New Foundations for Physical Geometry

Tim Maudlin

(First Draft: Please do not cite)

Socrates: I do not insist that my argument is right in all other respects, but I would contend at all costs in both word and deed as far as I could that we will be better men, braver and less idle, if we believe that one must search for the things one does not know, rather than if we believe that it is not possible to find out what we do not know and that we must not look for it.

Plato, *Meno* 86c

Introduction

The thesis of this book is both simple and audacious.

It is so simple that the basic claims can be boiled down into two sentences. First: the most fundamental geometrical structure that organizes physical points into a space is the line. Second: what endows space-time with its geometry is time. The remainder of this volume does nothing but elucidate these two sentences. Everything flows from them in such a straightforward way that I am almost convinced that the reader could close the book forthwith and, with sufficient patience and diligence, reconstruct most of the content from these two propositions.

As for the audacity, acceptance of either of these propositions demands the rejection of widely held and deeply entrenched alternatives. Consider a collection of objects that we wish to regard as forming not merely a set (which it does automatically) but as forming a *space*. Organizing the set into a space requires something more than the set-theoretic structure. What, at the most primitive level, is this "something else"?

For over a century, the mathematical subject devoted to this question has been *topology*. In topological theory, the fundamental structure that organizes a set into a space- organizes it so that notions such as the continuity of a function and the boundary of a set can be defined- is the *open set*. One specifies the topology of a set by specifying which of its subsets are open sets. All of the topological characteristics of the space are then determined by the structure of the open sets.

Topology is often called "rubber sheet geometry" because it describes geometrical characteristics of a space that are preserved under "stretching" the space without tearing or pasting. It is not obvious what could be meant by "tearing" or "pasting" a space, but the salient point is that topology concerns geometrical structure independent of distances. Intuitively, stretching can change the distances

between points, but will not change, for example, whether one point is enclosed by another set of points. As we will say, topology concerns the *sub-metrical* structure of a space. Standard topology asserts that the opens sets confer this structure on the space.

I will offer an alternative mathematical tool, a different way of understanding sub-metrical structure. This requires constructing a competitor to standard topology, which I call the theory of *Linear Structures*. Simply put, specifying the Linear Structure of a space amounts to specifying which sets of points in the space are lines. In terms of the lines, notions such as the continuity of a function, the boundaries of a set, and the connectedness of a space are defined. These definitions sometimes render different verdicts than the standard topological definitions, and have a wider sphere of natural application My burden is to show that we do better, when considering the geometrical structure of a physical space, by thinking in terms of the Linear Structure of the space than in terms of its open sets. If I am right, then the standard mathematical tools used for analyzing physical geometry have, for over a hundred years, been the wrong tools.

So the first task to be tackled is a purely mathematical task: to present, from its foundations, a new method of analyzing sub-metrical structure. Anyone familiar with the tremendous scope and complexity of topology will appreciate the audacity mentioned above. Topology is the subject of hundreds of books and many thousands of papers in mathematics. Recovering or recasting the results of standard topological analysis in terms of Linear Structures would be the work of several lifetimes. So all that can be done here is to lay the foundations, to show how the most basic concepts defined in standard topology can be given alternative definitions in the theory of Linear Structures. The first part of this book is devoted to this task, and will not cover even as much territory as the most elementary introduction to standard topology. If I am able to convince the reader of the value of this new approach, it will not be by seeing farther than the standard theory, but by looking deeper. I will try to show that the definitions and analyses available in the theory of Linear Structures offer a better understanding of geometrical structure, and allow for definitions that more closely capture the intuitive notions we are

trying to explicate, than do the standard definitions. We *understand* geometrical structure better if we think in terms of lines rather than open sets.

Even if one comes to share this assessment, still the magnitude of the task I am suggesting may render the undertaking slightly absurd. It is rather like noticing that the Empire State Building would have been better situated had it been built a few blocks over and turned on an angle. One may agree with the appraisal, but still be reluctant to go to the trouble to reconstruct on better foundations. Maybe standard topological theory is not the best way to understand physical geometry, but it is still *good enough*. Thomas Kuhn observed: "As in manufacture, so in science–retooling is an extravagance to be reserved for the occasion that demands it".[1] Persuasive arguments that such an occasion has arisen are hard to come by, and the more extensive the retooling, the more persuasive they must be. Following common practice when confronting such problems, I will resort to both a carrot and a stick.

The stick consists in a critique of standard topology. Of course, the issue is not a *mathematical* one: standard topology is a perfectly well-defined mathematical subject with rigorous and wide-ranging results. Rather, the critique is conceptual. A formalized mathematical subject such as topology is devised in the first place to capture, in a clear and precise language, certain informal concepts already in use. It is only because we begin with some grasp of a subject like geometrical structure that we seek strict definitions in the first place. Those formalized definitions can do a better or worse job of capturing the informal concepts whose names they inherit. It may be tempting to think that this is a purely *semantic* debate, in the pejorative sense of that term: after all, if someone wants to *define* a word like "continuous" or "connected" or "boundary" using the resources of standard topology, who's to object? As long as the definition is given, one can regard the term as nothing but an abbreviation, a concise way to refer to the defined concept. Such an approach makes the whole project of criticizing formal definitions appear wrong-headed.

[1] Kuhn 1996, p. 76

But the situation is subtler than that. Certain mathematical terms are not chosen arbitrarily, but are used because we already have some understanding of them. Long before the formal theory of topology was developed, mathematicians had something in mind when they characterized a function as continuous or a space as connected. Their concepts may have been somewhat imprecise, but everyone would have accepted some clear instances of continuous and discontinuous functions. For example, the sine function is a continuous function and the step function is not. And beyond these particular examples, notions such as continuity would be explicated by informal definitions. So when the topologist seeks to define "continuity" in her proprietary technical language, she is not entirely free. The definition must be shown to correspond–to the extent that a formally defined notion can correspond to a more informal and fuzzy one– with the concept one began with. If it does not, then the formal theory has failed in its aim.

In the first chapter of the *Physics*, Aristotle characterized the method of science as starting from those things that are clearer and more knowable to us and proceeding to those things that are clearer and more knowable in themselves. The mathematical elucidation of geometrical structure must proceed in the same way: one starts with the familiar, though somewhat obscure, and proceeds to the clearly and exactly defined. The fundamental axioms and definitions are presented in a more rigorous technical vocabulary, and then the initial notions are defined– and illuminated– by means of the technical notions. One returns to the starting point with a deeper understanding. But if one of the tasks is to explicate those initial concepts, then one should carefully consider whether the technical definitions have done justice to the original concepts, at least where their application was clear and uncontroversial.

Different readers will likely have wildly divergent reactions to these criticisms of standard topology. In particular, readers already familiar with the standard definitions –especially mathematicians or physicists who commonly use the standard theory– will have so internalized the standard definitions that *those definitions express what they now mean by terms such as "continuous"*. These readers will have to make an effort to recall the original, somewhat amorphous, concepts

that stood in need of clarification. And given the utility of the formalized notion, such readers are likely to see no point in trying to capture some more naïve notion. On the other hand, readers with little background in standard topology have the double task of learning the standard definitions and evaluating criticisms of them at the same time. They may be more open to accepting the critique, but also less concerned about it in the first place. So I will not rest too much weight on these shortcomings of the standard theory. But I will point them out nonetheless.

Perhaps a more effective line of attack concerns the scope of application of the standard theory. Topology was initially developed as a tool for describing certain spaces, a central example being Euclidean space. In particular, the spaces most naturally suited for topological treatment are *continua* (we leave to later sections the discussion of exactly what this means!). But the single most important object of which we need a geometrical account is *physical space* (or *space-time*), and there is no guarantee that physical space is a continuum. Indeed, many physicists believe that at a fine enough scale physical space is discrete rather than continuous. If standard topology if not an effective tool for articulating the geometrical structure of discrete spaces, then it may not be well suited for the primary requirements of physics. It would, in any case, be preferable to have an account of geometrical structure that can be applied with equal ease to discrete and continuous spaces. The theory of Linear Structures can be.

