has brought her life's work together in the multi-volume book *Before Writing*, volume I being *From Counting to Cuneiform*, University of Texas Press, 1992. In particular, she has "deciphered" the mysterious tokens found throughout the Middle East from roughly 8000 BCE to 3000 BCE, finding a simple method of accounting which merges seamlessly with highly developed cuneiform accounts in the 3rd millennium.

[8] See Chapter 5 in Richard Mattessich, *The Beginnings of Accounting and Accounting Thought*: *Accounting Practice in the Middle East (8000 B.C to 2000 B.C.) and Accounting Thought in India (300 BCE to the Middle Ages),* Garland Publishing, 2000.

[9] See Chapter 6 in Mattessich, *Op. Cit.,* which is based on the book *Modern Accounting Concepts in Kautilya's Arthasastra* by Anjan Bhattacharyya, Firma KLM, Calcutta, 1988.

clearly aware of how accounts must sometimes show a deficit and that people may carry a negative net worth.

Although the Arthaśātra does not mention negative numbers explicitly, they appear full blown in Brahmagupta's treatise *Brâhma-sphuta-siddhânta* (628 CE). The development of mathematics in India in the first millennium CE is connected much more strongly to astronomy than to accounting. For much of this period, treatises covering both mathematics (*ganita*) and astronomy (the motion of the sun, moon and planets and their positions at a given time and place in the sky) and called *siddhantas* were composed. Many of these were in verse, highly compressed and cryptic, meant to be memorized and handed down generation by generation from teacher to student.

The *Brâhma-sphuta-siddhânta* includes two Chapters on mathematics which are a compendia of the mathematical concepts and techniques which had been developed over previous centuries. Here we find all the correct rules for arithmetic with negative numbers and in it *positive numbers are referred to as "fortunes", negative numbers as "debts"*. It appears that accounting led naturally to an arithmetic in which negative numbers took their natural place. Here are some quotations, showing first the rules we laid out in table I and then, significantly, going on to describe how to multiply negative numbers[10]:

> *[The sum] of two positives is positive, of two negatives, negative; of a positive and a negative [the sum] is their difference; if they are equal, it is zero. The sum of a negative and zero is negative, of a positive and zero positive, of two zeros, zero.*
>
> *[If] a smaller [positive] is to be subtracted from a larger positive, [the result] is positive; [if] a smaller negative from a larger negative, [the result] is negative; [if] a larger from a smaller, their difference is reversed – negative becomes positive and positive negative.*
> . . . .
> *The product of a negative and a positive is negative, of two negatives positive, and of positives positive; the product of zero and a negative, of zero and a positive, or of two zeros is zero.*
>
> *A positive divided by a positive and negative divided by a negative is positive; a zero divided by a zero is zero; a positive divided by a negative is negative; a negative divided by a positive is negative.*
>
> *Chapter 18, verses 30–34*

The only oddity seems to be his confident assertion that $0/0 = 0$. The rest is as clear and modern as one could wish for. It would be wonderful to know what considerations led Indian mathematicians in the late centuries BCE or the early centuries CE to these conclusions – especially for the multiplication of negative numbers. The predominately oral transmission of knowledge in the Vedic

[10] We quote from the translation by Kim Plofker in her book, *Mathematics in India, 500 BCE – 1800 CE*, Chapter 5, p.151.

tradition – and perhaps the difficulty of preserving perishable writing materials through yearly monsoons – has not left us with any record of these discoveries. They just appear full blown in Brahmagupta's summary. R.Mattessich has developed at length the idea that it was the highly developed tradition of accounting which led to the full understanding of negative numbers[11] but unfortunately no evidence for this plausible conjecture exists.

As in China, having negative numbers opened the way to deeper studies of algebra itself. Perhaps the deepest of these was the Indian work on Pell's equation $x^2 - Ny^2 = m$, especially finding solutions for $m = 1$. Brahmagupta himself made the first huge step, discovering the multiplication law arising from the factorization

$$x^2 - Ny^2 = \left(x + \sqrt{N}y\right)\left(x - \sqrt{N}y\right).$$

More exactly, he showed how from solutions of the equation for m_1, m_2, one gets one for their product $m = m_1 \cdot m_2$. Some centuries later, Jayadeva found a complete algorithm for constructing solutions with $m = 1$.

We find reflections of the Indian use of negatives in their astronomy too. As stated, the main goal of these scholars was not to develop mathematics for its own sake but to apply mathematics to predict the positions of the sun, moon and planets. An epicyclic theory is used and, for the planets, both a 'slow' and 'fast' correction is added to the mean motion of the planet (in our terms, one is due to the ellipticity of their orbit, the other to the shift from a heliocentric to a geocentric description). David Pingree[12] has hypothesized that through the intermediary of the Indo-Greek empire, some version of the pre-Ptolemaic Hipparchan theory of planetary motion reached India. What is quite striking is that in making these corrections the sine function *in all four quadrants* is understood. Hipparchus had computed tables of *chords*, which are fundamentally unsigned positive quantities. The Indian tradition shifts to sines (actually 'Rsines', sines multiplied a large radius and rounded to the nearest integer) and then it is natural to extend them from the first quadrant to the full circle. Here is a quote from the *Brâhma-sphuta-siddhânta*, Chapter 2, verse 16 describing the corrections made by adding or subtracting appropriate sine function corrections to the mean position:

> *(In successive quadrants) (in the slow case) negative, positive, positive, negative correction, otherwise in the fast case. (The sum) of two positives (is) positive, of two negatives (is) negative, of positive and negative (is their) difference, of equals (positive and negative is) zero.*[13]

[11] See Chapter 7 in Mattessich, *Op. Cit*.

[12] David Pingree, The History of Mathematical Astronomy in India, in *Dictionary of Scientific Biography*, Charles Gillespie editor, Scribner, 1978, pp.533–633.

[13] Translation by J. Shah (personal communication).

It would be nice if they had drawn a graph of the correction in all quadrants, i.e. of the sine function, to clarify this verse, but that was clearly not their *modus operandi*. But further evidence that the sine function was seen as being extended to more than one quadrant comes from the rational approximation of the sine in the first two quadrants given by Bhaskara I (7th century CE)[14]:

$$\sin(\theta) \approx \frac{16 \cdot \theta(\pi - \theta)}{5\pi^2 - 4\theta(\pi - \theta)},\ 0 \leq \theta \leq \pi$$

This is an extraordinarily accurate approximation which would be hard to come up with if they had not grouped the first and second quadrant together.

