Annals of Mathematics and Philosophy
***vol.* 1, n° 1**, 2023

$M \times \Phi$

What Makes Mathematicians Believe Unproved Mathematical Statements?

Timothy Gowers[1]

Abstract. This paper considers the reasons mathematicians give for making probabilistic judgments about unproved mathematical statements, and discusses how one might interpret and justify such judgments more formally. Following Pólya, I argue that we update our probabilistic judgments in a broadly Bayesian way, while to explain what they mean in the first place, I argue that they are referring not so much to the truth of the statements as to the likely existence or otherwise of reasons for them. The link between the two is provided by a "no-miracle" principle, which says that a surprising mathematical statement will not be true unless it is true for a reason. This principle applies only to statements that are sufficiently natural, so the paper also sets out criteria for a statement to be more or less natural.

§ 1. — Introduction.

A feature of mathematics that marks it out from all other disciplines is that it has a formal notion of proof, which makes it possible, at least in principle, to justify a mathematical statement not just beyond all reasonable doubt, but beyond any doubt at all.[2] As many people have commented, justifications as they appear in typical research articles are not formal proofs but more like blueprints that are sufficiently detailed to convince experts that formal proofs exist. Some of these blueprints are very complicated and use ideas

[1] This article is a written version of two seminars I gave at the Collège de France.

[2] Unless one wishes to take the drastic step of doubting either some very basic axioms or some very simple rules of deduction.

that are understood by only very few mathematicians, so it can be reasonable to be uncertain about whether they are in fact correct. Nevertheless, there is a large body of mathematical knowledge that is sufficiently simple and widely understood that we can be sure that it will not be overthrown in the way that, say, Newton's theory of gravitation was overthrown by Einstein's. There is also a smaller but rapidly growing body of knowledge that *has* now been completely formalized with the help of computers.

Because proofs have become the gold standard, at least for pure mathematicians, in the sense that a statement is "accepted" if and only if somebody has proved it, it is tempting to divide mathematical statements into three classes: definitely true, definitely false, and don't know. (One might also wish to divide the last class further, into statements with proofs or disproofs that have not yet been discovered, and statements that are undecidable.) However, this is a very crude classification that does not give a complete picture of mathematical practice, since mathematicians have varying degrees of belief in different unproved statements. For example, they are extremely confident that Goldbach's conjecture is true and that π is a normal number, they are quietly confident but with not quite 100% certainty that the Riemann hypothesis is true, they think that almost certainly P$\neq$NP (though a few outliers think that this confidence is misplaced), and think it is probably not possible to factorize a large integer in subexponential time but would not be unduly surprised if it turned out to be possible. (I shall discuss all these examples in more detail later.)

It is a simple empirical fact that mathematicians make probabilistic judgments about mathematical statements. They do not actually assign precise probabilities to them, but they use words and phrases like "probably", "almost certainly", "unlikely", and so on, and if pressed might even be ready to translate those into odds that they would be prepared to accept in a bet. And yet this fact raises many questions. A particularly basic one is what it even means to say something like that a statement is "probably true". A second is a descriptive one: what causes mathematicians to be more confident in the truth of some unproved statements than others? A third is whether mathematicians are behaving rationally when they make the judgments they make. And a fourth is *why* mathematicians make these judgments.

My focus in this article will be the second question, though it is related to all the others. But before I get on to it, I would briefly

like to tackle the fourth, since I believe that it has an easy answer, though one that is also easy to overlook.

Contrary to what one might at first think, assessing the likelihood of as yet unproved statements is not just something that mathematicians do for amusement when they are talking shop with colleagues. Rather, it is a fundamental part of the research process. Solving a mathematics problem is extremely difficult and requires a big investment of time. In order to use one's time wisely, one has to make judgments about which statements are likely to be true, which are likely to have easy proofs and which not, which are likely to have interesting consequences, and so on.

That is not to say that trying to prove a false statement is always a complete waste of time. For some problems there is virtually no difference between trying to find a proof and trying to find a counterexample, since a good way to find a proof of a true statement is to try to disprove it and to examine why one's attempts fail. But for some problems that approach gets you nowhere. Even then it may be useful to try to prove a false statement. For example, early in my career I spent a long time trying to prove a theorem about Banach spaces that would have solved a famous open problem. I failed in my attempt, and later a highly ingenious counterexample was discovered. However, my time was not at all wasted: some of the ideas I had were extremely useful to me for solving other problems.

Nevertheless, the point remains that judging the likely truth of a statement is a very important skill for a mathematician. So far I have discussed entire research problems, but probabilistic judgments become even more important when they are part of the *process of solving* such problems. Typically, in order to solve a problem, one must use a top-down approach, which will involve a lot of guesswork. For instance, if one is trying to prove that $P \implies Q$, one often tries to identify a statement R for which one can prove that $P \implies R$ and that $R \implies Q$. There may be several candidates, in which case one must decide which one to try first, and having made the decision one must continually think about whether and at what point to abandon that line of attack and try another. In general, to be successful at research one must search efficiently for proofs, and that requires a sophisticated search strategy in which probabilistic judgments, whatever they might be, play an essential role. For this reason, understanding how such judgments are made and to what extent they are justified is not merely of philosophical interest: it is also of practical interest to mathematicians.

The rest of the paper is organized as follows. I begin with an informal discussion of several major open problems and some of the reasons mathematicians have given for what they believe the answers to be. I follow this with a discussion of how mathematicians update their beliefs in the light of changes to what they know: some basic rules for this were put forward by Pólya, which, as David Corfield has pointed out, it is natural to regard today as a form of Bayesian reasoning. In the following section I try to explain what probabilistic judgments about mathematical statements could possibly mean given that their truth values (in so far as they have truth values) are fixed. For this purpose I formulate two closely related principles that I call the "no-coincidence principle" and the "no-miracle principle". This is followed by a discussion of naturalness of mathematical statements, which I argue is central to any understanding of how we make probabilistic judgments. Issues arise here that are very similar to those raised by Goodman in his famous "new riddle of induction". In order to deal with Goodman-type paradoxes I formulate another basic principle that I call the "nice-formula principle". In the next section I argue that naturalness in mathematics is closely related with levels of abstraction — roughly speaking, the more it is possible to formulate a statement or definition without resorting to constants (as opposed to parameters that vary according to context), the more natural it is. I finish with a brief summary of the argument put forward over the course of the preceding sections.

§ 2. — Some examples of unproved statements.