Sticks, however, will never be enough to drive mathematicians and physicists out of the precincts of standard topology. Even if the standard approach is somehow flawed, they will reasonably demand a viable alternative. So the onus of persuasion must rest with the carrot: the theory of Linear Structures must be sufficiently intriguing in its own right to attract interest. I cannot claim unbiased judgment here, but I can attest that playing with the theory is a tremendous amount of fun. One is given a set of primitives (the lines), and then has to try to fashion reasonable definitions of other geometrical notions in terms of them. Often it is not obvious how to do this, and many alternative strategies present themselves. For example, once the set of lines in a space has been specified, how can one define what it means for a set of point to be open, or closed, or for one set of points to be the boundary of

another, or for a set to be connected, or for a function from one space to another to be continuous? There is no mechanical algorithm for producing such definitions, nor any indisputable standard by which a proposed definition can be evaluated. One wants the definitions to be natural, and to yield intuitively correct results, but one also wants the definitions to *lead to interesting theorems*. That is, the properties invoked in the definitions need to be exactly those properties from which other interesting results can be derived. But the fecundity of definitions is only established by the productions of proofs. A fascinating dialectic therefore develops: one proposes a definition, and then sees whether interesting proofs using the defined properties are forthcoming. If the proofs require slightly different properties, then the definitions can be adjusted.[2] Given the nature of the dialectic, one is always left uncertain whether better definitions are not possible: one needs the definitions to generate the proofs, but only gets a sense of how fecund the definitions are once the proofs are available. If the foregoing description seems too abstract, I can recommend only that the reader try for himself: once the basic axioms of a Linear Structure have been specified, try to construct definitions of terms like "open set" or "continuous function". I hope that especially mathematicians and physicists will give this a try, and see how easy it is to become hooked. For the feeling of productive conceptual play is the ultimate carrot that I have to offer.

The first feat of audacity, then, is to contend that the most well-entrenched approach to the formal analysis of geometrical structure should be forsaken, in some contexts, for completely new one. In the spirit of fair play, the second thesis should be as outrageous to physicists as the first is to mathematicians. For if one accepts the use of Linear Structures to articulate sub-metrical geometry, then the foremost *physical* question that confronts us is: what accounts for the Linear Structure of physical space-time? I claim that the geometry of space-time is produced by time.

[2] For a delightful discussion of this dialectic in the search for formal definitions of informal concepts, see Imre Lakatos's *Proof and Refutations* (19??). My own experience in trying to formulate definitions in terms of the Linear Structure corresponds exactly to Lakatos's description.

Why should such a claim be considered audacious? Because it reverses the common wisdom about the theory of Relativity. Relativity is often taken to imply that time is "just another dimension" like a spatial dimension, so the notion that there is anything physically special about time (as opposed to space) is outmoded classical thinking. Relativity is said to postulate a "four-dimensional block universe" which is "static", and in which the passage of time is just an illusion. Einstein himself wrote, near the end of his life, that "[f]or those of us who believe in physics, this separation between past, present and future is only an illusion, however tenacious".[3] In short, Relativity is commonly characterized as having *spatialized time*, that is, of having put the temporal dimension on an equal physical footing with the spatial dimensions, and of having thereby robbed time of any fundamental different from space.

My contention is just the opposite: the theory of Relativity shows, for the first time in the history of physics, how to *temporalize space*. In Relativity, but not in any preceding classical theory, one can regard time as the basic organizing structure of space-time. *In a precise sense, space-time has geometrical structure only because it has temporal structure, and insofar as there is spatial geometry at all, it is parasitic on temporal structure*. The argument to this conclusion is straightforward: the (sub-metrical) geometry of space-time is determined by its Linear Structure, and the Linear Structure of a Relativistic space-time is determined by its temporal structure. So rather than somehow demoting time from its position in classical physics, Relativity promotes time to a more central position. This thesis will be the topic of the second part of the book.

Having touted the outrageousness of the book's central claims, let me now calm the waters. Regarding the physical thesis, we should immediately note that the special geometrical role of time in structuring space-time is not, at a technical level, at all contentious. Physicists will not dispute that an intimate relation exists between lines and temporal notions in Relativity, while no such relation exists

[3] Letter to Besso's family- find exact citation.

between lines and spatial notions. The only real bone of contention here will be the significance of that fact.

With regard to the mathematical claim, let me reiterate that there is nothing wrong *per se* with standard topology as a tool of mathematical analysis. Many questions can be properly and insightfully addressed by topological analysis. The weakness of standard topology emerges chiefly when treating the specific subject of *geometrical space*. But what exactly do I mean by that term?

Metaphorical and Geometrical Spaces

In the right context, almost any collection of objects can be considered to form a "space". For example, if one is studying Newtonian mechanics, e.g. Newton's theory of gravity applied to points particles, it is natural to speak of "the space of solutions" of Newton's equations of motion. Each "point" in this space, each individual element, describes the motions of a set of particles governed by Newtonian gravity given a set of initial conditions. There is an intuitive sense– which can be made technically precise– in the which various solutions can be "closer" or "farther" from one another, and hence an intuitive sense in which the whole set of solutions can be thought of as having a "geometry". But this sort of talk of a "space" is evidently metaphorical. This "space" is, in an obvious sense, a *metaphorical* space, it is just a way of talking about the solutions and a measure of *similarity* among them. Analogously, philosophers are wont to speak of "logical space" as the set of all possible worlds. But this set also only forms a "space" in a metaphorical sense: space talk is just a picturesque means of discussing various ways and degrees that individual possible worlds are similar to one another.

In contrast, consider Euclidean space, the subject matter of Euclidean geometry. Euclidian space is an abstract object in the way that all mathematical objects are abstract. But Euclidian space is not just metaphorically a space. When we say that one point in Euclidian space is "closer" to another than it is to a third, we are not suggesting that the first point is more similar to the second than to the third in any way. Indeed, intrinsically the points of Euclidian space are all exactly alike,

they are all, in themselves, perfectly identical. The points of Euclidian space, unlike the "points" of the space of solutions to Newton's equations, really are points: they have no internal structure. The "points" of the space of solutions only form a (metaphorical) "space" because they are highly structured, and different from one another. The points of Euclidean space, being all intrinsically identical, only form a space because of structure that is *not* determined by their intrinsic features. Euclidian space is therefore an instance of what I mean by a *geometrical* space.

Euclidean space is the most important historical example of a geometrical space, but mathematicians have studied many other geometrical spaces. The various non-Euclidean spaces studied in Riemannian geometry are geometrical spaces in my sense, as is Minkowski space-time and the space-times that are solutions to the equations of General Relativity. In contrast, the set of integers, or of real numbers, do not form a geometrical space: the elements are not all intrinsically alike and the "geometrical" notion of, e.g., proximity is determined by the different intrinsic natures. The sense in which the number 2 is "closer" to the number 1 than it is to the number 100, or sense in which the number 2 "lies between" 1 and 100, has to do with the arithmetic nature of these objects, not with any extrinsic structure that unites them.

Attempting to give explicit definitions is a dangerous business. So I implore the reader first to reflect on the particular examples I have just given to understand how I mean to use "geometrical" and "metaphorical" when characterizing spaces. There is an evident difference between Euclidean space and the "space" of solutions to Newton's equations, and it is this difference I mean to mark. For the purposes of this book, just these examples should make the distinction clear enough.

Why is the characteristic mentioned above not necessary and sufficient: a geometrical space is a space in which the points are all intrinsically alike? This fits the mathematical spaces commonly studied under the rubric "geometry". But in addition to these mathematical spaces, we also want it to turn out that physical space– the space (or more properly space-time) that we actually inhabit, the space we walk around in– counts as a geometrical space. But the points of that space are not all intrinsically alike. For example, some points may be occupied by matter, or

by certain fields. Even so, those differences of material content do not analytically determine the geometry of the space we live in.[4] Simply put, all the facts about the intrinsic structure of two points in physical space-time do not determine their geometrical relation to one another. In the argot of philosophy, the geometrical structure of real space-time does not supervene on the intrinsic features of the points of space-time. So physical space-time, the space-time we live in, is a geometrical space.