Another natural place for using negative numbers is for coordinates, e.g. to measure the celestial latitude (perpendicular to the ecliptic), or the declination (perpendicular to the celestial equator), of a planet or star. Tradition, however, sanctifies describing latitudes and declinations as north/south instead of positive/negative and this is hard to change. But this latitude must often be put into formulas when converting from celestial coordinates to horizon based coordinates, e.g. when calculating the very important rising times of planets. At this point, rules for negatives again must be used. Here is an example from Brahmagupta's *Khandakhâdyaka,* Ch.6, verse 5[15]

> *Multiply the* <u>*north*</u> *celestial latitude by the equinoctial shadow and divide by 12; apply the quotient taken as minutes negatively or positively to (the longitude measured from) the orient and occident ecliptic points. When the celestial latitude is* <u>*south*</u>*, apply the resulting minutes to the same points positively or negatively.*

In modern terms (see Figure 3), he is computing (*longitude KA*) ± (*latitude KV*) × tan(ϕ), where ϕ is the observer's latitude and distinguishing the cases where longitude is measured eastwards or westwards and where the planet's latitude is north or south.

An explicit interpretation of negatives as coordinates on a number line occurs later in the work of the 12th century Bhaskara II (so-called to distinguish him from the earlier 7th century Bhaskara I). He wrote an immensely popular textbook on Algebra, the *Lîlâvatî*[16]. The title was said by a Persian translator to be the name of Bhaskara's daughter and, although this is not made explicit in the book, it is full of verses addressed to the "beautiful one", "the fawn-eyed one", etc. Present day texts are so drab in comparison!

The remarkable passage is in verse 166 and again it is given without any fanfare stating that a new interpretation of negative numbers is being given. But, to my

[14] Bhaskara I, *Mahabhaskariya*, Ch. 7, verses 17–19.

[15] *The Khaṇḍakhâdyaka of Brahmagupta*, with the commentary of Bhaṭṭotpala, edited and translated by Bina Chatterjee, World Press, Calcutta, 1970, p.122–3.

[16] We follow the classic translation by H. T. Colebrooke, first published by in 1817 and subsequently reprinted in numerous editions.

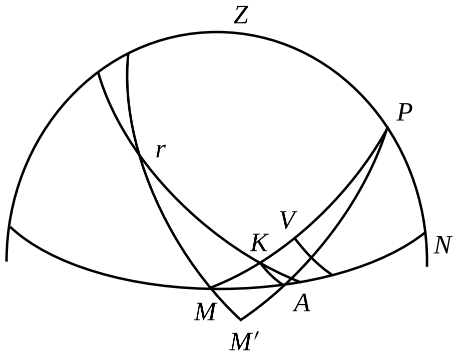


Figure 3. Diagram for calculation of rising time. Planet at V, rKA ecliptic, P north celestial pole.

knowledge, it is the first occurrence of the "number line", of using positive and negative numbers as coordinates on either side of an origin. Bhaskara is in the middle of a discussion of triangles and, specifically, given the three sides a, b, c of the triangle with a distinguished side c, the base, how to find the altitude and the position of the foot of the perpendicular dropped on the base. If you let x be the distance from one endpoint of the base to the foot, then $(c - x)$ is the distance from the other endpoint to the foot and Pythagoras's theorem tells us:

$$a^2 - x^2 = \text{altitude}^2 = b^2 - (c - x)^2$$

which gives us:

$$x = (a^2 + c^2 - b^2)/2c$$

In verse 166, he poses the problem:

> *In a triangle, wherein the sides measure ten and seventeen and the base nine, tell me promptly, expert mathematician, the segments, perpendicular and area.*

and his formula gives him $x = -6$, $c - x = 15$ (see Figure 4). Aha: what to do? Well, if you draw this triangle, you find the foot of the perpendicular lies outside the base. So what does Bhaskara say?

> *(The result 6) is negative, that is to say, in the contrary direction. Thus the two segments are found 6 and 15. From which, both ways too, the perpendicular comes out 8.*

This is stated so casually, as if it were common wisdom, that one can only conclude that this way of thinking about negative distances was well-known in his

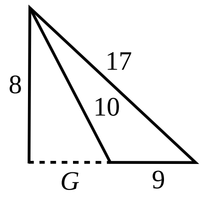


Figure 4. A triangle with a perpendicular falling outside the base, Bhaskara II

time. Nonetheless, as we will see, it doesn't occur in Europe before the work of Wallis near the end of 17th century.

3. The Shunning of Negative Numbers, From Al-Khwarizmi to Galileo

I now turn to the Arab and Western treatment of negative numbers. To keep the story within bounds, I will pick a small selection from the many figures who might be discussed, those who seem to me key figures in the story or who exemplify a particular stand.

Al-Khwarizmi (c.790–c.840)

It is repeated everywhere that the Indians invented zero and place notation and that the Arabs learned it from them and later transmitted this to Europe. It's bizarre that such a misunderstanding should be widespread but in fact, the Babylonians invented place notation (albeit using base 60) and their arithmetic was used by many Greeks, e.g. Ptolemy. I hope I have made the case that the most substantial arithmetic discovery of the Indians – and independently the Chinese – was not merely that of zero but the discovery of negative numbers. Sadly this discovery was not absorbed in any but a superficial way by the Arabs.

Al-Khwarizmi (whose full name was Abu Ja'far Mohammad ibn Musa Al-Khwarizmi) was familiar with Indian mathematics and astronomy and apparently with Brahmagupta's *Brâhma-sphuta-siddhânta* written some 200 years earlier. He worked under the patronage of the caliph Al-Mamun about whom he says "*That fondness for science, . . . , that affability and condescension which he* (the caliph) *shows the learned . . . has encouraged me to compose a short work on calculating by Completion and Reduction .. such as men constantly require in cases of inheritance, legacies, partition, law-suits and trade . . .*"[17] His book on Algebra is entitled *Al-jabr w'al muqabala* which refers to the operations of completion and reduction with

[17] *The Algebra of Mohammed ben Musa,* Translated by Frederic Rosen. Facsimile reprint of 1831 edition by the Oriental Translation Fund, London, Adamant Media Corporation 2002.

which he simplified his equations. These were relations between an unknown, its square and constants, given in prose. Nearly half of his book concerns incredibly complex inheritance cases.