In this section I shall look at several major unsolved mathematical problems. For some of them there are answers that are widely believed to be true, while for others there is much greater uncertainty. I shall discuss the kinds of reasons that mathematicians give, or in some cases actually have given, for their assessments of these problems. In later sections I shall try to draw some general conclusions from these and other examples.

2.1. Goldbach's conjecture. Goldbach's conjecture is the assertion that every even number greater than 4 is the sum of two odd primes. For example, 1984=1979+5, and both 1979 and 5 are prime numbers. Goldbach's conjecture is one of the most famous open

problems in number theory, but most mathematicians, at least if they have thought about the problem for a bit, are extremely confident that it is a true statement. What explains this confidence?

This question was considered in detail in an article by Alan Baker[3] on the use of scientific induction in mathematics. Goldbach's conjecture has been verified for all even numbers up to 4,000,000,000,000,000,000. Baker asked to what extent this kind of verification increased mathematicians' confidence in a statement, and to what extent it *should* increase their confidence, and concluded that in both cases the answer was not much.

One of the reasons he gave for this conclusion was that for any fixed n, the integers smaller than n are "minuscule", in the sense that almost all integers are far bigger. To a mathematician, this objection has force only if one has reason to suspect that the minuscule (in this sense) integers are not representative, and that varies from problem to problem. For instance, if I were to learn that the maximum possible value of a simple parameter associated with n-vertex graphs was $n-1$ for all n up to 10, then I might well regard that as very convincing evidence that it was always $n-1$, reasoning that any behaviour that would lead to a more complicated function would almost certainly have shown up by then. But there are examples in number theory of phenomena that show up only for very large numbers, so in general I would want to check further for a problem like Goldbach's conjecture.

Probably the reaction of most mathematicians, if all they knew about Goldbach's conjecture was that it was true up to 4,000,000,000,000,000,000, would be to think that it is probably true, but that it might be one of those funny problems, which are quite rare but which definitely exist, where the smallest counterexample is very large. However, as Baker pointed out, we actually know a lot more than this. Confidence in the truth of Goldbach's conjecture was greatly increased when tables were produced that showed not just *that* every even number up to a certain point was a sum of two odd primes, but also *in how many ways* that was the case. This revealed that, despite some fluctuation, on average the number of ways steadily increased as the even number increased. Therefore, small even numbers, to the extent that they are unrepresentative, are unrepresentative in the right direction: the evidence suggests that if you take a very large even number, then not only will it be

[3] *Is there a problem of induction for Mathematics?* in M. Potter (ed.), Mathematical Knowledge, Oxford University Press, pp. 57-71 (2007)

expressible as the sum of two primes, but it will be expressible in a huge number of ways.

However, Baker did not give the whole story, as there is an even stronger reason to be impressed by these tables, one that is so compelling as to leave virtually no room for doubt in the truth of Goldbach's conjecture.

A striking fact about the primes is that they occur somewhat sporadically, behaving in many ways as though they have been chosen randomly, subject to certain constraints. A sign of this is that all attempts to produce a non-artificial formula for the nth prime have failed, to the point where no mathematician seriously believes that such a formula exists (which is of course another example of a probabilistic judgment).

What does "as though they were chosen randomly" mean? One of the most famous theorems in mathematics, the prime number theorem proved independently by Hadamard and de la Vallée Poussin, states that the number of primes up to n is approximately $n/\log n$, so to a first approximation one might say that the primes look like a set that you would obtain if for each n you were to choose it with probability $1/\log n$, making all choices independently. However, such a set would contain a roughly equal number of even and odd numbers, which is certainly not the case for the primes. A more accurate random model of the primes is to choose each n as follows. If it has a small divisor such as 5 (the limit of what counts as small grows slowly with n), then do not choose n. Otherwise, choose n with a probability somewhat larger than $1/\log n$ (which can be calculated precisely) to compensate for the numbers that have been discarded.

This model can be used to predict, for a given large even number n, the approximate number of ways of writing n as a sum of two primes. For instance, if n is a multiple of 3 and p is a prime not equal to 3, then p is not a multiple of 3, which guarantees that $n-p$ is also not a multiple of 3, so p's being prime is positively correlated with $(n-p)$'s being prime. By contrast, if n is of the form $3m+1$, then the only way for p and $n-p$ both to avoid being multiples of 3 is if p is of the form $3m+2$, which is true for only about half of all primes. With this kind of reasoning one ends up hypothesizing that on average multiples of 3 can be written as a sum of two primes in about twice as many ways as non-multiples of 3. More generally, one can write down an approximate formula for the number of ways that n should be representable as a sum of two primes,

in terms of which small factors it has. And this formula turns out to be remarkably accurate. The tables show not just that large even numbers can be written in *many* ways as a sum of two primes, but that they can be written *in almost exactly the number of ways one would expect* if the random model of primes is a good one.

Why should this increase our confidence? The answer is twofold. First, the experimental evidence is now confirming a significantly more general statement (namely that the random model of the primes makes accurate predictions). The more general a statement is, the more diverse opportunities it has to be false, and therefore the more impressed one is when it is confirmed. Secondly, the predictions made by this more general statement are much more precise, and therefore again it is correspondingly easier for the predictions to fail, and therefore more impressive when they turn out to be correct.

2.2. The normality of π. A positive real number x is said to be *normal* if its decimal expansion looks random in the sense that every sequence of digits occurs with the frequency one would expect if the digits had been chosen at random. More precisely, suppose that the decimal expansion of x is $x_0.x_1x_2x_3x_4\dots$, where x_0 is a non-negative integer and each of $x_1, x_2, \dots$ is an integer between 0 and 9. Then given any sequence of digits, such as 137, and any positive integer n, one can look at all the sequences $x_ix_{i+1}x_{i+2}$ with $1 \leqslant i \leqslant n$ and count how many of them are the sequence 137. If the x_i have been chosen randomly, then each triple $x_ix_{i+1}x_{i+2}$ has a $1/1000$ chance of being 137, so one would expect the proportion of triples that give 137 to converge to $1/1000$ as n tends to infinity. The number x is said to be normal if that happens for every small sequence: that is, each digit occurs with a frequency that converges to $1/10$, each pair of digits with a frequency that converges to $1/100$, and so on.

It is widely believed that π is a normal number, which would have amusing consequences such as that your date of birth must occur somewhere in its decimal expansion. However, proving the normality of π is a wide open problem — it may even be an unprovable statement. To give an idea of our level of ignorance, it is not even known that there is not some point beyond which all digits are equal to 0 or 1, which would be an extremely strong refutation of the conjecture that π is normal.