(That is, of course, assuming that physical space-time contains points at all. All of our discussion so far, and in the remainder of the book, presupposes that physical space is a point set. This assumption may be challenged, and both mathematicians and philosophers have discussed the possibility that space that contain no points but only finite regions.[5] Still, all of the constructions of standard topology presume point sets, and all of the axioms of the theory of Linear Structures will presume point sets. Whether those axioms could be modified in a natural way to treat of pointless spaces is a question best left for another time.)

The distinction between metaphorical spaces and geometrical spaces that I have tried to draw may seem very unnatural to mathematicians. Mathematical practice has systematically ignored, and positively disguised, that distinction for some time. As an example, both mathematicians and physicists commonly refer to three-dimensional Euclidean space as R^3, and to R^3 as three-dimensional Euclidean space, as if these were just different names for the same mathematical object. But according to my usage, three-dimensional Euclidean space is a geometrical space and R^3 (the set of ordered triples of real numbers) is a metaphorical space. It is true that the latter can be used to *represent* the former, but the two cannot be identified. For several millennia, the actual space we live in was believed to be a three-

[4] Why the qualification "analytically"? According to General Relativity, the geometrical structure of space-time is physically affected by the distribution of matter, in accordance with Einstein's field equation. But still, the space-time has a geometrical structure, which can be mathematically specified independently of the matter fields: it is just the geometrical structure that stands on one side of the equation. So as far as the geometry is concerned (as opposed to the physical laws that produce the geometry), the points of space-time are all the same.

[5] See Caratheordory 1963 and Skyrms 1993.

dimensional Euclidean space, but no one ever imagined that the actual space we live in consisted in ordered triples of real numbers. Such a proposal makes no sense whatever.

Similarly, the Euclidean plane– the object that Euclid himself studied and proved theorems about– is not R^2 (the set of ordered pairs of real numbers). R^2 has a tremendous amount of structure that the Euclidean plane lacks. R^2 has an origin, the unique "point" which is the ordered pair <0,0>. Given a "point" in R^2, there is a fact about whether one or both of the real numbers in it are negative, or irrational, or whether the first number is an integral multiple of the second. Given any two "points" in R^2 (i.e. any two ordered pairs of real numbers), there are many well-defined mathematical facts about them. They either do or do not have the same real number in the first slot, or in the second slot. Visualizing the elements of R^2 as points in a Euclidean plane, this provides for a natural notion of an "*x*-direction" and a "*y*-direction". Furthermore, there are well-defined arithmetical operations on the elements of R^2. One can "add" or "subtract" or "multiply" the elements of R^2 by adding or subtracting or multiplying the respective real numbers in the "points".

In contrast, the points of the Euclidean plane have none of this structure. The Euclidean plane is homogeneous and isotropic. It has no "origin", nor any "*x*-direction" or "*y*-direction". And it makes no sense to "add" or "subtract" or "multiply" points in the Euclidean plane. In fact, it would be hard to imagine two objects as different as R^2 and the Euclidean plane! Through most of the history of mathematics, mathematicians would never have thought of even comparing a fundamentally arithmetic (and set-theoretic) object such as R^2 with a fundamentally geometrical object such as the Euclidean plane. So the confusion between the two, which is now so prevalent, is of relatively recent historical vintage.

It is of utmost importance that we make a forceful separation between the arithmetical and geometrical objects here. Our object of analysis is physical space (or space-time), and our conjecture is that physical space is a geometrical space. *If this is correct, then the most appropriate mathematical object to use to represent physical space is a (mathematical) geometrical space, not a metaphorical space*. For if physical space is a geometrical space, it is best understood by means of abstract

geometrical spaces. Even if it is possible to use a metaphorical space such as the "space" of ordered triples of real numbers to represent a geometrical space, this imposes an inappropriate screen of mathematical representation between us and the object we are interested in. Physical space is not made of numbers of any kind, nor of elements for which arithmetical operations are defined. So the use of numbers to represent physical space is not a natural one, and will entail careful consideration of which mathematical aspects of the representation correspond to physical aspects of the object represented. It may seem simple and obvious to note that R^2 has a unique origin while physical (and abstract Euclidean space) do not, but if one constantly employs arithmetical objects and metaphorical spaces when trying to represent non-arithmetical geometrical objects, the opportunities for confusion are endless.

The impulse to use arithmetical objects to represent geometrical structure has had an extremely rocky history, going back to the origin of formal mathematics. A brief (and selective, and highly partial) review of that history will help illuminate the present situation, and clarify the principles from which these new foundations for geometry have been constructed.

A Light Dance on the Dust of the Ages

The ancient Greeks divided the field of mathematics into two main divisions: arithmetic and geometry. Arithmetic was the theory of numbers and geometry the theory of space and is parts. By "numbers" (*arithmoi*), the Greeks meant only the positive integers, what we call the counting numbers. Numbers could be added and multiplied, and a smaller number could be subtracted from a larger. Zero was not a number, nor was there any notion of a negative number. One number could not always be divided by another, since there were no "fractional" numbers. The arithmetic unit, the one, was considered completely indivisible and partless. As Socrates says of arithmetic:

> It leads the soul forcibly upward, and compels it to discuss the numbers themselves, never permitting anyone to propose for

> discussion numbers attached to visible or tangible bodies. You know what those who are clever in these matters are like. If, in the course of the argument, someone tries to divide the one itself, they laugh and won't permit it. If you divide it, they multiply it, taking care that one thing never be found to be many parts rather than one.[6]

The subject matter of geometry included points, (straight) lines, curves, plane figures, solids, etc. Lines, plane figures and solids were instances of *magnitudes*. In contrast to numbers, magnitudes were taken to be infinitely divisible: a magnitude could always be divided in half, or into any number of equal or unequal parts. The prime examples of magnitudes were straight lines. Lines could be added to one another, and a shorter subtracted from a longer, but the only operation analogous to multiplication does not yield another line: it yields the rectangle that has the two lines as sides. Since space is three-dimensional, three lines can be "multiplied" to form a solid (a rectangular prism), but no similar construction would correspond to "multiplying" four lines. This contrasts with the multiplication of numbers, since the product of two numbers is another number of exactly the same kind.

Magnitudes and numbers, then, had rather little in common for the Greeks. With respect to divisibility, one might even contend that they were fundamentally opposed in their natures. Since numbers and magnitudes could both be added, and the smaller subtracted from the larger, certain principles applied to both fields. That is why the axioms of Euclid's *Elements* include propositions such as "equals added to equals are equals": the axioms were principles that governed both geometry and arithmetic, while the postulates were properly geometrical.

A more interesting commonality between arithmetic and geometry is provided by Eudoxus' theory of proportion. Numbers stand in ratios to one another, and magnitudes of the same kind (such as straight lines) stand in ratios to one another, and *a pair of numbers can stand in exactly the same ratio to one another as a pair of lines do*. The theory of proportions is presented in Book V of the *Elements*.

[6] *Republic* 525 e, translation by G. M. A. Grube, revised by C. D. C. Reeve (Plato 1992).

Since both numbers and magnitudes can stand in ratios, we begin to see how one might naturally use numbers to represent magnitudes (or magnitudes to represent numbers). We might, for example, be able associate numbers to lines in such a way that the lines stand in exactly the same ratio to one another as their associated numbers do. If there is one way to do this, there are many (for example, doubling all the numbers leaves their ratios unchanged, so the doubled numbers would do as well as the originals), which gives rise to what will much later be called a gauge freedom.

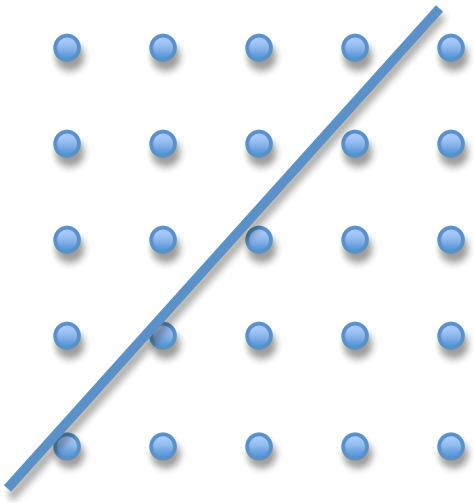
But for the Greeks, numbers would be of only very limited utility as representatives of magnitudes. For, as the Pythagoreans had discovered, magnitudes can stand in ratios that no pair of numbers stand to one another. The famous example is the ratio of the diagonal of a square to one of its sides: no two integers display exactly this proportion. Such pairs of magnitudes were called "incommensurable", since no number of copies of the one, laid end to end, would exactly equal any number of copies of the other.