I find three things especially striking in this book. Firstly, negative numbers appear only once, in a section on multiplication whose goal appears to be to explain the identity

$$(a-c)\cdot(b-d)=ab-ad-bc+cd$$

and justify it by geometry, just as in our discussion of "Method II" for multiplying negative numbers. But then they are never mentioned again. The second striking thing is that quadratic equations always have positive coefficients and thus belong to three types:

1. $ax^2+bx=c, a, b, c>0$ (referred to as "roots and squares are equal to numbers")
2. $ax^2+c=bx, a, b, c>0$ ("squares and numbers are equal to roots")
3. $ax^2=bx+c, a, b, c>0$ ("roots and numbers are equal to squares")

This separation of cases continues down through the whole European tradition through De Morgan. An equation, in short, must be an identity between two positive numbers. Thirdly, he discusses exactly the same problem that Bhaskara II was to take up: finding altitudes of triangles whose sides are given. But, unlike Bhaskara, all the examples he treats have the foot of the perpendicular *inside* the base so this big clue about negatives never comes up.

Leonardo of Pisa (1170–1250)

Leonardo of Pisa was one of the first Europeans to master the Arab arts of calculation, including the use of Indian symbols and place notation. He wrote a remarkable book, his *Liber Abaci* (Book of Calculation), in which the rules for all the basic arithmetic operations are laid out in great detail and exhaustively illustrated by numerical examples. This occupies the first half of his book which is essentially what we would call a primer. But he deals exclusively with the arithmetic of positive integers and positive fractions. His section on subtraction is entitled *On the Subtraction of Lesser Numbers from Greater Numbers*.

As in the Indian tradition, accounting was one of the principle stimuli for the development of arithmetic in the Middle Ages and much of the book deals with the arithmetic of money, goods and possessions. The second half of the book treats a huge number of "word problems" involving goods and money. He is following a curious tradition going back to Diophantus (and found in Chinese and Indian works also) of what, to modern eyes, are quite bizarre artificial "word problems" involving a group of people who, after exchanging various sums of money, have sums satisfying some linear relationships. Here is an example[18]:

> *Three men had pounds of sterling, I know not how many, of which one half was the first's, one third was the second's and one sixth's was the thirds; as they wished to have it in a place of security, every one of them took from the sterling some amount, and of the amount that the first took he put in common one half, and of it that the second took, he put in common a third part, and of that which the third took, he put in common a sixth part, and from that which they put in common every one received a third part, and thus each had his portion.*

In modern algebra terms, if S is the sum of sterling and x_1, x_2, x_3 are the sums which the three men took, so that $\left(x_1/2 + x_2/3 + x_3/6\right)$ is what "*they put in common*", then the last sentence "*each had his portion*", sets up three equations:

$$\frac{x_1}{2} + \frac{1}{3}\left(\frac{x_1}{2} + \frac{x_2}{3} + \frac{x_3}{6}\right) = \frac{S}{2}$$
$$\frac{2x_2}{3} + \frac{1}{3}\left(\frac{x_1}{2} + \frac{x_2}{3} + \frac{x_3}{6}\right) = \frac{S}{3}$$
$$\frac{5x_3}{6} + \frac{1}{3}\left(\frac{x_1}{2} + \frac{x_2}{3} + \frac{x_3}{6}\right) = \frac{S}{6}$$

This is only one of hundreds of such problems. He develops methods of laying out the coefficients in rows and manipulating the numbers to get the answer. In the above, the 'answer', is the smallest set of relatively prime x's which solve these three homogeneous equations in 3 unknowns. Leonardo has a rather awkward and special version of the Chinese algorithm for solving linear equations in many unknowns.

Now most of his problems are set up so all the numbers which occur are positive. But not all! First of all, negative numbers can arise in the course of the calculation. He then says things like[19]:

> [he is in the middle of an algorithm]. . . *and from the* 240 *you subtract* 288 *leaving minus* 48, *and this I say because the* 288 *cannot be subtracted from the* 240; *from this* 48 *you take* 1/3 *for the* 1/3 *of the second position; there will be minus* 16 . . . .

He is getting close to the red and black rods of the Chinese, but these examples are few and far between and are not pursued very far. In a few other cases, the answer itself is negative. For example, after solving the problem described in the first quote, he varies the proportions of S owned by the three men to 1/2, 2/5 and 1/10. In this case, the solution is $x_1 = 326$, $x_2 = 174$ and $x_3 = -30$. The setting of the problem, that all the x's are amounts of money, comes to his rescue. The third man, he says,

[18] Leonardo of Pisa, *Liber Abaci*, p.415 of the English translation by L. Sigler, *Fibonaccis' Liber Abaci*, Springer-Verlag, 2002.

[19] Ibid, p.419.

does not take anything from the sum *S* which they share but instead puts in an additional 30 pounds of his own “proper” money: there were 470 pounds in all, and when they “wanted to have it in a place of security”, the third man *added* 30, the first man took 326 and the second took 174. When money is concerned, negative quantities can always be given a simple meaning!

Leonardo is making the first tentative steps towards enlarging the number system to include negatives. With money, he is comfortable with assets and debts, giving and taking. But his examples are few and he never makes explicit rules for extending arithmetic.