So why do we believe that π is normal? Certainly the experimental evidence is quite convincing: if one examines a table of the first thousand digits of π, say, they really do look pretty random, and statistical tests have been carried out much further than this without throwing up any anomalies. Furthermore, there is no particular reason to think in this context that the "minuscule" number of digits we can examine form an unrepresentative sample.

If pressed further, I myself would use the following two-step argument.

1. If π is not normal, then there will have to be some *reason* for any bias that might occur.
2. The kinds of reasons I can imagine are almost all ones that would lead to a bias showing up very clearly within the first million digits.

This argument falls far short of a proof, of course. When I said that there would have to be a reason for any bias, I did not rigorously rule out the possibility that π might "just happen" to have only finitely many 7s in its decimal expansion. But all my mathematical experience tells me that if such a remarkable phenomenon were to occur, there would in fact be an explanation for it.

The second part of the argument is slightly less convincing, since there are some notorious examples of phenomena that show up only for surprisingly large integers, despite having very natural explanations. One is the remarkable fact that

$$e^{\pi\sqrt{163}} = 262537412640768743.99999999999925\ldots.$$

It would be very reasonable to suppose, after calculating the expansion of $e^{\pi\sqrt{163}}$ up to twelve decimal places and obtaining the answer 262537412640768743.999999999999, that in fact it is equal to the integer 262537412640768744, and the justification would be very similar: it is hard to imagine a reason for a pattern like this that continues for twelve decimal places without continuing for ever.

However, there is such a reason, and in fact it is not all that strange, though it depends on some advanced number theory. It turns out that one can expand $e^{\pi\sqrt{163}}$ in a natural way as a series where the first two terms are $(640320)^3$ and 744, and all remaining terms are extremely small, and that is why one obtains a number that is very close to an integer without actually being an integer.

There are many examples throughout mathematics of simply defined numbers that turn out to be very large (and there are even

theoretical reasons, related to Gödel's theorem, to expect this to be the case), so the mere fact that a pattern has been observed to continue for a long time is not always overwhelming evidence in favour of the hypothesis that it continues for ever. However, there is a difference between the example of $e^{\pi\sqrt{163}}$ and the hypothetical example of a failure of π to be normal. In the first case, there is a pattern that lasts for a long time before coming to an end. But for π to fail to be normal, one would require the opposite phenomenon: a pattern that only *starts* after a long time. The decimal expansion of π has been calculated up to 62.8 trillion digits without showing any sign of a pattern, and it is very hard to imagine what a proof could conceivably look like that would establish that some bias crept in after that point.[(4)]

2.3. $R(k,k)^{1/k}$ **tends to a limit.** A central theorem in combinatorics, Ramsey's theorem, asserts that for every k there exists an n such that if all the edges of a complete graph with n vertices are coloured either red or blue, then there must be k vertices entirely linked by red edges or k vertices entirely linked by blue vertices. The smallest such n is denoted $R(k,k)$. It is known that $2^{k/2} \leqslant R(k,k) \leqslant 4^k$. These upper bounds are very far apart, but it is a major open problem to improve substantially on either of them.

It is tempting to assume that there must be some constant C, lying between $\sqrt{2}$ and 2, such that $R(k,k)$ is roughly equal to C^k for all k. Plausible candidates for C are $\sqrt{2}, 2$, and 4. To be more precise about this assumption, what seems very likely to be true is that there is some C such that for any $\alpha < C$ and $\beta > C$, $R(k,k)$ lies between α^k and β^k when k is sufficiently large. A more concise way of saying this is to say that the quantity $R(k,k)^{1/k}$ converges to a limit C as k tends to infinity.

Almost all combinatorialists are (if my feelings are anything to go by) confident that this is true. The alternative is that as k gets larger and larger, the Ramsey number $R(k,k)$ "oscillates", in the sense that there are two constants $A < B$ such that infinitely often $R(k,k) \leqslant A^k$ and infinitely often $R(k,k) \geqslant B^k$.

[(4)] Strangely enough, there might be more hope of proving that the base-16 expansion of π is not normal, owing to a remarkable formula of Bailey, Borwein and Plouffe that enables that expansion to be calculated extremely rapidly. However, this possibility has been investigated and the best one can say at the moment is that their formula reduces the question of the normality of π to a simpler-looking but still wide open question, the probable answer to which would imply that π is normal even in base 16.

Why does this alternative possibility seem so unlikely? One reason is that there are many examples of exponential growth rates that occur naturally and are known not to oscillate in this sense, and none (or none that I can think of) that do oscillate. However, this is not a completely convincing argument, because the examples in question, of functions f such that $f(k)^{1/k}$ converges, are often examples for the same reason. A function is called *submultiplicative* if $f(ab)\leqslant f(a)f(b)$ for every a and b, and *supermultiplicative* if $f(ab)\geqslant f(a)f(b)$ for every a and b. It is not too hard to prove that if a function f is either submultiplicative or supermultiplicative, then $f(k)^{1/k}$ converges to a limit, and this is the reason that many functions that occur naturally have this property. However, the function $f(k)=R(k,k)$ is not submultiplicative, and it is not obviously supermultiplicative either (though so few Ramsey numbers are known that one cannot rule out the possibility), so if $R(k,k)^{1/k}$ converges it will probably have to be for a different reason.

So why, despite there being no particular reason to suppose that $R(k,k)$ is submultiplicative or supermultiplicative, are mathematicians so confident that $R(k,k)^{1/k}$ converges? This seems to be another case where, *pace* Alan Baker, there is no reason to suppose that minuscule natural numbers are unrepresentative, so if $R(k,k)$ is about C^k for some very large (by human standards) k, then probably $R(k,k)$ is about C^k for all larger k as well.

This argument raises questions in its turn, since one could replace C^k by another function $D(k)$ defined by a formula such as $(2+\cos(\pi\log_2 k)/10)^k$, which I have designed so that $D(k)=(2.1)^k$ when k is equal to an even power of 2 and $D(k)=(1.9)^k$ when k is equal to an odd power of 2. If it could be established that $R(k,k)$ was approximately equal to $D(k)$ for some very large k, would that suggest that $R(k,k)$ was approximately $D(k)$ for all larger k? Clearly not, but why not?

The answer must be that in some way the hypothesis that $R(k,k)\approx D(k)$ is an artificial one, which leads to the question of how to distinguish between natural and artificial hypotheses. This is a mathematical version of Goodman's new riddle of induction, to which I shall return later.