It is often reported that the Pythagoreans discovered "irrational numbers", or that $\sqrt{2}$ is irrational, but this is an anachronism. They never recognized what we call rational numbers, much less irrational numbers, and would not have understood "irrational" as an adjective applicable to any individual mathematical object. A magnitude, e.g. the diagonal of a square, is neither "rational" nor "irrational" in itself: it is either commensurable or incommensurable *with another magnitude*. The fact that the side and diagonal of a square are incommensurable cannot be attributed to anything peculiar about either the side or the diagonal taken individually. Both the side and the diagonal are commensurable with some other magnitudes and incommensurable with others.

What the discovery of incommensurable magnitudes showed was that *the structure of ratios among magnitudes is intrinsically richer than the structure of ratios among numbers*. That is, the field of geometry presents an inherently more extensive mathematical universe than does the field of arithmetic, as the Greeks understood it. Perhaps this realization lies behind the legend that the Pythagoreans, being on a

ship at sea when one of their circle first proved the existence of incommensurable magnitudes, threw the hapless discoverer overboard. The Pythagoreans famously wished to reduce the fundamental essence of all things to number, but geometric investigation demonstrated this to be impossible.

On the other hand, the geometrical universe does contain natural representatives of numbers, viz., collections of points. The Greeks commonly visualized numbers as points arranged in spatial patterns. "Square" numbers could be arranged to form a square, "triangular" numbers arranged to form a triangle, and similarly for "pentagonal" and "hexagonal" numbers. So while we think of "squaring" as having to do with the multiplicative structure of numbers (and so can easily extend the notion to any "power" of a number), the Greeks thought of "squaring" and "triangling" in essentially geometrical terms. We no longer have any use for notion of triangular numbers, but the Greek visual approach allowed one to see relations between sorts of numbers without the use of algebra. For example, it is immediately obvious that every square number greater than 1 is the sum of two consecutive triangular numbers, for example, 25 = 10 + 15:



Using points to represent arithmetical units would strongly reinforce the idea that the unit is indivisible, and so make the invention (discovery?) of fractional numbers all the more difficult.

Although geometry provides natural representations of numbers, collections of points are not numbers: geometrical points have the property of location, which numbers lack. But if the Greeks had had a mind to try to reduce mathematics to one field instead of two, their only choice would have been to try to reduce arithmetic to geometry, rather than geometry to arithmetic. In addition, the body of geometrical results available to the ancients was much more impressive than the body at arithmetical results: it is not by accident that the most influential mathematical text in history was Euclid's *Elements*, and not the *Arithmetica* of Diophantus.[7] So it is hardly surprising that for nearly two millennia geometry took pride of place in mathematics. And it would have been obvious to any mathematician that a geometrical problem could not be stated or solved in the language of numbers, since the geometrical universe had more structure than the numerical universe.

If one desired to translate geometrical problems into the language of numbers, *one would have to invent (or discover) more numbers*. Progress in this direction (at least in the West) was quite slow and deliberate, with much discussion of the "reality" of any numbers beside the positive integers. This progress is the topic of the next chapter of our tale.

The Proliferation of Numbers[8]

As we have seen, Plato explicitly denied the existence of fractional numbers: the numerical unit had no parts and could not be divided. Of course, for practical purposes (in business and construction and astronomy, for example) fractions were commonly required. The use of what we call rational numbers therefore infiltrated almost imperceptibly into theoretical mathematics. It would be hard to say exactly when rational

---

[7] An autobiographical note: when I first heard of "Diophantine equations", they were described as equations for which only positive integer solutions were accepted. I was, at the time, extremely puzzled: if one seeks the solution of an equation, why care whether the solution happens to be an integral or not? But from Diophantus' perspective, he was simply seeking a numerical solution, without any restriction at all. If the equation had no integer solution, it had no solution.

[8] For more a more complete account of the history discussed here, see Kline 1972, from which this is largely drawn.

numbers were recognized as numbers since this requires making a careful distinction between the *ratio* 1:2 (which had a perfectly good pedigree in Eudoxus' theory of proportion) and the *number* ½. As Morris Kline reports of the Hellenistic period:

> [T]he Alexandrians used fractions as numbers in their own right, whereas mathematicians of the classical period spoke only of ratios of integers, not parts of a whole, and the ratios were used only in proportions. However, even in the classical period genuine fractions, that is, fractions as entities in their own right, were used in commerce. In the Alexandrian period, Archimedes, Heron, Diophantus, and others used fractions freely and performed operations with them. Though, as far as the records show, they did not discuss the concept of fractions, apparently these were intuitively sufficiently clear to be accepted and used.[9]

Kline's comment that fractions were "intuitively sufficiently clear" could be made more precise. What is characteristic of numbers is that they are subject to the arithmetical operations of addition, subtraction, multiplication and division. These operations on rational numbers can be reduced to the addition, subtraction, and multiplication of integers: A/B + C/D = (AD + CB)/BD, A/B – C/D = (AD - CB)/BD, A/B x C/D = AC/BD, A/B ÷ C/D = AD/CB.[10] Since irrational numbers cannot be represented by fractions compose of integers, arithmetical operations on irrational numbers could not in any obvious way be reduced to operations on integers. It would be quite a long time after this period before irrational numbers were tolerated, and until this step is taken, there are no prospects for describing geometrical problems in arithmetical terms.

By 1500, borrowing from the algebraic traditions of the Hindus and Arabs, some mathematical work appeared to trade in irrational numbers. Algebraic manipulation would yield symbols that at least seemed to denote various roots of integers that, as we would say, are irrational numbers. But the more careful writers were acutely aware of the gap between notation and mathematical reality. We can write down "0/0", which looks

[9] Kline 1972, p. 134

[10] The observation that arithmetical operations on the rational numbers can be defined in terms of arithmetical operations on the integers is made by Dedekind in *Continuity and Irrational Numbers* section III (1963, p. 10). As we will see, this was of utmost importance to Dedekind.

like a symbol for a rational number. Nonetheless, we would deny that there is any such number. Mathematicians adopted the same attitude toward algebraic calculations that we would interpret as yielding irrational numbers. Michael Stifel, for example, used symbols for roots quite freely, but when it came to recognizing irrational numbers he balked:

> Since, in proving geometrical figures, when rational numbers fail us irrational numbers take their place and prove exactly those things which rational numbers could not prove…. we are moved and compelled to assert that they truly are numbers, compelled, that is, by the results which follow from their use–results which we perceive to be real, certain, and constant. On the other hand, other considerations compel us to deny that irrational numbers are numbers at all. To wit, when we seek to subject them to numeration.… we find that they flee away perpetually, so that not one of them can be apprehended precisely in itself…. Now that cannot be called a true number which is of such a nature that it lacks precision…. Therefore, just as an infinite number is not a number, so an irrational number is not a true number, but lies hidden in a kind of cloud of infinity.[11]

Kline continues:

> A century later, Pascal and Barrow said that a number such as √3 can be understood only as a geometric magnitude; irrational numbers are mere symbols that have no existence independent of continuous geometrical magnitude, and the logic of operations with irrationals must be justified by the Eudoxian theory of magnitudes. This was also the view of Newton in his *Arithmetica Universalis*.[12]

We will see presently how the issue of irrational numbers was ultimately settled in the centuries that followed.

The debate concerning irrational numbers was paralleled by a less surprising debate about negative numbers. Just as algebraic manipulation of symbols can yield apparently irrational solutions to equations, so can it yield apparently negative solutions

---

[11] From Stifel's *Arithmetica Intega* (1544), cited in Kline 1972, p. 251.
[12] Kline 1972, p. 252

(and even apparently imaginary solutions). But should such "solutions" be taken seriously? Blaise Pascal thought the notion of subtracting 4 from 0 to be absurd on its face. Antoine Arnauld found an objection to negative numbers in the theory of proportions: If –1 exists, then the proportion 1: –1 is the same as the proportion –1 : 1 (as we might say, 1/–1 = –1/1). But if –1 is supposed to be less than 1, how can the proportion of the greater to the lesser be the same as the proportion of the lesser to the greater?[13] Euler would later argue that negative numbers were greater than (positive) infinity, illustrating how problematic the notion could be.