Nicole Oresme (1323–1382)

Nicole Oresme was a mathematically inspired scholastic, working in Paris in the mid-14th century. He made a giant stride taking geometry beyond Euclid. In his great book, *Tractatus de configurationibus qualitatum et motuum* (*Treatise on the configurations of qualities and motions*)[20], he proposed considering all intensities which varied in time and whose values at different times could be compared by a proportion. To any such quality, he proposed constructing a *graph*. First he took a line segment, called the *subject*, whose points represented the interval of time over which the quality was varying. This, in itself, was a radical departure from Euclid: now space was being used *analogically*, as a substitute for time. Then he proposes erecting line segments perpendicular to the subject whose lengths had the same proportions as the qualities being graphed:

> *Therefore, every intensity which can be acquired successively ought to be imagined by a straight line perpendicularly erected on some point of the space or subject of the intensible thing, e.g. a quality. For whatever ratio is found to exist between intensity and intensity of the same kind, a similar ratio is found to exist between line and line, and vice versa. . . . Therefore, the measure of intensities can be fittingly imagined as the measure of lines.* (Oresme, I.i)

He talks about graphing many things (although he never gathers data or actually goes beyond making simple cartoons of his graphs – see Figure 5). In particular, he discusses graphing velocity, temperature, pain and grace (of a soul). Some of these are clearly positive quantities by nature, e.g. pain and grace. He is interested in contrasting intensities which are constant (graph (a) in figure), intensities which vary at a constant rate (graph (b) in the figure) and intensities which are more complex (graphs (c) in figure). For example the grace of a soul ‘*occupied by many thoughts and affected by many passions*’ will be difformly difform – his name for type (c). On the other hand, velocities can clearly change sign and, as for temperature, he even considers there to be complementary intensities of hotness and coldness. For

[20] Translations are from Marshall Clagett’s translation, *Nicole Oresme and the Medieval Geometry of Qualities and Motions,* University of Wisconsin, 1968.

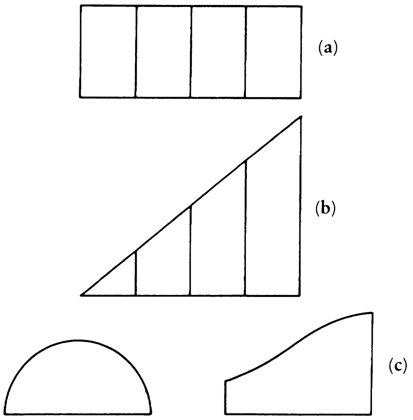


Figure 5. Oresme's examples of graphs

temperature, hotness might have a graph with values $f(x)$ and coldness a graph with values $C - f(x)$. In other words, he adds a suitable positive constant so as to make every intensity positive everywhere.

Because his graph is the whole area, not simply the curve at the tips of the his line segments, he cannot have a graph which goes from positive to negative, crossing the 'subject'. This is especially striking because at one point he makes a catalog of various types of difformly difform graphs: but no graph in the catalog is, for example, regularly oscillating like a sine wave. He even hints at the fact that the area of the graph of velocity is the distance traveled, the fundamental theorem of calculus, but to make his picture, the velocity cannot change sign: no backtracking. Oresme has gone beyond Euclid in a striking way but he cannot make the further leap of allowing negative values for an intensity.

Luca Pacioli (1445–1517)

Pacioli's importance is not due to his discoveries but to the fact that he wrote an encyclopedic work *Summa de arithmetica, geometria, proportioni et proportionalita* which summarizes the contemporary knowledge of arithmetic, geometry and especially accounting. The work's greatest influence was due to its description of double-entry book keeping which was a key step in the expansion of the international business enterprises which characterized the Renaissance. Here we find a small number of linear equations involving amounts of money whose solution is negative. As in Leonardo, when the result was a negative number, it is described as a debt. In one case, the price of an egg comes out negative – owning the egg puts you in debt so the sellers are paying you to take their eggs.

Sesiano (*op.cit.*), however, tracked down one isolated instance of a problem in Pacioli's writings which is more exciting. There is an untitled manuscript, written for his students in Perugia, which survives in the Vatican[21]. A standard class of problems (going back to Babylonian times) involves dividing a number into two parts which satisfy some quadratic condition. After solving some such problems with positive solutions, he comes to what he calls the *bellissimo caso.* This example asks you to divide 10 into two parts *the difference of whose squares* is 200. The reader may like to check that the answer is $10 = 15 + (-5)$. Here is a problem not only in pure numbers one of which is negative but requiring squaring this negative number. Although an obscure and forgotten footnote to history, it seems that the young Pacioli ventured briefly into uncharted territory in a truly original way. It is unfortunate that in his *Summa*, he did not pursue these ideas.

Girolamo Cardano (1501–1576)

The only reason to include Cardano is that he wrote the book *Ars Magna*[22], so we can analyze how he thought, how he looked on negative as well as imaginary numbers. The solution of cubic equations was due to Scipione del Ferro, Professor of Mathematics at Bologna around 1515, and the solution of the quartic to Cardano's student Ludovico Ferrari. Cardano himself was an arrogant man, a compulsive gambler, who led a wild life of ups and downs. That he computed the odds of various sorts of gambling was arguably his greatest mathematical achievement.

If Al-Khwarizmi had spun out the solutions for quadratic equations in to many different cases, Cardano really went to town describing how to solve 13 distinct cases of cubic equation (and 44 types of derivative cases). Why so many? Because (a) the coefficients all had to be positive and (b) the equation had to equate a positive quantity to another positive quantity. The many sections are entitled things like "*On the cube and square equal to the first power and number, generally*". Nonetheless, he did recognize that some of his equations had negative solutions: these he called

> "*fictitious (for such we call that which is a debitum or negative)*"

but he does very little with such roots, ignoring them systematically. But in the later Chapter, "*On the rule for postulating a negative*", he does explore a bit what algebra

[21] Cod. Vat. lat. 3129.

[22] Quotations are from the 2007 Dover reprint *The Rules of Algebra: (Ars Magna)*, translated by T. Richard Witmer.

can do for you if you admit negative roots. His example of a problem requiring negative numbers is this:

> *The dowry of Francis' wife is 100 aurei more than Francis' own property, and the square (?) of the dowry is 400 more than the square of his property. Find the dowry and the property.*

This works out to give Francis –48 aurei of property, that is, he is in debt 48 aurei, but fortunately is getting a dowry of 52 aurei. Here he correctly identifies the negative solution with a debt. This is an excellent illustration although squaring a sum of money is a pretty weird thing to do.

There would little else to say except for the curve ball that was thrown to Cardano: for all cubic equations which have only one real root, del Ferro's formula worked like a charm. But if there were three real roots (the other possibility, known as the *casus irreducibilis*), it gave an apparently meaningless result. His formula for the roots of the equation $x^3 + ax + b = 0$ is:

$$x = \sqrt[3]{\left(b/2 + \sqrt{-D/4.27}\right)} + \sqrt[3]{\left(b/2 - \sqrt{-D/4.27}\right)}, \text{ where } D = -4a^3 - 27b^2$$

D, the discriminant, is equal to the square of the difference of all pairs of distinct roots, hence it is positive if all the roots are real. So we need to find the a square root of a negative number even though in the end we only want the real number x. Cardano struggled unsuccessfully with what this might possibly mean.

