CHAPTER 13

Knowledge, Symbols, and Understanding

Kenny Easwaran

Like in any human endeavor, there are different kinds of knowledge one can have in mathematics. Someone reliable tells you that something is true, and you believe it, and this gives you a basic kind of knowledge. Or you work things out using a method you know to be reliable and convince yourself that this thing must be true. But most satisfying is when you manage to understand *why* it is true, by seeing the claim in a light that makes it clear that it *had* to be true. Ted Chiang's story, "Division by Zero", is about the destabilizing effect that occurs when this deepest level of knowledge is undone.

13.1 I.

Let's illustrate these different levels of knowledge with an example. There is a formula that you may or may not remember from an algebra class, that $b^2 - a^2 = (b + a)(b - a)$. That is, for any two numbers, a and b, the difference between their squares, b^2 and a^2, is a number that can be factored into the difference of the two numbers and the sum of the two numbers.

K. Easwaran (✉)
University of California, Irvine, CA, USA
e-mail: k.easwaran@uci.edu

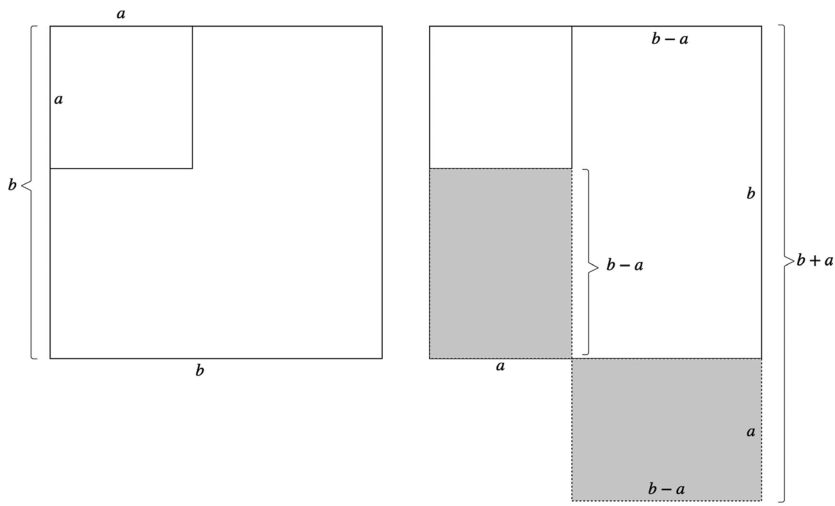
D. Friedell (ed.), *The Philosophy of Ted Chiang*,
https://doi.org/10.1007/978-3-031-81662-8_13

Assuming you trust me, or remember from an algebra class, you now have the first level of knowledge of the truth of this equation.

You can get a deeper level of knowledge by seeing the algebraic reasoning yourself. (Or you can skip the next two paragraphs to avoid the algebraic formalism and go directly to the geometric understanding, or skip all the way to the next section if you are happy remaining at the level of pure trust, both about the mathematical result and about the way the different levels of mathematical knowledge feel.)

We can start with the right side of the equation, $(b + a)(b - a)$, and expand it algebraically. This is a multiplication of two quantities formed by addition and subtraction, so we can apply the distributive rule of multiplication. Applying it to the left term, if you add two things together and then multiply by something else, the result must be the same as multiplying each of the two terms by that thing and then doing the addition, so $(b + a) \times (b - a) = b \times (b - a) + a \times (b - a)$. Each of the resulting terms is now something times a difference of two terms, and so this is the same as multiplying that thing by each term and then take the difference, yielding $(b \times b - a \times b) + (b \times a - a \times a)$. Looking at the middle two terms, the commutative rule tells us that $a \times b = b \times a$, so since one of them is negative and the other is positive, they must cancel. Thus, $(b + a)(b - a) = b \times b - a \times a$. But this term on the right just is what $b^2 - a^2$ means, so we have verified that the equation must hold.