Let's consider how algebra can appear to lead one directly to irrational, and negative, and even imaginary numbers. The general solution of the quadratic equation $Ax^2 + Bx + C = 0$ is $x = (-B \pm \sqrt{(B^2 - 4AC)})/2A$, and by judicious choice of the numbers A, B and C, the "solution" can be made to be irrational or negative or imaginary. Why not take these solutions at face value? Methodologically, the problem is that algebra is a form of *analysis* (in the classical sense) rather than *synthesis*. In analysis, one begins by supposing there is a solution of the problem (which is represented by a symbol such as *x*, the unknown), and one then operates with this supposed solution as if it were a number: adding it, subtracting it, multiplying by it, and so on. The Greeks recognized that such a procedure, while useful as a heuristic, had no logical foundation as a method of proof. For the original supposition, viz. that a solution exists at all, might be false and hence all of the supposed arithmetic operations with the unknown chimerical. Classical Greek practice only accepted as rigorous *synthetic* proofs, in which the solution is constructed via accepted operations from the data.

In the case of the quadratic equation, there is available a clear geometric interpretation of what is being sought: one is asking for the points where a given parabola intersects a given line. Choice of the numbers A, B, and C are a means of specifying both the parabola and the line. And in those cases where the solution is imaginary (i.e. when $4AC > B^2$), the parabola fails to intersect the line at all. So the

[13] Kline 1972, p. 252

analytical presupposition, viz. that a solution exists, fails; it is hardly surprising that the formal algebraic manipulation yields nonsense.

Descartes and Coordinate Geometry

It is tempting to think that the amalgamation of arithmetic and geometry (and the acceptance of irrational numbers) must have been accomplished by Descartes. For as we understand it, coordinate geometry essentially involves naming geometrical points by means of ordered sets of numbers, and for there to be enough numbers to go around, we need irrational as well as rational numbers. As it turns out, this claim is historically inaccurate: Descartes, despite our use of the phrase ”Cartesean coordinates”, did not invent modern coordinate geometry. Descartes’ accomplishment, instead, was the importation of *algebraic method* into geometry. Since algebra had been developed for the solution of arithmetical problems, this importation required a means by which geometrical magnitudes could be handled as if they were numbers. Descartes explains precisely how this is to be done in the first paragraph of *La Geometrie*:

> Any problem in geometry can easily be reduced to such terms that the knowledge of the lengths of certain straight lines is sufficient for its construction. Just as arithmetic consists of only four or five operations, namely, addition, subtraction, multiplication, division and the extraction of roots, which may considered a kind of division, so in geometry, to find required lines it is merely necessary to add or subtract other lines; or else, taking one line which I shall call unity in order to relate it as closely as possible to numbers, and which can in general be chosen arbitrarily, and having given two other lines, to find a fourth line which shall be to one of the given lines as the other is to unity (which is the same as multiplication); or, again, to find a fourth line which is to one of the given lines as unity is to the other (which is equivalent to division); or, finally, to find one, two, or several mean proportionals between unity and some other line (which is the same

> as extracting the square root, cube root, etc., of the given line). And I shall not hesitate to introduce these arithmetical terms into geometry, for the sake of greater clearness.[14]

We have noted that it is characteristic of numbers that they can be multiplied and (at least sometimes) divided to yield other numbers, but no such operation exists for geometrical magnitudes. The closest one can come to multiplying two lines is to form a rectangle with the lines as sides, but this yields a plane figure rather than another line. Since multiplication and division are used in algebra, Descartes needed a way to make sense of these operations when applied to magnitudes rather than numbers. His solution employs the theory of proportions and, critically, the introduction of an arbitrary magnitude to play the role of unity.

Here's the idea, presented very informally. Let the arbitrary unit length be denominated U. Given other two lengths A and B, we seek a fourth length C that can be regarded as the product of A and B *relative to the choice of U as unity*. Since all of these lines stand in ratios to one another, we can straightforwardly ask for a line that stands in the same ratio to A as B does to U, i.e. C : A :: B : U. In modern notation, we are tempted to write C/A = B/U, and by "cross multiplying", CU = AB. If we regard U as unity (that is, as the "multiplicative" identity), then CU = C, and we get C = AB. (A clearer geometrical explanation: construct a rectangle with sides A and B. Now construct a second rectangle with the same area, one of whose sides is U. The length of the remaining side is the sought-for length C.) Note that the operations of "multiplication" and "division" of lines only have content relative to the arbitrarily chosen standard of "unity", so these are not the same sort of operations as their arithmetical namesakes. But the conventions allow Descartes to use the language and methods of algebra to describe geometrical problems. Descartes is then careful to also describe how to translate the algebraically described solutions back into a geometrical construction, so the desired line can ultimately be constructed.

Where, in this new method, is modern coordinate geometry? Bluntly put, nowhere. Descartes never assigns coordinate numbers to points: he rather uses

[14] Descartes 1952, p. 295

algebraic symbols to denote geometrical magnitudes and explains what geometrical meaning is to be given to apparently arithmetic operations. When Descartes identifies a point on the plane by means of two quantities x and y (as he does, e.g., when solving a problem that goes back to Pappus and Apollonius and Euclid[15]), the x and y do not denote numbers, they denote lines. Modern readers are apt to be confused since modern notation uses these same variables to range over coordinate numbers, but this was not any part of Descartes' method. If Descartes were to derive a solution of the form $x = \sqrt{2}$, he would not conclude that x is an irrational number, but that x is a line of such that the square built on it has twice the area of the square built on the arbitrarily chosen line he called "unity". It would then be easy to construct such a line.

Newton was critical of the Cartesean methodology, and his criticism clearly indicates the nature of Descartes' method. In *Arithmetica Universalis* Newton wrote:

> Equations are expressions of arithmetical computation and properly have no place in geometry except insofar as truly geometrical quantities (that is, lines, surfaces, solids and proportions) are thereby shown equal, some to others. Multiplications, divisions and computations of that kind have been recently introduced into geometry, unadvisedly and against the first principles of this science….Therefore these two sciences ought not to be confounded, and recent generations by confounding them have lost that simplicity in which all geometrical elegance consists.[16]

Neither Descartes nor Newton would have recognized the existence of irrational numbers or negative numbers, and hence neither would had had the resources to employ coordinate geometry as we understand it. The judicious application of algebraic method to geometrical problems does not require any expansion of the numerical universe, much less the whole universe of real numbers. So for Descartes and Newton, arithmetic and geometry remained fundamentally different fields of

[15] Descartes 1952, p. 301
[16] Newton 1707, p. 282, cited by Kline 1972, p. 318

mathematics: the modern amalgamation of arithmetic and geometry had not yet occurred.

John Wallis and the Number Line

The simplest instance of modern coordinate geometry–geometry that associates numbers with geometrical points–is the numerical coordinatization of a one-dimensional geometrical space: a line. If there are enough numbers to coordinatize a line, then ordered sets of numbers can be used to coordinatize the plane, or Euclidean three-dimensional space. Conversely, if one has not recognized enough numbers to coordinatize a line, then there will be no prospect of converting geometrical problems into properly arithmetical ones. Furthermore, if there really are enough numbers to replicate the structure of a line, then there must be irrational numbers such as √2, and negative numbers as well. But the idea that there are enough numbers to coordinatize a line is just the idea that there is a *number line*. That is, if one can associate numbers with a geometrical magnitudes in the way the number line suggests, then there are enough numbers to replicate the ratio-structure of the magnitudes. And that means that there must be more than just the rational numbers.

I conjecture that the extensive use of the number line makes it difficult for us to recover the historical puzzlement concerning irrational and negative numbers. Nowadays, children are introduced to the number line as early as they are introduced to arithmetic at all. And with the number line before you, you can just see that there must be numbers such as ½, and √2, and even –3. If there are as many numbers as there are points on the line, then the results of geometry can be used to prove things about the existence of numbers.