His one attempt to deal with these complex expressions is in the same Chapter, "*On the rule for postulating a negative*" mentioned above. Here he considers problems which have complex roots, such as the following:

> *Divide 10 into two parts the product of which is 40.*

The usual quadratic formula gives the two parts as $5 + \sqrt{-15}$ and $5 - \sqrt{-15}$. This is also the answer his math gives him and which he puts in writing in his book but he doesn't attribute much meaning to it. He makes his famous comment:

> *So progresses arithmetic subtlety, the end of which, as is said, is as refined as it is useless.*

At the end of this Chapter, he gives a third type of example where he reasons incorrectly with products of a real and an imaginary. In a later edition, he added an appendix *De aliza regula liber* in which he flirted with the idea that maybe $(-1)^2 = +1$ was wrong. Why not try $(-1)^2 = -1$? Between 'fictitious' negative numbers and useless imaginaries, you get the sense that Cardano was at sea.

Galileo (1564–1642)

Perhaps mathematicians were stuck thinking that negative numbers were fictitious but surely physicists who were actually measuring things in the real world, had a clearer view? Arguably, Galileo's great contribution to physics was his recognition that momentum was a key property of objects, that it was constant when no forces were acting and that the force of gravity acting on projectiles and falling bodies changed their momenta at a constant rate, not their positions. As an old man, when the Pope commuted his sentence for heresy to house arrest, he wrote down these theories in his *Dialog concerning Two New Sciences*[23]. He starts off with his foil Simplicio getting put down again and again by Galileo's mouthpiece Salviato. But by the Fourth Day, Galileo lapsed into a more standard Euclid-style exposition and puts out the centerpiece of his theory: the demonstration that a projectile follows a parabolic arc under the force of gravity. Here was something he had actually experimented with and he was on solid ground, theoretically as well as experimentally. Figure 6 is an excerpt from his notebooks working on projectiles.

The central assertion in these dialogs is that gravity endows the projectile with a constant downward acceleration. Thus its vertical velocity will be positive going up, zero at the peak and negative coming back down. It is a linear function changing from positive to negative. The math couldn't be simpler – *if you are willing to use negative numbers*.

What does Galileo do? His main result is:

Theorem 1. *A projectile which is carried by a uniform horizontal motion compounded with a naturally accelerated vertical motion describes a path which is a semi-parabola.*

Note that he uses a semi-parabola: the half of the parabola in which height is a monotone function of time. Considerably later, after a long discussion of the time and distance of the semi-parabolic arc carrying the projectile to the ground, he reverses time without any discussion and concludes that the rising phase of a projectile is also a semi-parabola.

The discussion continues on optimal angles at which to fire guns. But the astonishing point is that he never talks about the whole parabolic arc, with ascending and descending halves and how there is constant downward acceleration throughout the path. *All* the diagrams in the book resemble the figure from his notes: a semi-parabola with some auxiliary chords and tangents. He analyzes the geometry of the semi-parabola and the physics of a falling body and then asserts without any discussion that one can reverse the direction of motion from a fall to a climb – nothing else. That the velocity changes at the apex from positive to negative is not stated anywhere.

[23] Quotations are from the 1956 Dover edition, *Dialogues Concerning Two New Sciences, translated by* by Alfonso De Salvia Henry Crew.

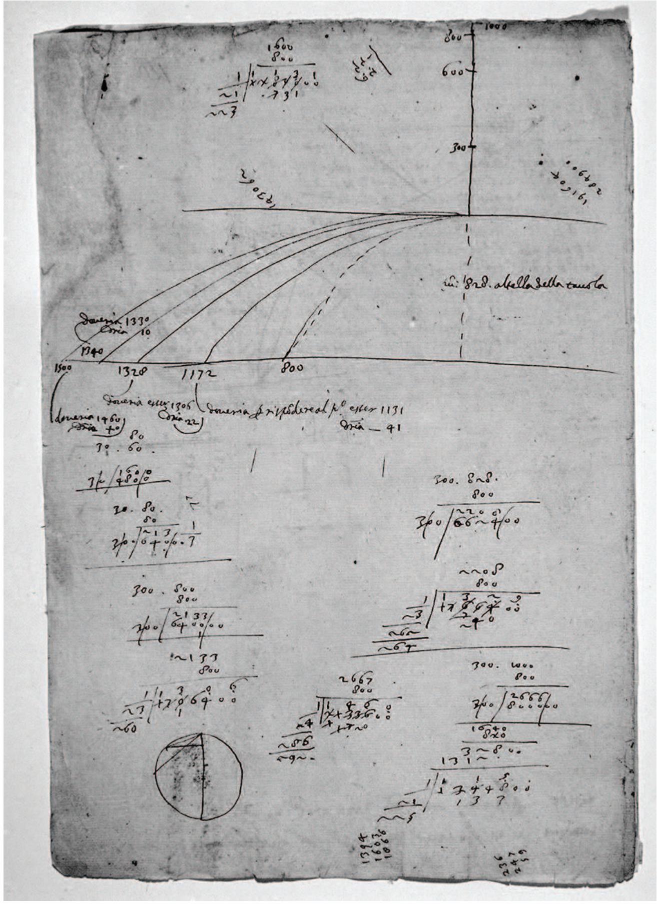

1000
800
600
300
1330
1340
1500
1328
1172
800

Figure 6. Galileo's notes on projectiles

Fermat (1601–1665)

Fermat and Descartes, at essentially the same time, had the idea of introducing coordinates into the plane and connecting geometric loci with polynomial equations in two variables. Plane curves are not confined to the positive quadrant, so one might expect that their logic would have pushed them to allow their variables to take on both positive and negative values. But no! Their coordinates were only in a positive quadrant and the other parts of a curve were treated separately if at all.

Below are two figures from Fermat's paper on the subject, *Ad Locos Planos et Solidos Isagoge*, (*Introduction to Plane and Solid Loci*). Incidentally, *plane* loci meant lines and circles, *solid* loci meant the other conic sections, terminology which dates from Greek times.