2.4. The average end-to-end distance of a two-dimensional self-avoiding walk. A two-dimensional self-avoiding walk of length n is a path of length n in the two-dimensional integer grid $\mathbb{Z}^2$ that starts at $(0,0)$, moves at each step to one of the four neighbouring points,

where the neighbours of (x, y) are $(x+1, y), (x-1, y), (x, y+1)$ and $(x, y-1)$, and do not visit any point more than once. Despite this simple definition, self-avoiding walks are very hard to analyse, and many basic questions about them are still unanswered.

One of these is to determine the average end-to-end distance of an n-step self-avoiding walk chosen at random from all such walks — that is, to work out how far, on average, the last point of the walk is from $(0, 0)$. It is known to be at least $n^{1/4}$ and a slight improvement to the trivial upper bound of n is also known. But that leaves a very large gap.

However, mathematicians are extremely confident that the average is in fact around $n^{3/4}$, and have been so for a long time. Why is this?

The reason is that there are arguments in favour of this conclusion that fall short of being rigorous proofs but that are nevertheless convincing. Pure mathematicians might call them heuristic arguments, but for many physicists they are completely satisfactory. They belong to a class of arguments where certain mathematical operations are carried out that are "invalid" (one famous example being to establish a formula that depends on dimension, which has to be an integer, and then to let the dimension "converge to zero"), despite which, they give precise predictions that are then confirmed by experimental evidence. This evidence can then be regarded as confirming not just the particular estimate of the average end-to-end distance of a self-avoiding walk, but also the more general hypothesis that there is some explanation, yet to be uncovered, for why physicists' methods seem to work so well.

Over the years, this confidence, which was already high, has become even higher after remarkable work of pure mathematicians that proves the $n^{3/4}$ estimate, as well as many related estimates that had been predicted by physicists, subject to a general hypothesis known as conformal invariance — a plausible hypothesis that asserts that the macroscopic behaviour of a self-avoiding walk (and several other related models) has a certain kind of symmetry. The conformal invariance hypothesis plays a role here that is somewhat similar to the role played for Goldbach's conjecture that the random model of the primes makes accurate predictions, with two notable differences: the conformal invariance hypothesis is a more precise statement, and the fact that one can deduce from it precise information about the behaviour of self-avoiding walks is far from obvious.

2.5. P versus NP. The P versus NP problem was formulated in 1971 by Stephen Cook, and independently in 1973 by Leonid Levin, and rapidly gained the status of being one of the most fundamental and important unsolved problems in mathematics and theoretical computer science. The basic object of study in theoretical computer science is the notion of an algorithm, and a basic question is whether, for a given computational task, there is an algorithm that can perform the task efficiently. A useful definition of "efficiently" is that of a polynomial-time algorithm, which means an algorithm for which there exists a polynomial P such that if the input to the algorithm has size n, then the time taken by the algorithm is at most $P(n)$. For example, standard long multiplication of two n-digit numbers is efficient in this sense, as you have to do roughly n^2 simple arithmetical operations. (There are also cleverer methods that are significantly more efficient, the current record being an algorithm that takes time closer to $n \log n$.)

To prove that there is an efficient algorithm for a given task, all you have to do is find such an algorithm. What appears to be far harder is to prove that there is *not* an efficient algorithm for some task. This general problem was brought into sharp focus by Cook and Levin, who observed that there is a large class of important computational problems that are all of equivalent difficulty in the following sense: if you can find a polynomial-time algorithm for one of them, then you can use it to create a polynomial-time algorithm for any other one.

This class of problems is closely related to a class of problems called NP (which stands for "non-deterministic polynomial time"). Loosely speaking, a problem belongs to NP if there is a polynomial-time algorithm for checking whether a proposed answer is correct. For example, an important problem in cryptography is the following: you are given a product n of two large prime numbers p and q and your task is to determine what p and q are. Nobody knows of an efficient way to do this, but if you are given two numbers p and q, then you can use long multiplication to check efficiently whether pq really does equal n.

The big unsolved problem is whether P equals NP. That is, if a computational task has the property that checking whether an answer is correct can be done efficiently, does that mean that finding the answer can also be done efficiently? The factorization example suggests that the answer is probably no: there just seems to be no reason for finding an answer to be anything like as easy as checking that an answer is correct once found.

Cook and Levin observed the remarkable fact that many natural problems in NP are what we call *NP-complete*, which means that if you can solve one of these problems then you can solve all problems in NP. One of the most famous such problems is the travelling salesman problem: you are given a collection of towns, and roads linking those towns, and the problem is to determine whether there is a route that visits each town exactly once before returning to its starting point. This belongs to NP, since if somebody shows you a route, you can easily check whether it visits each town exactly once. Much less obviously, it is an NP-complete problem, which means that if there is an efficient method for determining whether such a route exists, then that method can be converted into an efficient method for any other problem in NP. For example, it could be converted into a method for factorizing products of two large prime numbers. (By contrast, the factorization problem itself, though in NP, is almost certainly not NP-complete.)

The fact that there are many NP-complete problems and that finding an efficient algorithm for just one of them would prove that P=NP might seem to tip the balance back in favour of P and NP being equal after all. But that is not how mathematicians see it. A typical view, which I share, is that P and NP are almost certainly not equal, but that the level of certainty is not quite as high as it is for the other problems I have discussed so far. Rather than formulating my own reasons for this view, I shall leave that task to the well-known and philosophically inclined theoretical computer scientist Scott Aaronson. In an article entitled "$P \stackrel{?}{=} NP$" he writes the following.

> To my mind, however, the strongest argument for $P \neq NP$ involves the thousands of problems that have been shown to be NP-complete, and the thousands of other problems that have been shown to be in P. If just one of these problems had turned out to be both NP-complete and in P, that would've immediately implied P = NP. Thus, we could argue, the $P \neq NP$ hypothesis has had thousands of chances to be "falsified by observation." Yet somehow, in every case, the NP-completeness reductions and the polynomial-time algorithms "miraculously" avoid meeting each other — a phenomenon that I once described as the "invisible fence".

He then goes on to mention a particularly striking example of this invisible fence in action, related to a central example of an NP-complete problem, called 3-SAT. An instance of 3-SAT is a collection of variables $x_1, \dots, x_n$, each of which can take the value True or False, and also a collection of "clauses", each of which is the OR of three variables or their negations. For example, a possible clause is $x_2 \vee \neg x_5 \vee x_8$, which is true if and only if either x_2 is true or x_5 is false or x_8 is true. The problem is to determine whether there is a choice of values for the variables that makes all the clauses in the given collection true. This problem is a central example, because many proofs of NP-completeness proceed by showing that a problem can be reduced to 3-SAT, and since 3-SAT is already known to be NP-complete, that shows that the given problem is also NP-complete.