Of course, from a logical point of view all of this is perfectly question-begging. *If* there are numbers enough so that they can be put into a natural correspondence[17]

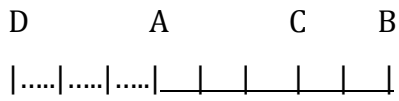
[17] By "natural correspondence" I mean that there is an isomorphism between the relation of greater and lesser defined on the numbers and the linear order of points on the line.

with points on a geometrical line, then there must be more than just the rational numbers. And if you don't believe that there are irrational and negative numbers, you won't believe that there *is* a number line. But the use of the number line as a pedagogical tool would seem to be an indication that irrational and negative numbers have been accepted into the arithmetical universe. Which raises the question: when was the number line invented?

Standard references attribute the concept of the number line to John Wallis in his *Treatise on Algebra* (1685). If this attribution were correct, then there would be evidence of the acceptance of irrational and negative numbers from that date, and therefore acceptance of enough numbers to be used as numerical coordinates of geometrical points. Unfortunately, a careful examination of Wallis work reveals that far from accepting more than the traditional collection of numbers, he was concerned instead to explain how one can invest phrases with geometrical meaning even when they have no arithmetical content.

Following the observation that algebraic arguments can apparently lead to both negative and even imaginary roots, Wallis remarks:

> But it is also impossible, that any Quantity (though not a Supposed Square) can be *Negative*. Since that it is not possible that any *Magnitude* can be *Less than Nothing*, or any *Number Fewer than None*.
>
> Yet is not that Supposition (of Negative Quantities,) either Unuseful or Absurd; when rightly understood. And though, as to the bare Algebraick Notation, it imports 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 he had Advanced (upon the whole march) when at C? or how many Yards he is now forwarder than when he was at A? I find (by subducting 2 from 5,) that he is Advanced 3 Yards. (Because + 5 – 2 = + 3.)



But if, having Advanced 5 yards to B, he then retreat 8 Yards to D; and it be asked, How much is he Advanced when at D, or how much Forwarder than when he was at A: I say –3 Yards. (Because + 5 – 8 = – 3.) That is to say, he is advanced 3 Yards less than nothing.

Which in propriety of Speech, cannot be, (since there cannot be less than nothing.) And therefore as to the line AB *Forward*, the case is Impossible.

But if (contrary to the Supposition,) the Line from A, be continued *Backward*, we shall find D 3 Yards *behind* A. (Which was presumed to be *Before* it.)

And thus to say, he is *Advanced* – 3 Yards; 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.

Which doth not only answer Negatively to the Question asked. That he is not (as was supposed,) Advanced at all: But tells moreover, he is so far from being Advanced (as was supposed) that he is Retreated 3 Yards; or that he is at D, more backwards by 3 Yards, than he was at A.

And consequently – 3, doth as truly design the point D; as + 3 designed the point C. Not Forward, as was supposed; but Backward, from A.

So that + 3, signifies 3 Yards Forward; and – 3, signifies 3 Yards Backward: But still in the same Straight Line. And each designs (at least in the same Infinite Line,) one Single Point: And but one. And thus it is in all Lateral Equations; as having but one Single Root.[18]

[18] Wallis 1685, p. 265

Wallis goes on to apply this same train of reasoning to the imaginary roots of equations. Just as the production of (apparently) negative roots shows that a presupposition of the problem was wrong (e.g. that the man ultimately advanced rather than retreated), so the production of imaginary roots shows that a presupposition was wrong (i.e. that the point sought lies on a particular line rather than somewhere else in the plane). And just as the negative root, properly interpreted, can indicate where the point sought actually lies, so can the imaginary root. In essence, Wallis goes on to invent something like the complex plane.

But in all of this explication, Wallis does *not* accept the reality of negative numbers or imaginary numbers. He says explicitly at the outset that no number can be less than zero. Rather, all that Wallis argues for, we might say, is the utility of negative numerals. Algebraic manipulation can produce a result such as "–3". That sign does not denote any number at all: an indication that the problem was posed based on a false presupposition. But in the problem discussed, the solution to be sought was never a number in the first place: it was a point in space, viz. the location of the man after his perambulation. And under appropriate interpretation, the sign "–3" can be understood to denote exactly that point. In exactly the same way, Wallis could remain unperturbed if algebraic arguments yielded the result "√2": he could explain how that symbol could denote a particular point in the line.

In sum, if we require that the number line be composed of numbers, Wallis did not invent it. Since he did not accept negative numbers, he could not have imagined that any collection of purely numerical objects could properly represent the geometrical line.

A similar attitude toward negative numbers persisted for the following several centuries. George Peacock, for example, in his *Arithmetical Algebra* (1842) makes a distinction between "arithmetical" and "symbolic" algebra. In arithmetical algebra, letters always stand for numbers. So in arithmetical algebra, the symbol "$a - b$" does not always represent something: it fails to represent anything if $b > a$, since then b cannot be subtracted from a. In such a case, writes Peacock, "we might call the quantity represented by $a - b$ *impossible*, if by the use of such a term with such

an application, we should merely deny the *possibility* of obtaining any conceivable numerical result, when the number *a* was less than the number *b*."[19] In symbolic algebra, though, one can always "subtract" one letter from another: the result is simply the arrangement of symbols "*a* – *b*". Symbolic algebra specifies rules for the manipulation of symbols, inspired by similar valid rules in arithmetical algebra, but without the presupposition that the symbols denote particular numbers. So in symbolic algebra, the string of signs "– 3" could be the correct outcome of a series of sanctioned manipulations without any thought that there exists a negative number for the outcome to denote.

Both Wallis and Peacock acknowledge three relevant universes of objects: the numbers (subject matter of arithmetic), the magnitudes (subject matter of geometry) and the *mathematical symbols*. Some symbols, such as "3", denote numbers. Some symbols, such as "– 3", fail to denote numbers, but can still (under the correct interpretation) denote magnitudes (or points).

Without irrational and negative numbers, any attempt to construct a numerical *doppelgänger* for a geometrical space (such as R^2 for the Euclidean plane E^2) is bound to fail. Without irrational and negative numbers, the usual coordinatization of the plane by means of numbers is impossible. So even as late as 1842, a fundamental gap in mathematics between the theory of number and the theory of magnitude remained. There may have been loose talk about irrational and negative numbers, but no rigorous arithmetical foundation for them existed. This challenge was taken up in 1872 by Richard Dedekind.

Dedekind and the Construction of Irrational Numbers

If there is any single work that can serve as a useful foil to the project of this book, it is Dedekind's 1872 essay *Continuity and Irrational Numbers*. Dedekind is magnificently clear about the goals of his work and the gaps he perceives in contemporary mathematical practice. He was dissatisfied by any appeal to

[19] Peacock 1842, p. 7

geometrical intuition when trying to establish proofs in arithmetic, in particular proofs in differential calculus.[20] Dedekind's primary goal is to develop an account of continuity that is logically independent of any geometrical notions. He is acutely aware that the notion of a numerical domain (in particular, the rational numbers) being discontinuous is usually explicated by means of the Euclidean line. He regards this as an illegitimate intrusion of geometrical concepts into arithmetic. The theory of numbers ought to be developed entirely from within its own resources, using (as we would say) nothing but the positive integers and set theory. Dedekind articulates a criterion of continuity and discontinuity that can be applied directly to numbers with no reference to geometrical lines, and shows that by this criterion the rational numbers are discontinuous. He then sets about constructing new numbers– the irrational numbers– such that together with the rationals they form a continuous domain.

One aspect of Dedekind's approach is his constructivism. For him, numbers are creatures of the human mind, and he would not be able to make any sense of questions about whether irrational numbers (or rational numbers, for that matter) "really exist". The counting numbers are determined by methods of enumeration, and from those one can define the basic operations of addition, subtraction, multiplication and division (where possible). Since the positive integers are not closed under these operations, we create new numbers, including fractional and negative numbers, specifying the basic operations on them in terms of the basic operations on the counting numbers. This closes the set under all operations save division by zero, and constitutes a *body of numbers* (*Zalhkörper*).[21] One could see why the counting numbers in themselves are in some sense "incomplete" since they are not closed under the basic arithmetic operations, but in what sense could the set of rational numbers be recognized as "discontinuous"?