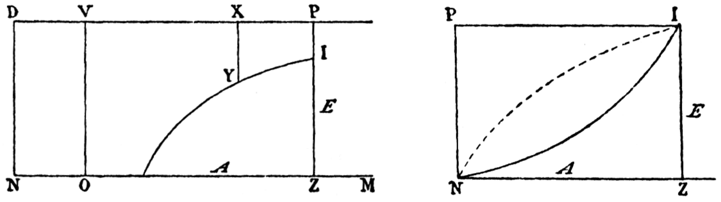


Figure 7. Two diagrams from Fermat's *Isagoge*

In these figures N is the origin, NM is the x-axis (although Fermat used the letter A, not x because his variables were vowels), ND or NP is the y-axis (the letter E for Fermat), $x = NZ$, $y = ZI$ so I is the point with coordinates (x, y). On the left, he is describing the locus of the equation:

$d^* + x \cdot y = r \cdot x + s \cdot y$ (which he writes Dpl.$+ A$ in E aeq. R in $A + S$ in E).

Here $s = NO, r = ND$ and d^* is a constant area, so we have a rectangular hyperbola, centered at V, with asymptotes VO and VP. The curious point is that he draws only this small part of the hyperbola, cutting it off on the x-axis. He also cuts it off at the plotted point I. On the right, he is describing a parabola with equation:

$$x^2 = d \cdot y \qquad \text{(which he writes } Aq \text{ .aequatur } D \text{ in } E\text{)}$$

Again, he cuts the locus off at his axes (and at I).

Descartes' treatment is similar, except that he does say in the text that there are multiple orderings possible for the relevant points on the axes and that you must set up different equations depending on the directions and ordering of both the variable point and the constants in the construction. The goal is to make both sides of your equation sums of positive quantities, just as in Al-Khwarizmi and Cardano's work on quadratic and cubic equations. Note that this is how Fermat's version of the equation for the hyperbola reads.

4. Clarifying the Muddle: Wallis and Newton

So when did European mathematicians begin to make their peace with negative numbers? The first treatment which seems to me quite modern is that of John Wallis (1616-1703), Professor of Mathematics at Oxford. He published his *Treatise on Algebra*[24], written in English, in 1685. This was just two years before Newton published by his earth-shaking *Principia Mathematica* and well after Newton had done his major work in mathematics. In his mathematical notes, where he used

[24] Available online at http://eebo.chadwyck.com through subscribing universities.

algebra and coordinates, Newton was equally modern in his treatment of negative numbers, putting them on equal footing with positive numbers. So we should attribute the first clear European view of negative numbers to Wallis and Newton equally.

In Chapter 16, *Addition, Subduction, Multiplication and Extraction of Roots in Specious Arithmetic*, Wallis defines negative numbers as nicely, simply and clearly as you could wish (here '*Specious*' is Viete's term for arithmetic with variables given by letters):

> *To these Notes, Symbols or Species are prefixed (as occasion requires) not only numeral figures, but the signs* $+$ *and – (or plus and minus), the former of which is a Note of Position, Affirmation or Addition; the other of Defect, Negation or Subduction: According as such Magnitude is supposed to be, or to be wanting. And where no such Sign is, it is presumed to be Affirmative and the sign* $+$ *is understood.*
>
> *And accordingly these Signs are still to be interpreted as in a contrary signification. If* $+$ *signify Upward, Forward, Gain, Increase, Above, Before, Addition, etc. then – is to be interpreted of Downward, Backward, Loss, Decrease, Below, Behind, Subduction, etc. And if* $+$ *be understood of these, then – is to be interpreted of the contrary.*

In this quote, the capitalization is his. With this understanding of negatives, how does he justify the rule for multiplying negatives? Here is what he says:

> *For the true notion of Multiplication is this, to put the Multiplicand, or thing Multiplied (whatever it be) so often as are the Units in the Multiplier. . . . and this, whatever the thing Multiplied, Positive or Negative: for there may well be a Double Deficit as a Double Magnitude; and* $-2A$ *is as much the Double of –A as* $+2A$ *is the Double of A. . . .*
>
> *But in case the Multiplier be a Deficit or Negative quantity; suppose* -1*; then instead of Putting the Multiplicand so many times, it will signify so many times to Take away the Multiplicand. . . . so that* $+$ *by – makes –; But to Multiply –A by* -2 *is twice to take away a Defect or Negative. Now to take away a Defect is the same as to supply it; and twice to take away the Defect of A is the same as twice to add A or to put 2A . . . : So that – by – (as well as* $+$ *by* $+$*) makes* $+$*.*

As far as I know, this is the first place in Western literature in which the rule of signs is not merely stated but explained so clearly. After this, when he gets to writing out the formulae for roots of equations, he no longer has to separate all these cases which we saw in Al-Khwarizmi and Cardano. For the quadratic he writes:

$$\begin{aligned}\text{Given the equation, } x^2 \pm 2bx &= \pm c^2\\ \text{the roots are } x \pm b &= \sqrt{\pm c^2 + b^2}\end{aligned}$$

(I have only changed his variable from a to x and noted squares by using e.g. c^2 for his cc.) Note that he follows Euclid is making all terms homogenous – so that, for example, x, b, c can all be lengths and the equation relates an area to an area. For this reason, he needs the symbol $\pm$ in front of the c^2.

Finally, Wallis gives what I believe is the first explicit use of the full number line, positives to the right, negatives to the left, in Western literature:

> *Yet is it not that Supposition (of Negative Quantities) either Unuseful or Absurd when rightly understood. And though, as to the bare Algebraick Notation, it import a Quantity less than nothing: Yet, when it comes to a Physical Application, it denotes as Real a Quantity as if the Sign were +; but to be interpreted in a contrary sense.*
>
> *As for instance: Supposing a man to have advanced or moved forward (from A to B) 5 yards; and then to retreat (from B to C) 2 yards; If it be asked, how much had he Advanced (upon the whole march) when at C? I find . . . he has Advanced 3 Yards. But if, having Advanced 5 Yards to B, he thence retreat 8 Yards to D; and it then be asked, How much is he Advanced when at D, or how much Forwarder than when he was at A: I say –3 Yards. . . . That is to say, he is advanced 3 Yards less than nothing. . . . (Which) is but what we should say (in ordinary form of Speech), he is Retreated 3 Yards; or he wants 3 Yards of being so Forward as he was at A.*