In the absence of an efficient algorithm for solving 3-SAT, one can set one's sights a little lower and try to find an efficient algorithm that will find an assignment of values to the variables that makes as many of the clauses true as possible. And an algorithm is known that will find values that make approximately 7/8 of the clauses true.[(5)]

A deep result of Johan Haståd shows that unless P=NP, one cannot do better than this. In other words, for any fraction α greater than 7/8, if there were an efficient algorithm for finding an assignment that satisfies at least a proportion α of the clauses, then P would equal NP. In other words again, the problem of finding a satisfying assignment for a proportion α of the clauses is NP-complete: the invisible fence appears at exactly 7/8.

2.6. The Riemann hypothesis. The Riemann hypothesis is regarded by many as the single most important unsolved problem in mathematics (though experts in number theory will often qualify this by pointing out that for many applications the Riemann hypothesis on its own is insufficient and what is needed is a more general form of it). Like the P versus NP problem, it is one of the Clay Millennium Problems, for which a million dollars is offered for a solution. It states that the zeros of the Riemann zeta function, which is defined for complex numbers with real part greater than 1 by the formula $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$ and on the rest of the complex plane apart from 1 by analytic continuation, are all either "trivial zeros" that occur at negative even integers or have real part equal to 1/2.

[(5)] It is even possible to say very roughly how it works. First note that if you choose the values at random, then each clause has a 7/8 probability of being true. But there is also a known derandomization procedure that can convert this simple observation into a deterministic algorithm.

The general attitude amongst experts towards the Riemann hypothesis is that there are certainly arguments that support the belief that the Riemann hypothesis is true, but there are also arguments against it, with the result that opinions vary, and while most mathematicians (or at least this is my impression) think it is probably true, they certainly do not rule out the possibility of its being false.

One argument in favour of it is that it has been checked with the help of computers that the imaginary part of any counterexample would have to be larger than a very large number that has steadily increased over the years and currently stands at about 3×10^{12}.

A second argument is that there is something very natural about the statement. A sign of this is that it has many equivalent formulations that look very different. One of these is related to the heuristic principle mentioned earlier in connection with Goldbach's conjecture, namely that the primes are distributed as randomly as they can be given certain obvious constraints. As mentioned earlier, the prime number theorem asserts that the density of primes close to n is approximately $1/\log n$. The Riemann hypothesis is (as Riemann himself showed) equivalent to the statement that the error term in this approximation is not much bigger than $\log n/\sqrt{n}$, which is the sort of error term one would expect if one were choosing the primes according to the random model mentioned before.

In an introduction to the Riemann hypothesis for the Clay Millennium Prizes, the famous analytic number theorist Enrico Bombieri, wrote a section entitled "Evidence for the Riemann hypothesis". He mentioned the computational evidence (which at the time was less advanced than it is today), but also a subtler argument, which is that there are reasons to believe that a counterexample, if it exists, should be in the vicinity of a number $1/2 + it$ for which $|\zeta(1/2 + it)|$ is particularly large. Such peaks are known to exist, but are also very infrequent, which raises the possibility that there is very large counterexample but no smaller one. The fact that this possibility is at least imaginable would seem to make the computational evidence mentioned above less persuasive. However, Bombieri went on to mention that Andrew Odlyzko, one of the leading experts on computation of zeros of the zeta function, had looked at zeros in regions where counterexamples would be more likely to be found (for instance, he had found 175 million consecutive zeros with imaginary parts around 10^{20}) and not found any. Thus, the Riemann hypothesis had in fact passed more stringent tests.

Other arguments given by Bombieri are that there are results showing that zeros with real parts that are not close to 1/2 are rare, and also results showing that at least 40% of the zeros have real part equal to 1/2. At first sight, this may not seem like very convincing evidence: why could it not be the case that a certain proportion of zeros have real part 1/2 but not all of them? The answer to this is quite interesting. The zeta function arises very naturally, as I have already stressed, so one would expect that the set of zeros of the zeta function would also look natural. But in that case one would expect either that they would all be on the critical line (that is, the vertical line of complex numbers of real part 1/2) or that they would have real parts that were scattered around 1/2 in a somewhat random-like way. It is much harder to conceive of a natural set that would be largely, but not completely, confined to the critical line.[6]

A final and important reason is that the Riemann hypothesis is now just one of a large number of analogous statements about different kinds of "zeta function", some of which are not open problems but major theorems. This raises the possibility of proving the Riemann hypothesis by considering classes of functions and not just one isolated function — a technique that has been very fruitful in much of mathematics. (There are also some analogues that turn out to be false, but they are not sufficiently closely related to the Riemann zeta function to dash this hope.) It also raises the possibility of using another proof-finding technique, namely analogy construction: one looks at the proofs for the other zeta functions and tries to solve problems of the form "This zeta function is to the Riemann zeta function as this element of the proof is to what?" and in that way one hopes to build up a proof. Bombieri writes, somewhat cautiously,

> In our opinion, these results in the geometric setting cannot be ignored as not relevant to the understanding of the classical Riemann hypothesis; the analogies are too compelling to be dismissed outright.

Another mathematician who is on record as saying that he believes strongly in the truth of the Riemann hypothesis is Peter

[6] I did not fully appreciate this point until, quite by chance and after having written a first draft of this article, I overheard my colleague Aled Walker explaining over lunch to a non-mathematician why he believed the Riemann hypothesis. He gave this reason as the one he found most convincing, so I asked him why, and he responded with something like the argument above.

Sarnak. In an interview in 2012 he mentions two points that back up his view (though he doesn't say that explicitly). One is that of the huge number of interesting consequences that have been discovered of the Riemann hypothesis, a significant number have subsequently been proved by methods that avoid appealing to it. Thus, rather like the statement that P$\neq$NP, it makes non-obvious predictions, several of which have now been rigorously confirmed and none of which have been disproved. His second reason (also alluded to by Bombieri) is that there is a fascinating connection between the distribution of the zeros of the Riemann zeta function and the distribution of the eigenvalues of a random matrix. The latter is quite well understood, and has led to predictions about the former that would have been impossible to guess, and which have been strongly supported by computational evidence. Thus, as with non-rigorous arguments about self-avoiding walks, one has the feeling that "something is going on" even if it is not yet fully understood.

Andrew Odlyzko's view after his computational work is completely agnostic: he thinks that the Riemann hypothesis could be true and recognises that there is evidence in favour of it, but would not be especially surprised if there turned out to be a large counterexample (which Bombieri seems far less willing to countenance, saying that it "would create havoc in the distribution of prime numbers"). Some clue as to the reason for his attitude can be found in an unpublished paper where he writes,

> The main conclusion that can be drawn from the data in this paper is that in many respects the zeta function reaches its asymptotic behavior slowly, so that even the neighbourhood of the 10^{20}th zero does not represent what happens much higher.