So far, this is just a mechanical use of the formalism of algebra. If we've understood the rules of how the terms can be manipulated, we can use them to verify that the result is correct. But we still don't have a deep, intuitive understanding of it.



We can get more understanding by looking at a picture.

In the left part of the diagram, we see two squares, one whose side length is b (which thus has area b^2), and one nested inside it whose side length is a (and thus has area a^2). The difference of the squares, $b^2 - a^2$, is the area of the L-shaped region left behind when the a square is removed from the b square. On the right part of the diagram, we have kept the same shape but have also shaded in two rectangles. Looking at those shaded rectangles, we see that they are the same size, both of dimensions a by $(b - a)$, so they both have area $a \times (b - a)$. So the L-shaped region made up of the big white rectangle and the left grey rectangle must be the same area as the longer rectangle made up of the big white rectangle and the lower grey rectangle. But that latter, longer rectangle has dimensions $b + a$ by $b - a$, so its area is $(b + a) \times (b - a)$. This may give us a deeper understanding of the equation.

In fact, this diagram can even give us insight into the formal algebraic manipulations themselves. Notice that the big white rectangle has dimensions $b \times (b - a)$, and the smaller grey rectangle has dimensions $a \times (b - a)$. The fact that the rectangle has the same area, whether you rotate it so that the long side is horizontal or the long side is vertical, corresponds to the commutative rule, saying that $a \times (b - a) = (b - a) \times a$. Putting the rectangle on the bottom, to give us the longer rectangle, tells us that $(b + a) \times (b - a) = b \times (b - a) + a \times (b - a)$, as our first application of the distributive rule showed. The second set of applications of the distributive rule corresponds to seeing that the area of the white rectangle on the right is the area of the large square minus the left strip, and seeing that the area of the grey rectangle on the left is the left strip minus the small square. The final equation corresponds to noticing that the area of the L-shaped region is the whole square minus the left strip, plus the left strip minus the small square.

Hopefully, following this geometric explanation has helped provide some substantive insight and understanding beyond the formal symbol manipulation provided in the algebraic argument. This is by no means a full explanation—mathematicians who work on these sorts of issues in algebra and geometry will often bring in many other different areas of mathematics as well, such as number theory, set theory, topology, complex analysis, and so on, each one of which brings a different aspect of understanding to a result that can often be translated between them. We have just gotten a taste of this by looking at one result through two different branches of mathematics.

13.2 II.

Many of the important points in the story "Division by Zero" depend on the difference between mathematical insight and mechanical formalism. Renee prides herself on mostly working at the deeper level of insight, while many other mathematicians, including her department chair Peter Fabrisi, mostly work at the level of formal manipulation. Many central philosophical issues about mathematics involve this distinction, and the story alludes to many of them.

Intuitive understanding of a mathematical result is definitely more satisfying. But it's also not systematic—each piece of understanding relies on someone figuring out just what to focus on (which diagram, or which configuration of shapes, or which point) in order to get it. The formalism, instead, just provides a set of rules and a method for following them, that is guaranteed to get the answer if you work at it long enough. Math classes (particularly introductory classes whose students will go on to work on something else) often focus on the formalisms, because having a method that eventually finds the answer in all cases is more practical. But mathematicians themselves would usually rather have the insight that frames each particular result in its own unique way, rather than the ugly and unilluminating brute force method. We can think of the difference between a

Sherlock Holmes, who looks around aimlessly until he identifies the one clue that gets to the core of a case, and an insurance claims investigator who works down a checklist and eventually comes to a good enough conclusion about what must have happened.

In addition to the fact that the mechanical formalism is often systematic, it also usually both makes it easier to explain your reasoning to someone else and easier to verify that no mistake has been made. Although we like to fantasize about a Sherlock Holmes who convincingly identifies the true crux of the matter, in reality we're more comfortable with the investigator who can show us the checklist than with the wannabe Sherlock Holmes who may just have a convincing-sounding explanation that hides some hidden flaw.