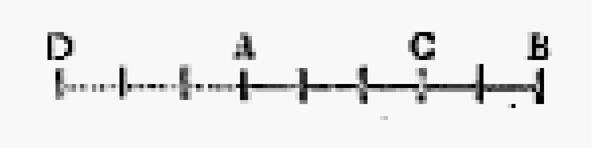


Figure 8. Wallis's illustration of the "number line"

Newton, as one would expect, had a full command of negative numbers and all their uses. He wrote lecture notes on arithmetic, algebra and geometry at some point, presumably early in his career. They were first published (without his approval) in 1707 and later translated into English with the title *Universal Arithmetick*. Here he introduces negative numbers at the very beginning with the following sentences[25]:

> *Quantities are either* Affirmative*, or greater than nothing; or* Negative*, or less than nothing. Thus in human affairs, possessions or stock may be called affirmative goods, and debts negative ones. And so in local motion, progression may be called affirmative motion, and regression negative motion; because the first augments, and the other diminishes the length of the way made. And after he same manner in geometry, if a line drawn in a certain way be reckoned for affirmative, then a line drawn the contrary way may be taken for negative.*

Later on, he discusses multiplication and is very clear that pure numbers arise as ratios of quantities with the same dimension and one can either multiply a quantity with a dimension by a pure number, getting another such quantity or multiply two

[25] Page 3 of the second edition published in 1728.

pure numbers. He states the rule for the sign of the product simply as "*. . . making the product* Affirmative *if both factors are Affirmative or both Negative; and* Negative *if otherwise.*" Unfortunately, he says nothing about why one should believe in this rule.

Whereas Fermat had given a systematic study of quadratic equations in two variables showing that they all defined conic sections and Descartes had introduced several cubic equations giving new curves (notably the "Cartesian parabola" and his "Folium"), Newton went on to look at all possible cubics, in an article entitled "Curves" in *Lexicon Technicum* by John Harris published in London in 1710. He classified them into 72 types and sketched them. *Without hesitation, he used all four quadrants of the plane* and plotted all roots (x, y), positive and negative. Here is an example:

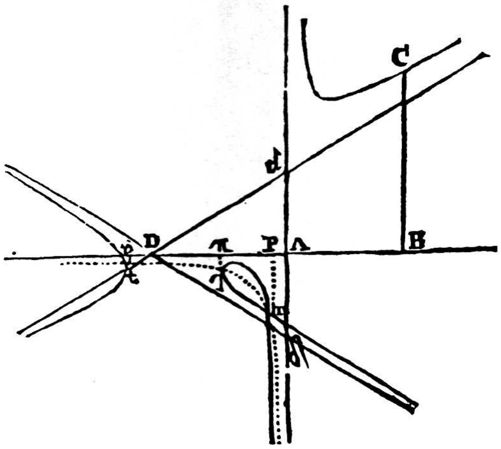


Figure 9. One of the 72 types of cubic curves plotted by Newton

After Wallis and Newton's work, a modern arithmetic with negative numbers was widely accepted in Continental Europe, where there was an explosion of mathematical research during the Enlightenment. In England, curiously, the resistance to negative numbers continued for some 150 years, culminating in De Morgan. A long debate ensued between those who accepted them and those that didn't, a story which is beautifully described in Pycior's book that we have cited. In the end, De Morgan and Hamilton founded the general theory of fields and negative reals took their place in the greater world of complex numbers and quaternions.

5. Two Factors in the World View of 15th–17th Century Europe

I hope I have proven my point that Europe in the 16th and 17th centuries resisted expanding their numbers to include negatives in a way which calls for some explanation. China and India both seem to have moved naturally to this bigger domain of numbers when the occasion presented itself. I want to make the case that the European reticence was due to two factors. The first was the overwhelming importance of Euclid in defining what is and what is not mathematics and the fact that negative numbers had no place in Euclid's view of mathematics. The second is that, at the time negative numbers should have been accepted, imaginary numbers cropped up too and the idea arose that both negative and imaginary numbers had the same twilight existence. It was because of negatives that square roots had a problem, so maybe it was best to consider them both as second class citizens of the world of numbers.

Euclid's *Elements* were written in the newly founded school/library at Alexandria around 300 BCE and integrated the mathematical ideas of Theaetetus, Eudoxus and many others in a systematic treatise. It is written in a monolithic theorem/proof style not seen again in the History of Mathematics until the collective 'Bourbaki' composed their treatise in the 20th century. It was translated into Arabic in the 8th century CE and from Arabic into Latin in 12th century. As a result, it came to define what mathematics is for every generation of Arabs and Europeans, arguably until Newton and the Enlightenment when concepts with no roots in the Elements began to take center stage.

But what is Euclidean mathematics? There are roughly three parts to the *Elements*: Books I–VI on plane figures, Books VII–X on number theory and irrationals and Books XI–XIII on three dimensional geometry. What numbers occur in the *Elements*? Here's a list:

1. "*magnitudes*": the length of a line, the area of a plane figure and the volume of a solid figure
2. positive integers implicitly as in "*The greater is a multiple of the less when it is measured by the less*" (definition 2, Book V) and explicitly as in "*A number is a multitude composed of units*" (definition 2, Book VII). Note that the number is still a length but, because he always has a "unit" around when studying numbers, it becomes in effect dimensionless.
3. ratios as in "*A ratio is a sort of relation in respect of size between two magnitudes of the same kind*" (definition 3, Book V).

Note that none of these concepts give numbers which can be negative or even zero. What sort of arithmetic does Euclid have for these numbers?

1. Magnitudes are clearly added and subtracted (so long as the result remains positive), lengths are multiplied to give areas and volumes, etc. But "units" are only introduced in Book VII and there are no actual calculations and certainly no approximations (e.g. for π).
2. Positive integers are also added and subtracted and multiplication is defined in "*A number is said to multiply a number when that which is multiplied is*

added to itself as many times as there are units in the other and thus some number is produced".