He then goes on to give some idea of why this slow convergence occurs.

2.7. The difficulty of factorizing large integers. Modern cryptography, and in particular internet security, rely heavily on protocols that would be insecure if integer factorization turned out not to be hard. And while it is widely believed that the problem is indeed hard, that belief is nowhere near as strong as our belief that P$\neq$NP. The reason for this is connected with a fact that I mentioned earlier, which is that although the problem belongs to NP — if you give

me a factorization, I can check easily whether it is correct — it is not thought to be NP-complete. Let me briefly explain the reason for this.

First, one needs to know that search problems, where one is asked to find an object with certain properties, are closely related to corresponding decision problems, where one is merely asked to determine whether the answer to a certain question is yes or no. In our case, the search problem is to find a factorization of a given large integer n. The corresponding decision problem is to determine whether n has a factor less than some given k.

Clearly if one can solve the search problem, one can solve the decision problem. But the reverse is also true: if one has an efficient algorithm for determining whether n has a factor less than any given k, one can play a game of twenty questions, rapidly narrowing down the range in which there must be a factor until one finds it exactly.

As with the search problem, the decision problem is said to belong to NP if the answer can be demonstrated to be correct in polynomial time. That is clearly the case here: if you want to convince me that n has a factor m that is less than k, you just have to exhibit m and do a routine calculation to show me that n really is a multiple of m (as well as demonstrating that $m < k$, which is even easier).

However, something else happens here that is much rarer: if n does *not* have a factor less than k, you can also convince me of that in polynomial time. The basic idea is that you show me a complete prime factorization of n: that is, you tell me that $n = p_1 p_2 \dots p_r$ for some prime numbers $p_1, \dots, p_r$, and point out that all the p_i are at least as big as k. Of course, that does not convince me unless I am sure that the p_i are indeed prime numbers. It is not at all obvious that there is a quick way of convincing me that a given number is prime, but it turns out that such methods do exist. One method is to use a breakthrough result of Manindra Agrawal, Neeraj Kayal, and Nitin Saxena, who showed that one could determine in polynomial time whether a number n is prime. (Here "polynomial time" means that the time taken is bounded above by a polynomial function of the number of digits of n.) It was previously known that determining primality was in NP — the proof depended on the number-theoretic fact that if p is prime, then there will be a number a such that a^{p-1} is congruent to 1 mod p and no smaller power of a is congruent to 1 mod p.

A decision problem of the form "Does there exist x such that $P(x)$?" is said to belong to co-NP if a negative answer can be checked in polynomial time. So the decision version of the integer factorization problem belongs both to NP and to co-NP, in strong contrast to most natural NP problems and in particular to all known NP-complete problems (at least as far as we can tell — if P=NP then all these distinctions collapse).

Why does this cause us to believe that integer factorization is not NP-complete? Because that would have the extraordinary consequence that NP=co-NP. For example, given an instance of 3-SAT — that is, a collection of clauses of size 3 — we would be able to find a pair of integers n and k such that n has a factor less than k if and only if the collection of clauses can be simultaneously satisfied. Since integer factorization is in co-NP, that would mean that if the collection of clauses *cannot* be simultaneously satisfied, there would be a way of demonstrating that fact, by demonstrating that n does not have a factor less than k. So 3-SAT would also belong to co-NP (as would all other problems in NP by a similar argument).

Thus, problems that belong both to NP and to co-NP are believed to be "easier" than NP-complete problems. This reduces our confidence that integer factorization is hard.

There are three more facts that reduce our confidence still further. One is that although one might think that to factorize a product n of two large primes there is not much option beyond a brute-force search for the smaller of the two primes, which would take time roughly exponential in the number of digits of n, that is not in fact true: there are methods that are far from obvious that use advanced number theory and take a time that is exponential in something closer to the cube root of the number of digits. Though that is still a rapid growth rate, which makes the methods impractical for numbers with more than a few hundred digits, it is still a big improvement on the naive approach, and demonstrates that there can be algorithms that work in clever and unexpected ways. The second fact is that Peter Shor famously showed that a quantum computer *can* factorize large integers in polynomial time. While that is not strong evidence that a classical computer can do the same (in fact, it is normally taken as evidence of the additional power of quantum computers), some of the number-theoretic tricks that are used to prove the result are further demonstrations of the fact that this is a field where there are unexpected ideas to be found. (Another example of this is the astonishing primality

test mentioned earlier.) The third fact is that another problem that famously belongs to NP but does not appear to be NP-complete, the so-called graph isomorphism problem, was shown by László Babai in 2017, by a remarkable and unexpected argument, to be soluble in "quasipolynomial" time — that is, much faster than a typical NP-complete problem.[7] This demonstrates that at least some natural NP problems that seem hard can turn out, for highly non-obvious reasons, to be very much easier than NP-complete problems. This increases our perception of the likelihood that factorizing could be such a problem.

§ 3. — Some general reasons for finding mathematical statements plausible.

A common thread that runs through many of the justifications discussed in the previous section is what I think of as *pseudo-Bayesian* arguments. That is, one starts with some prior belief about the likelihood of a mathematical statement being true, does a suitable experiment, and updates one's beliefs accordingly. What makes this process pseudo-Bayesian rather than genuinely Bayesian is that, as mentioned earlier, the probabilities in question do not take numerical values — rather, they are vaguer notions such as "extremely likely" or "certainly possible", which mathematicians might or might not be willing to convert into very rough estimates such as "with probability at least 70%". (For example, I am ready to stick my neck out and say, very much as a non-expert, that I would give the Riemann hypothesis at least a 90% chance of being true.)

This kind of reasoning is discussed in detail by George Pólya in his 1954 book Mathematics and Plausible Reasoning Volume II: Patterns of Plausible Inference and more recently taken up again by David Corfield in his article Bayesianism in Mathematics, from the book Foundations of Bayesianism, edited by David Corfield and Jon Williamson. One pattern of reasoning Pólya mentions is, for example, the following. Suppose you wish to assess whether a statement A is true, and for the moment your level of confidence in A is medium. Now suppose you spot that A has a consequence B that

[7] A function $f(n)$ is said to be quasipolynomial if it is bounded above by $A\exp(C(\log n)^r)$ for some constants A, C and r. (If $r = 1$ then $A\exp(C(\log n)^r) = An^C$ and f is bounded above by a polynomial.)

would be extremely unlikely to be true "just by chance", and you then discover that B is true. This will greatly increase your confidence in A. Pólya presents this pattern as follows.