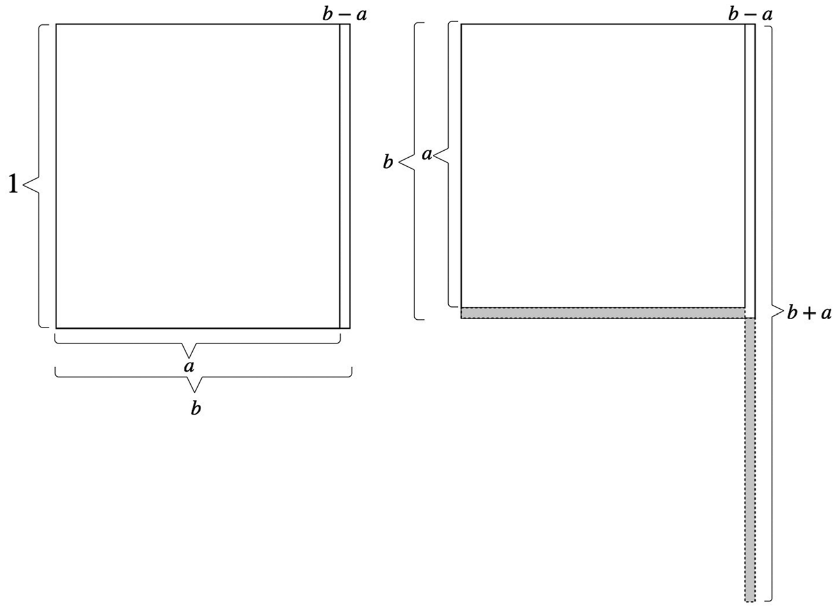
We often say that something is "more of an art than a science" when there is no checklist or recipe that can be followed to do it. But this is misleading about what research scientists actually do. What we refer to in that phrase as "a science" is a recipe or checklist that anyone can learn to follow (perhaps with a few years of university study). But researchers who actually develop these "sciences" are themselves working more in the mode of "an art". Just like Renee, a cutting-edge researcher in solar energy, or viral genetics, or computer engineering, will constantly look for new and creative insights into their field that they can derive from looking at it through a new lens, from a seemingly unrelated angle, in a way that no one can anticipate. One result of this collective research process is a set of formalized methods that can be taught to many people, so that every neighborhood can have mechanics and pharmacists and IT specialists that improve our lives.

This is not to say that only a select few people have the capacity for understanding or insight—just that the formalism is what we can systematically teach—while the understanding or insight is something that people have to organically develop. It can be a joy to work with a mechanic, or a doctor, or an accountant who has deep insight into the issues you are dealing with—but it isn't necessary if you just want someone who will fix your problem and give you a set of instructions to avoid running into it again.

13.3 III.

Now we can see what goes wrong when we let a formalism float free of the rules. The following argument is alluded to at the start of section 2 of the story "Division by Zero".

Let $a = b = 1$. Since $a = 1$, multiplying both sides by a yields $a^2 = a$. Similarly, since $b = 1$, multiplying both sides by b yields $b^2 = b$. Now subtract the two sides of $a^2 = a$ from the corresponding sides of $b^2 = b$, to yield $b^2 - a^2 = b - a$. By our earlier result, $b^2 - a^2 = (b - a) \times (b + a)$, so we have $(b - a) \times (b + a) = b - a$. Dividing both sides by $b - a$, we get that $b + a = 1$. But this means that $1 + 1 = 1$! Looking through more carefully, and plugging in the values $a = b = 1$, we can verify that every equation in our reasoning was correct, even up to $(b - a) \times (b + a) = b - a$. $((1 - 1) \times (1 + 1)$ is in fact 0×2, which does in fact equal $1 - 1$.) It only went wrong in the last step, when we divided by $b - a$, which meant that we were dividing by zero.