Dedekind reiterates the standard geometrical reasoning, which he is intent to reject. Starting with a line in Euclidean space, arbitrarily pick a point to count as the "origin" and arbitrarily pick a line segment to count as a "unit". Now mark out all the

[20] Dedekind 1963, p. 1
[21] Ibid, p. 5

points on the line whose distance from the "origin" is commensurable with the "unit". This set of points is isomorphic to the set of rational numbers: once one arbitrarily picks a direction from the "origin" to count as the "positive" direction, each rational number is associated with a unique point on the line. Note that under this association, the arithmetic order of the numbers is mirrored by the geometrical order of points on the line. This sort of construction, involving an arbitrary choice of "origin", "unit", and "positive direction", is familiar from the discussion of Descartes above. Now it is evident that the set of points on the geometrical line that results from this procedure is discontinuous in the straightforward sense that some of the point on the line are missing from the set. There is a point in the "positive" direction from the origin whose length is the same as that of the diagonal of a square built on the "unit", but that point will not be in the set. The set of points picked out by this procedure is incomplete in the direct sense of not containing all the points on the line. Furthermore, between any pair of points in the set, there will be missing points. Reasoning backward from the line to the rational numbers, then, one might try to argue that they too are "discontinuous". As Dedekind says "Since further it can be easily shown that there are infinitely many lengths that are incommensurable with the unit of length, we may affirm: The straight line *L* is infinitely richer in point-individuals than the domain *R* of rational numbers in number-individuals".[22]

But Dedekind rejects this argument. He insists that the notion of extensive magnitude it foreign to arithmetic: "I demand that arithmetic shall be developed out of itself."[23] This means first articulating a notion of continuity that makes no use of extensive magnitude, and then constructing enough new numbers to constitute a continuous *Zahlkörper*. And this new domain of numbers must be fundamentally arithmetical: the basic arithmetical operations on them must be reduced to operations of integers. Dedekind claims that before his own work, even

[22] Ibid, p. 9
[23] Ibid. p. 10

mathematicians who accepted irrational numbers could not prove things about their basic arithmetic properties: no one had proven, for example, that $\sqrt{2} \bullet \sqrt{3} = \sqrt{6}$.[24]

One might wonder, though: even granting a definition of continuity according to which the rationals can be shown to be discontinuous, why should one feel obligated to construct a continuous *Zahlkörper*? Having strictly separated the theory of number for the theory of extensive magnitude, as the Greeks had done, why not let each go its separate way?

One answer is contained in the Preface to the First Edition of *Essays on the Theory of Numbers*:

> In speaking of arithmetic (algebra, analysis) as a part of logic, I mean to imply that I consider the number-concept entirely independent of the notions of space and time, that I consider it an immediate result from the laws of thought. My answer to the problems propounded in the title of this paper is, then, briefly this: numbers are free creations of the human mind; they serve as a means of apprehending more easily and more sharply the difference of things. It is only through the purely logical process of building up the science of numbers, and by thus acquiring the continuous number-domain that we are prepared accurately to investigate our notions of space and time by bringing them into relation with this number-domain created in our mind.[25]

That is, Dedekind's view is that to think accurately and clearly requires thinking by means of numbers, so one needs a continuous number-domain to think clearly about a continuous spatial or temporal domain.

Dedekind does not believe that space and time are certainly continuous. He points out that all of the constructions possible in Euclidean geometry (i.e. using only straightedge and compass), and hence all Euclidean proofs, could be carried out on a discontinuous space.[26] And he says directly:

---

[24] Ibid. p. 22. Dedekind reiterates this claim on p. 40, in the Preface to the First Edition of *Essays on the Theory of Numbers*.
[25] Ibid, pp. 31-2
[26] Ibid, p. 37-8

> If space has at all a real existence it is *not* necessary for it to be continuous; many of its properties would remain the same even were it discontinuous. And if we knew for certain that space is discontinuous there would be nothing to prevent us, in case we so desired, from filling up its gaps, in thought, and thus making it continuous; this filling up would consist in a creation of new point-individuals and would have to be effected in accordance with the above principle [viz. that to every Dedekind *Schnitt* there should correspond a unique point].[27]

So Dedekind is unsure whether space exists at all, and if it does, whether it is continuous. Still, space might be real, and it might be continuous (even Euclidean), so we require sharp mathematical concepts by which we can consider the possibility. And those concepts must ultimately be arithmetical.

(Dedekind's talk of "filling up gaps" in space by the "creation of new point-individuals" is either sloppy or confused. If space is real, then it is not up to us to create points of space by an act of our mind, and whatever fiction we create will not be point-individuals in real space. He might think that the structure of any discontinuous space can be embedded into a continuous space, so that we could regard a real discontinuous space as part of a fictitious continuous space. But he has neither proven this contention, nor shown that it would make things clearer or more perspicuous to "fill up the gaps" this way.)

Of course, Dedekind had other reasons to pursue a continuous *Zahlkörper* beside adequacy to represent space and time. He regarded calculus as fundamentally arithmetic, and needed proofs about continuity to get results in infinitesimal analysis. But for our purposes, the connection to space and time is essential, so we will keep that aspect in focus.

Let's review Dedekind's "purely logical" account of the discontinuity of the rational numbers. The rational numbers (positive and negative) come equipped with an arithmetic order, represented by the "greater than" symbol ">". This can be

[27] Ibid, p. 12

arithmetically defined, since $a > b$ iff $a - b$ is positive, $a < b$ iff $a - b$ is negative, and $a = b$ iff $a - b$ is zero. This relation over the rationals has several formal properties that Dedekind points out: first, it is transitive; second, since the rationals are dense in the order, there are infinitely many numbers that lie between any pair of distinct numbers. The third property is the most important:

> III. If a is any definite number, then all numbers in the system R fall into two classes, A_1 and A_2, each of which contains infinitely many individuals; the first class A_1 comprises all numbers a_1 that are $< a$, the second class A_2 comprises all numbers a_2 that are $> a$; the number a may itself be assigned at pleasure to the first or second class, being respectively the greatest number of the first or the least of the second. In every case the separation of the system R into two classes A_1 and A_2 is such that every number of the first class A_1 is less than every number of the second class A_2.[28]

Such a separation of the numbers into two classes, each member of one being less than every member of the other, is called a cut (*Schnitt*). Clearly, to every rational number there correspond a pair of cuts (depending on which class the number is assigned to). But, as Dedekind points out, it is not the case that to every cut there corresponds a rational number that generates it, an obvious example being the cut that assigns a rational number to A_1 if its square is less than or equal to 2 and to A_2 if its square is greater than 2. The essence of continuity for Dedekind is this: to every element in the ordered set there corresponds a cut (really, a pair of "equivalent" cuts) *and to every cut there corresponds a member of the set that produces it*.[29] The rationals are not continuous because there are cuts that are not generated by any rational number.

Given his constructivism, it is an easy matter for Dedekind to produce a bigger set of logical elements: simply stipulate that every cut in the domain of

[28] Ibid., p. 6
[29] Ibid, p. 11

rationals should itself be an element of the new domain.[30] That leaves a lot of work to do. First, one must extend the ordering relation > to cover the new domain. Since each cut can be specified by just A_1 (all the rational numbers not in A_1 are in A_2), this is not very hard. In essence, the element corresponding to the set A_1 is greater than the element corresponding to A'_1 just in case for every member of A'_1 there is a greater member of A_1.[31] Once the new domain is ordered, one must show that it is continuous, by showing that every cut in *it* is produced by one of its elements, the elements now being cuts in the reals. By showing this, Dedekind proves that he has produced a continuous domain.

To make it a *number* domain (*Zahlkörper*) he now needs to define the basic arithmetic operations for the new elements. Again, it is not hard to see the strategy. To add the element that corresponds to A_1 to the element that corresponds to B_1, simply add each of the members in A_1 to each of the members in B_1. This produces a set of rationals, and all one need to do is prove that this new set itself corresponds to a cut. Similar definitions of subtraction, multiplication and division can be devised. Since the arithmetic operations on the rationals reduce to arithmetic operations on the integers, Dedekind has produced a continuous number domain, the real numbers. The equation $\sqrt{2}\bullet\sqrt{3} = \sqrt{6}$ can now be rigorously proven since the multiplicative structure of the reals has been specified.