3. Adding and multiplying ratios is the main goal in the extremely abstract Book V, which is said to be the work of Eudoxus. Book V begins with defining when two ratios are equal. For any ratio given by two lines A and B, he considers which multiples satisfy $nA > mB$ and which satisfy $nA < mB$. Of course, this is the 'cut' Dedekind re-introduced in the 19th century to *construct* real numbers from rationals. Here Eudoxus doesn't need to define real numbers – they are ratios given by geometry. What he needs to do is to define equality of ratios and he does this by requiring that their associated cuts are the same. Addition and multiplication of ratios are both implicit in that (a) if a line segment A is divided into two parts B and C then $A : D$ is going to the sum of $B : D$ and $C : D$ and (b) $A : C$ is to going to be the product of $A : B$ and $B : C$. What is not at all clear is that addition and multiplication are *well-defined* operations on the equivalence classes called ratios. This is exactly what is asserted in Proposition 24, Book V (for addition) and in Proposition 22, Book V (for multiplication) after a long and subtle sequence of intermediate steps. One stands amazed at Eudoxus' mathematical skills.

How about algebra, identities and formulas with the arithmetic operations? Euclid studies at length in Book II what people call 'geometric algebra', a series of propositions which amount to algebraic identities such as

$$(a+x)^2 + (a-x)^2 = 2a^2 + 2x^2$$

which is essentially the content of Proposition 9, Book II. Now what about the solutions of quadratic equations? This seems to be essentially what the lengthy and confusing Book X is all about. As Heath points out in his introduction to Book X, Euclid's classification of *binomials* and *apotomes* can be read as a systematic study of all the *positive* roots of all possible quadratic equations. This sets the stage for the separation of cases in treating roots of polynomial equations in all the works we have reviewed.

All in all, if you are going to start with Euclid, you are not going to be predisposed to introduce negative numbers in to your calculations. He has gone to extraordinary lengths to reduce arithmetic and algebra to geometry and thoroughly inoculate it against negatives. It is worth looking briefly at what else was known at 300 BCE which Euclid did *not* put in his book. There is apparently an unbroken tradition starting in Babylon in 1800 BCE and continuing through Ptolemy of calculating with the sexagesimal equivalent of decimals and approximating e.g. $\sqrt{2}$ and π to many sexagesimal places. Moreover, there was also a tradition also going this far back of solving quadratic equations by algorithms – described in words but exactly equivalent to the quadratic formula. Euclid, in other words, distanced himself from a rich numerical tradition and consciously, it would seem, purified his version of mathematics.

The Europeans, then, had the benefit of this shining example of pure math and of the wonderful deductive logic on which they built. But it was hard to go beyond it in any radical way, to model other phenomena in the real world which cried out for negatives. Euclid was both the strength and the weakness of the European mathematical world of the 16th and 17th centuries.

But I think there is a second factor behind the slow acceptance of negatives which ought to be considered. As soon as one accepted -1, the algebra of the day thrust upon you formulae requiring its square root and this was truly inexplicable. The fate of -1 and i were inseparable. Cardano's book makes this very clear. We have already quoted from Chapter 37, near the end of his book, entitled *On the Rule for Postulating a Negative*. The Chapter starts with the sentence:

> *This rule is threefold, for one either assumes a negative, or seeks a negative square root, or seeks what is not.*

He is essentially equating three follies, all problematical. That he later entertained the idea that perhaps $(-1)^2$ ought to be equal to -1 shows how he viewed the problems as intertwined. Harriot (1560–1621) also played with both possibilities, as in the poem:

> *Yet lesse of lesse makes lesse or more,*
> *Use which is best keep both in store*
> . . . . . .

(Here '*lesse of lesse*' means multiplying -1 by -1 and he asks in line 1 whether this should equal -1 or $+1$).

Even if you didn't accept -1, the *casus irreducibilis* mentioned above, the case of cubic equations with three real roots, was a bone in the throat of algebraists. As long one of these roots was positive, you really ought to have a formula for this root. But the formula of del Ferro for solving cubics requires in this case that you take the square root of a negative number in an intermediate step. Of course, the imaginary parts of the resulting complex expressions will cancel at the end but not before. The full story of this problem is quite ironic. Viète in 1593[26] discovered that trisecting an angle was equivalent to solving the special cubic equation which belongs to the *casus irreducibilis*:

$$x^3 - 3x = b$$

[26] In his *Supplementum Geometriae*. A full treatment is in *Theoremata ad sectiones angulares* [Opera pp 287–304]

and he showed how to reduce the general *casus irreducibilis* to this special case. Thus he reduced a famous unsolved algebraic problem to a famous geometric one, unsolved in the sense that no ruler and compass construction was known (nor exists). At the same time, Bombelli proposed that Cardano's formula could make sense if you solved

$$\left(x + \sqrt{-y}\right)^3 = a + \sqrt{-b}$$

So trisecting an angle was related to taking complex cube roots – but no one put these together for a long time by finding the geometric meaning of complex numbers. Later we have Wallis, knowing the geometric meaning of negatives as the left half line, searching for a two dimensional geometric interpretation of imaginary numbers. There was a big clue on the table if anyone had linked Viète's trisection with taking cube roots of complex numbers. I believe it was Euler who finally worked out complex exponentials and made the link between these two. Oddly enough, even then Euler did not make explicit the geometric interpretation of complex numbers, leaving this to Wessel, Gauss and Argand.

Finally, there is also the issue of a psychological explanation for avoiding negative numbers. As Tversky and Kanneman have made popular, people are '*loss averse*', a loss of \$$x$ causes more pain than a gain of \$$x$ and they do not act rationally using mathematically correct expectations. The fear of loss is one of themes in Ionesco's bizarre play *The Lesson*, where a young woman comes for a tutoring lesson: she can add with proficiency but cannot subtract. The mathematician doesn't come off very rational either: he winds up killing her.

Mathematicians are attracted to Platonism, of believing that their discoveries are all insights into the eternal true world of mathematical facts. This example, the discovery of negative arithmetic and its incorporation into our numerical and algebraic toolkit, shows us that we must not be too literal. Yes, negative numbers were eventually accepted in the West as well as in China and India and all three cultures made the same math out of them. But there can be huge differences between cultures in the way mathematics unrolls. Euclid led the West down a certain path, dominated for many centuries by geometric figures and constructions. Other cultures were more practical and looked to solving concrete problems with approximate numbers. I think the discovery of calculus is another instance of this split: in India, studying the numerical table of sines led mathematicians to the idea of first and second differences and the fundamental theorem of calculus. But that is another story.