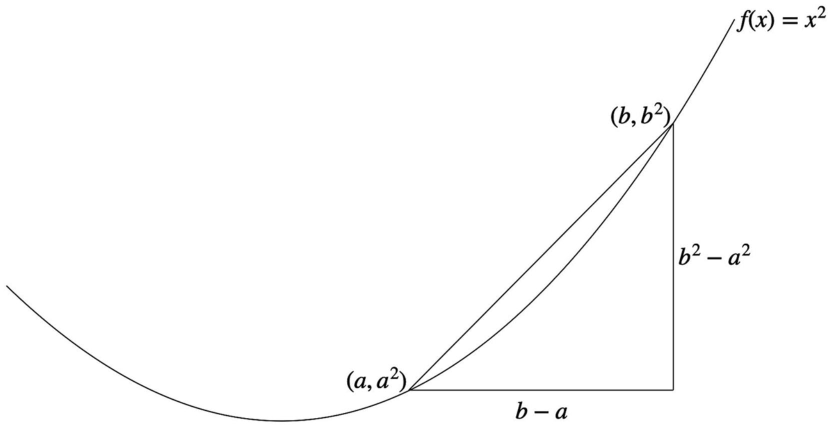


When we try to draw this out in a diagram, we can see what goes wrong.

This is what the diagram would look like if a and b were both very close to 1. On the left, we have rectangles of height 1, and width a and b, respectively. On the right, we have two nested squares, with side length a and b, respectively. When we claim that multiplying both sides of the equation $a = 1$ by a yields a new equation, $a^2 = a$, we are claiming that the square of side a on the right has the same area as the $a \times 1$ rectangle on the left. Similarly, the equation $b^2 = b$ claims that the square of side b on the right has the same area as the $b \times 1$ rectangle on the left. When we take the differences, we claim that the vertical sliver between the rectangles on the left has the same area as the L-shaped sliver between the squares on the right—which is the same as the area of the tall rectangle on the right. Since the two rectangles both have the width of $b - a$, we then claim that their height must be the same, so $1 = b + a$. We can see that when $a \neq b$ none of the steps in this reasoning are right—but when $a = b$ everything is in fact precisely right, even including the point where we say the two rectangles have the same area. But the reason they have the same area is that they both have area 0. Dividing by 0 corresponds to claiming that two rectangles with the same area and the same width must have the same height—which is true, as long as the width is not 0.

As the story says, the standard response is just to note that division by zero is not allowed—you break the formalism by doing that. But it is hard to notice this rule every time it matters—and some important parts of mathematical history are built on breaking it.

The central question of the mathematical subject we now know as "calculus" is about measuring the slope, not of a line, but of a general curve. Isaac Newton had an idea of how to do this, during the plague year of 1665 when he had to stay home from university. If we want to measure the slope of a curve that is given by some mathematical function, say, the parabola given by $f(x) = x^2$, we can approximate the slope at a point by finding two points near it and calculating the slope between them.



So, for instance, we can consider two points on this curve, (a, a^2) and (b, b^2), and calculate the slope of the line connecting them. The vertical separation between the two points is $b^2 - a^2$, and the horizontal separation is $b - a$, so the slope of this line is $\frac{b^2 - a^2}{b - a}$. As we have noted several times, $b^2 - a^2 = (b - a)(b + a)$, so when we divide by $b - a$, we get that this slope is $b + a$. Newton notes that when a and b are distinct numbers, we get the slope connecting two different points on the curve, which isn't quite what we want—so we should just let b and a be the *same* number if we want to calculate the slope of the curve *precisely* at that point. At the point (a, a^2), he says, the slope of the curve $f(x) = x^2$ is precisely $2a$. He did similar things to compute the slopes of other curves and put this together with his new theory of gravity to discover the shape of the solar system, and his method was central to all of physics for the next several centuries.