A implies B
B very improbable in itself
B true
A very much more credible

Pólya does not set out explicitly *why* A becomes very much more credible (though he provides persuasive examples), but it can be explained quite naturally in a pseudo-Bayesian way as follows. There are two possible explanations for the truth of B. One is that A is true, which happens with probability $\mathbb{P}[A]$. The other is that A is false and that B just happens to be true, which occurs with probability $\mathbb{P}[\neg A]\mathbb{P}[B|\neg A]$ (the second term stands for the conditional probability that B is true given that A is false), which is small. Putting this slightly more formally, Bayes's formula tells us that

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \wedge B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B|A]\mathbb{P}[A]+\mathbb{P}[B|\neg A]\mathbb{P}[\neg A]}$$
$$= \frac{\mathbb{P}[A]}{\mathbb{P}[A]+\mathbb{P}[B|\neg A]\mathbb{P}[\neg A]},$$

where the first three equalities constitute a proof of Bayes's formula and the last equality follows from the fact that B is a consequence of A, which implies that $\mathbb{P}[B|A] = 1$. Because $\mathbb{P}[B|\neg A]$ is small and $\mathbb{P}[A]$ is not particularly small, the bottom of the last fraction is only slightly bigger than the top, which means that the probability $\mathbb{P}[A|B]$, that is, the probability we should now assign to A given the evidence B, is close to 1.

I have still not made any attempt to make sense of the notion of the probability that a mathematical statement is true (or of the related idea that B might "just happen to be true"), but let us postpone thinking about this problem for the moment, and instead look back at some of the arguments in the previous section and try to understand them from a pseudo-Bayesian point of view.

Goldbach's conjecture provides a very good illustration of more than one of Pólya's patterns of plausible inference. A second principle, which is closely related to the first, he presents as follows.

A implies B B quite probable in itself B true
A just a little more credible

It too can be justified using Bayes's formula. Again we have that

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A]}{\mathbb{P}[A] + \mathbb{P}[B|\neg A]\mathbb{P}[\neg A]},$$

but this time $\mathbb{P}[B|\neg A]$, rather than being small, is fairly close to 1, which means that the bottom of the fraction on the right is only a little smaller than $\mathbb{P}[A] + \mathbb{P}[\neg A] = 1$, and therefore the entire fraction is only a little larger than $\mathbb{P}[A]$.

Now let A be Goldbach's conjecture and let B be a statement such as that 10,000,000 can be written as the sum of two primes. We certainly know that A implies B, but in order to apply one of Pólya's patterns of reasoning we need to decide how probable we think that B is "in itself". That is, how likely would it be that 10,000,000 could be written as a sum of two primes if Goldbach's conjecture is not true?

The answer to this question seems to depend quite a lot on how hard one thinks about it. Somebody with no mathematical experience at all might think "It could go either way" or even "It's quite hard to find primes, so it's unlikely that there will be two of them that add up to a large number like 10,000,000 just by chance." Such a person might well have their confidence in Goldbach's conjecture significantly increased by the information that it is true for the number 10,000,000.

But somebody with more mathematical background will realize that the second piece of reasoning above is quite wrong: it may be hard to find large prime numbers, but that does not mean that they do not exist, and in fact the prime number theorem implies that they exist in some abundance. Therefore, for any given large even number it is highly probable that there will be two primes that add up to it "just by chance", and therefore knowing that it is true for just one such number has almost no effect on our confidence in Goldbach's conjecture.

However, in a sense that is true only because our confidence in Goldbach's conjecture has been greatly increased in another way. Let R be the (not completely precise) statement that the primes are distributed in a random-like way. If R is true, then given our knowledge that there are many primes less than n, the probability that

any given large even number can be written as a sum of two odd primes is extremely close to 1. One can even quantify this probability (assuming a suitable random model) and reach the conclusion that the sum over all large m of the probability that $2m$ is not a sum of two primes is very small, and since Goldbach's conjecture has been checked for small m, this means that Goldbach's conjecture is very likely to be true.

What has happened here? We have established that if the random-distribution principle R is true, then almost certainly Goldbach's conjecture is true. But why should that increase our confidence in Goldbach's conjecture itself? On the face of it, since R more or less implies Goldbach's conjecture, it cannot be more likely to hold than Goldbach's conjecture itself, so should not increase our confidence in Goldbach's conjecture.

However, this objection fails: the point is the very simple one that if we observe that a statement A follows from a statement B that is very likely to be true, then it follows that A is very likely to be true, so our confidence in A may well become significantly higher than it was before we made the observation. An extreme example of this is if B is the final observation that completes a proof of A, but even if we do not know for sure that B is true, it may well be that for pseudo-Bayesian reasons we are more confident in B than we previously were in A.

In the case of R, one reason we can be more confident in it than we might have been in Goldbach's conjecture in isolation is that it makes stronger predictions: it implies for instance that there will be many ways of writing 10,000,000 as a sum of two primes and not just at least one. Those predictions are less likely to be true just by chance, so our confidence in R is increased by correspondingly more.

As I mentioned earlier, this is even more clear when we use R to make predictions about approximately how many ways there should be of writing 10,000,000 as a sum of two primes. Predictions of this kind turn out to be remarkably accurate, a fact that would be hard to explain without R, so this is another example of the first of Pólya's patterns mentioned above.

As I also mentioned, the extra generality of R allows one to perform a much wider diversity of tests. For example, one can use it to guess roughly how many primes p less than n there should be such that $p + 2$ is also a prime. Such pairs $(p, p + 2)$ are called *twin primes*, and the famous twin prime conjecture states that there are infinitely many of them. Although the twin-prime conjecture is

still open, the number of twin primes up to n closely matches what R suggests it ought to be, at least for the range of n we have managed to test experimentally. This provides what one might think of as an independent test of R, since instead of counting solutions to equations of the form $p+q=2m$ (where p and q are required to be prime) we are counting solutions to the equation $p+2=q$ in the range $q\leqslant n$ for varying n. There are many such tests, and R seems to pass all of them (with one or two subtle exceptions that are well enough understood that they do not shake our confidence in a suitably adjusted principle).

This is an important observation. As we shall see, for several of the examples considered earlier, our confidence in a statement increases if that statement is shown to follow from a more general statement that we have been unable to disprove. Since there are many more potential ways of disproving a more general statement, we are more confident in the general statement than we were in its consequence when the consequence was considered in isolation.