Note an important architectural feature of Dedekind's account. At one level, he is especially concerned to create a continuum of numbers, and so to construct objects for which the basic arithmetic operations are defined. But the definition of *continuity* makes no use at all of most of the arithmetic structure. In fact, the only arithmetic structure employed in the proof of continuity of the reals (or discontinuity of the rationals) is the linear order imposed on the numbers by the "greater than" relation. Given only that relation–and omitting every other fact about

[30] Two cuts correspond to the same element if they differ only in that the greatest member of A_1 in one cut is the least member of A_2 in the other.

[31] This needs refinement since two cuts can correspond to the same rational number, but the details need not detain us.

additive or multiplicative structure– it is determined whether the number domain is continuous or not.

This feature of Dedekind's approach is quite deliberate. After all, the primary intuitive example of continuity is the straight line in Euclidean space, and that object, as Dedekind constantly reminds us, is composed of point-individuals rather than number-individuals. Points in a Euclidean line cannot be added or subtracted or multiplied or divided, so those arithmetic properties had better not appear in the definition of continuity. What the Euclidean line has in common with the number domain is a linear order: by arbitrarily choosing a direction on the line, we can define a relation symbolized by ">" that has just the structure needed for the concepts of continuity and discontinuity to be applicable. Dedekind points out exactly this analogy in *Continuity and Irrational Numbers*.[32] This feature of his definition is essential if the notion of continuity is to be applicable to physical space just as it is to numbers. So Dedekind's concern for an explicitly arithmetical definition of real numbers stands in contrast with his need for a non-arithmetical definition of continuity.

But this very feature of Dedekind's approach raises a puzzle. If one can characterize continuity in non-arithmetical terms, why think that it is only by "acquiring the continuous number-domain that we are prepared accurately to investigate our notions of space and time by bringing them into relation with this number-domain created in our mind"? Why not investigate our notions of space and time directly, using the very logical resources that he has created, *without the intervention of numbers at all*? This would seem to be a superior method, since the numbers drag along with them a tremendous amount of non-geometrical structure that is liable just to confuse the situation.

Furthermore, the real numbers form a "space" (in particular a "line") only in the metaphorical sense. Talk of "position on the number line" is shorthand for talk about the intrinsic arithmetical natures of numbers: 2 sits "between" 1 and 3 whether thought of as elements of the set of real numbers, or of rational numbers,

[32] Ibid, p. 7

or of integers. In contrast, one geometrical point sits between two others on a line only in virtue of the structure of the rest of the line. Since the geometrical elements are point-individuals, one cannot appeal to their nature to determine their linear order. So Dedekind is suggesting that the only way we can accurately apprehend a geometrical space is through the mediation of a metaphorical space as a representation. What he fails to do is give us even the slightly reason to think this is true.

These problems, which may seem a bit abstract, pale in comparison with a difficulty the Dedekind leaves completely untouched. Having argued that we can only clearly grasp a geometrical continuum via a numerical continuum, he bequeaths us only a *one-dimensional* numerical continuum. And it is not at all obvious how his account of continuity could be extended to anything but a one-dimensional space. Defining continuity requires defining cuts a class of objects, and the cuts are defined by reference to the linear order of the elements. Points in a one-dimensional geometrical space do instantiate such an order, but points in a two-dimensional space like the Euclidean plane do not. And without a linear order among the elements, Dedekind's account of continuity gets no purchase.

So even were we to accept Dedekind's unargued premise that one must think with numbers in order to clearly grasp geometry, he has not provided the tools to accomplish this in any but the most trivial geometrical case. Nonetheless, he has formulated one of the keys for unlocking geometrical structure. Even better, his approach to understanding continuity operates at the sub-metrical level: it only appeals to the order of elements in a set, not to distances between them. If we can find a way to extend his method beyond one dimension, we will have a way to characterize geometry without appeal to metrical considerations. The main burden of the theory of Linear Structures is to articulate precisely what this extension requires.

Overview and Terminological Conventions

The remainder of this book falls into two parts. The first deals exclusively with mathematical topics. We start with an overview and critique of standard toplogy in Chapter 1, followed by a presentation of the basic axioms of the theory of Linear Structures in Chapter 2. Each of the succeeding four chapters takes up a set of concepts that are defined in standard topology and explains how they are to be defined using the resources of Linear Structures. These alternative definitions usually give somewhat different results. Part 2 applies the new mathematical tools to physics. This requires reformulating some familiar physical theories in the language of Linear Structures. For Newtonian physics (Chapter 7) the reformulation is quite straightforward. Relativity Theory is tackled in Chapter 8, where the mathematics is shown to fit quite smoothly with the physics. Time order emerges as the central organizing structure in Relativity, a result more extensively discussed in Chapter 9. In the final chapter, we will apply these tools to some specific physical problems, including the problem of evaporating black holes.

One of the most vexing decisions that had to be made in composing this manuscript concerns terminological conventions. The theory of Linear Structures constitutes a novel mathematical language. In terms of its basic concepts we seek to define such informal, intuitive notions as “connected space”, “open set”, “continuous function” and so on. But these very notions have already been given formal definitions in standard topology. And the definitions given in the theory of Linear Structures are not even extensionally equivalent to the standard definitions. So the opportunity for confusion is almost unlimited.

One might think that the proper thing to do would be to create an entirely new set of terminology: standard topology got there first and staked a claim to the words. But there are several problems with that suggestion. One is practical: a well-chosen nomenclature is a powerful mnemonic device. Arbitrary terms carry no associations that can assist in recalling what they mean. So, for example, in standard topology a topological space can be T_0, T_1, T_2, T_3, T_4, T_5, Hausdorff, Urysohn or Lindelöf. These names convey nothing about the properties they denote. There is a certain austere virtue to this: the names cannot possibly be misleading. But there is a correlative disadvantage, in that neither can they be leading. One must simply

memorize the associated conditions, which makes thinking about them more arduous.

But beyond the merely practical, there is a deeper methodological point here. Names are chosen for certain formally defined properties because they are supposed to be formal explications of particular informal, intuitive notions that we start with. For example, there is no doubt that Dedekind's condition that there be an element in a domain that generates every cut is a perfectly clear condition (once the notion of a cut has been defined). But Dedekind did not set out merely to invent some formally defined property, he set out to explicate the pre-existing notion of continuity. If that had not been his intent, we would have little interest in his constructions. And if his definitions did not yield the right results, if, e.g., the Euclidean line turned out not to be "continuous" according to his criterion, then we would reject the definition. When names for formally defined properties are merely arbitrary stipulations, there can be no possibility of criticizing them: we cannot fault Prof. Hausdorff for getting wrong what should be meant by a "Hausdorff" space. But if all the terms in a mathematical theory were mere stipulations of this kind, then we would have no idea whatever what the theory was attempting to be a theory *of*. It is exactly because we start out with some notion of connectedness and openness and continuity that we take interest in a formal definition of these properties. And the theory of Linear Structures is an attempt to give strict formal accounts of the very same informally grasped geometrical properties that standard topology attempts to explicate. The theory of Linear Structures has the same right as standard topology to these terms, and inherits the same responsibility by using them: to show that the formally defined concepts really do articulate the informal, intuitive notions that go by the same name. So it is not only useful for the theory of Linear Structures to use much of the same terminology as standard topology, it is methodologically necessary.

Avoiding the standard nomenclature is unacceptable, and using the standard nomenclature invites endless confusion. I have cut this Gordian knot by a simple, if unconventional, technique. When I come to introduce the theory of Linear Structures in Chapter 2, terminology proprietary to that theory will all be written in

this font (Matura M7 Script Capitals). There is thereby no danger of mistaking the formal terminology of this theory either with terms in standard topology or more informal notions. The reader is constantly kept aware that these terms are not equivalent to the standard ones: a function can turn out to be continuous but not **continuous**, or a space **connected** but not connected. I leave to the judgment of the reader, when such conflicts occur, which seems the more intuitive result. It is largely by such judgments that value of this undertaking will assessed.