But, did you notice? He divided by $b - a$, and then said that $b = a$, so he was dividing by zero. The philosopher (and Bishop of the Anglican Church of Ireland) George Berkeley noticed and wrote a treatise in 1734 attacking Newton for this (though not by name, since by then Newton was too well-respected by contemporaries for even a luminary of the church to go after). Just as atheists ridiculed the church for teaching that one and the same God is sometimes the father, sometimes the son, and sometimes the holy ghost, Berkeley ridiculed the "infidel mathematician" for teaching that $b - a$ is sometimes non-zero, when you divide by it, and sometimes zero, when you calculate at the end, and must thus be just the "ghost of a departed quantity".

Since Newton's formal tools for calculation worked and didn't seem to lead to contradiction, most mathematicians just ignored Berkeley and went on doing what they were doing. But by the early nineteenth century, contradictions were building up. Leonhard Euler had extended Newton's methods to compute various infinite sums and solved some long-standing problems—but also ended up with strange claims like $1 + 2 + 3 + \ldots = -1/12$. Augustin Cauchy used Newton's methods to "prove" that a certain construction always yields a continuous function—but Joseph Fourier showed how to construct the discontinuous sawtooth wave in this way. Giovanni Saccheri "proved" that Euclid's parallel postulate followed from the others—but Bernhard Riemann showed that it was possible to reject the parallel postulate without trouble.

By the late nineteenth century, these errors were being fixed. Karl Weierstrass had developed a new formalism to justify Newton's calculations that avoided division by zero, and he and others separated Euler's truths from his mistakes and showed where Cauchy and Saccheri had gone wrong. This set the stage in 1900 for David Hilbert to propose, as one of the most important open problems for mathematics the challenge of finding a formalism that would work for all of mathematics and prove that it is consistent. The introductions to many of the numbered sections of the story refer to key moments in this project—Russell and Whitehead's *Principia Mathematica*, which was a formalism sufficient to encompass all the mathematics of the time, Gödel's proof that only a stronger system could ever prove this system consistent, Gentzen's construction of such a stronger system.

13.4 IV.

If Hilbert's program had succeeded, then mathematics would have been consistent—but reduced to a formal system. The invention of computers would have meant that mathematics no longer existed as a field of research. There would be no more reason for the mathematics that is "more of an art than a science". But it would have been sound. The philosopher Imre Lakatos argued (1978) that early calculus mostly avoided contradiction because of the understanding that early mathematicians had, even though their formalism was unsound. But Hilbert's dream was a sound formalism that would have made understanding unnecessary.

Gödel, like many other mathematicians and philosophers since then, saw his results as preserving the importance of creativity and understanding in mathematics. But people whose lives depend on our ability to solve difficult mathematical problems (which is increasingly all of us, as mathematics becomes more central to biology, banking, ballistics, and everything else) may have wanted there to be a way to automate mathematics and let humans pursue creativity and understanding in music or poetry instead. Even despite the fact that there can be no universal computer system for mathematics, some mathematicians have argued in recent decades that it is important enough to get things right that we should start producing proofs that can be verified by computers, rather than proofs that humans can use for their own purposes. I have said more about why I think mathematicians generally reject this movement in two of my previous publications (Easwaran 2009, 2015).

The story of "Division by Zero" explores the opposite catastrophe and reaches a surprisingly similar conclusion, where there is nothing left of interest in mathematics. When mathematical understanding inevitably leads to contradiction, all that is left is meaningless symbol manipulation. This is often what many people assume that mathematics is, when they haven't experienced the understanding for themselves. But mathematicians themselves have always sought understanding and used symbols only as a way to keep their expressions more precise for communication.

References

Easwaran, K. 2009. Probabilistic Proofs and Transferability. *Philosophia Mathematica* 17 (3): 341–362.

———. 2015. Rebutting and Undercutting in Mathematics. *Philosophical Perspectives* 29:146–162.

Lakatos, I. 1978. Cauchy and the Continuum. *The Mathematical Intelligencer* 1 (3): 151–161.