Here is a list of general statements, not all of them precise, that played this kind of role in the discussions of specific problems earlier.

1. The normality of π follows from a more general principle that biases in decimal expansions do not occur without a reason, and that the reason is usually fairly simple. This more general principle is confirmed by many other examples: for instance, there is no discernible pattern to the decimal expansions of $\sqrt{2}$, e, π^2, and so on.

2. The statement that $R(k,k)^{1/k}$ converges to a limit follows from a general principle that the behaviour of large graphs is not very sensitive to the number of vertices. So if, for instance, $R(100,100)$ is close to 2^{100} and $R(150,150)$ is close to 2^{150}, then we would expect $R(200,200)$ to be close to 2^{200}. Again, there are many examples, and in some contexts even theoretical arguments, that support this general principle.

3. The statement that the average end-to-end distance of an n-step self-avoiding walk is roughly $n^{3/4}$ follows from the more general principle that the heuristic arguments of physicists are a good guide to the truth, even if they are not rigorous proofs. Once again, this general principle is confirmed by many other examples (though there have been occasional cases where pure mathematicians have disproved statements that physicists had confidently asserted).

4. Like Goldbach's conjecture, the Riemann hypothesis can be justified by the more general principle that the primes behave as a natural random model would predict. (Recall Bombieri's statement that the failure of the Riemann hypothesis would create havoc in the distribution of prime numbers.) It also follows from the generalized Riemann hypothesis, which is a precise mathematical statement. It also follows from the even more general, but less precise, statement that there is a large family of zeta functions that all satisfy appropriate versions of the Riemann hypothesis. (The lack of precision is that, as far as I know, there is not a clear conjecture about what the family is, though there are many important zeta functions that would belong to it.)

A second reason for increased confidence in a conjecture is if it can be shown to have an "otherwise unlikely" consequence that we can prove, or at least confirm experimentally. Often the reason the consequence is unlikely is simply that it is very precise. (Of course, we see this in science too, such as Eddington's famous measurement of the gravitational deflection of starlight, which provided a dramatic confirmation of Einstein's theory of general relativity.) Some examples of this related to the problems that we saw earlier are the following.

1. The general principle that the primes are distributed as if they are random makes precise predictions about the approximate number of ways of writing a large even number as the sum of two primes.

2. The arguments of statistical physicists concerning self-avoiding walks predict not just the average end-to-end distance but also a great deal more detailed information about what a typical self-avoiding walk looks like, all of which is confirmed by experiment.

3. A consequence of P not equalling NP is that even though it is easy to find an efficient algorithm that solves 7/8 of the clauses in an instance of 3-SAT, there is no such algorithm that will solve a very slightly larger fraction of the clauses. And indeed, no such algorithm has been found, and there is no serious hope that one will be found.

4. It is conjectured that the zeros of the Riemann zeta function do not just all have real part equal to 1/2 but that they are

> distributed in a way that closely resembles the distribution of eigenvalues of random Hermitian matrices (chosen in a natural way). This basic idea has led to many remarkably precise conjectures about the behaviour of the zeros that agree impressively with experimental evidence.

It is also worth reflecting on reasons that our confidence may be decreased. One example we have seen was Odlyzko's point that the Riemann zeta function "reaches its asymptotic behaviour slowly". Although this does not alter the fact that the experimental evidence concerning the zeros agrees extremely well with precise theoretical predictions, it raises the possibility that there might be some phenomenon that shows up only for zeros with very large imaginary parts, which would make that phenomenon difficult to detect experimentally.

Another example was the probable difficulty of integer factorization, where there were two main reasons that mathematicians are inclined to be a little cautious in their assessment. One was that there are known algorithms that perform related tasks efficiently in very unexpected ways, so the mere fact that there is no *obvious* way of factorizing efficiently is not completely compelling evidence: it is at least conceivable that there is a non-obvious way of doing it. The other was that factorization is a central example of a problem that is in NP but that is not believed to be NP-complete, and another prominent example of such a problem was recently shown to be soluble in quasi-polynomial time.

The second of these reasons is an example of a further general principle that mathematicians use when making probabilistic judgments about unproved statements, which is that *similar statements may well behave similarly*. We often make guesses about problems that are based on experience with similar problems, though obviously there is work to do in deciding what "similar" means here, as some similarities will be more relevant than others. The integer factorization problem and the graph isomorphism problem are superficially quite different: the relevant similarity is that they both belong to NP and to co-NP, which is a rather deep phenomenon.

Once again, Pólya has identified this pattern of reasoning. He presents it as follows.

A analogous to B B true
A more credible

He tentatively justifies it by saying that because the statements are analogous, one can imagine that they are both consequences of some more general statement H, and, as we have seen, generalizing a statement often makes it more plausible. Of course, if a third statement C analogous to A and B turns out to be false, this will reduce our confidence again, unless on closer inspection we find that C has some property that might lead us to expect it to behave differently, in which case we have the option of replacing the generalization H by a more specific generalization H' that includes A and B and rules out C.

§ 4. — What do probabilistic judgments about mathematical statements even mean?

The pseudo-Bayesian account cannot fully answer the question of why mathematicians have differing levels of confidence in different unproved statements, because Bayes's formula contains probabilities on both sides of the equals sign. So while it may give a good account of how we update our beliefs given our priors and our assessments of certain conditional probabilities, it does not explain how we arrive at the priors and conditional probabilities in the first place (except when they themselves are updated beliefs, but that just pushes the problem further back). At some point we need to grapple with the question of what is going on when we use probabilistic language in this very deterministic-seeming context.

At first glance it seems to make no sense at all to say of a specific (decidable) mathematical statement S that S is "probably true" since it is either definitely true or definitely false. It is not as though we repeatedly try to prove it, usually succeeding but occasionally finding a counterexample.

However, this objection is not very strong, as easy examples from outside mathematics show. If you toss a coin and keep it covered with your hand, then it is either definitely heads or definitely tails, but it still makes perfect sense for me to say that the probability that it is heads is $1/2$. Indeed, that makes sense even if you have surreptitiously looked at the coin to see how it landed.

The judgment in the case of the coin can be justified with a very basic frequentist interpretation of probability: all one has to note is that the event is one that could in principle be repeated many times, and one would expect the coin to come up heads half the time and

tails half the time. So even though the coin has now been tossed, before it was tossed it was reasonable to assign a probability of 1/2 to its coming up heads, and since I do not yet know the result it is still reasonable.

If we try to use reasoning of this kind for our judgment about S, we run into a problem, which is that S is not obviously an instance of a repeated event that has varying outcomes: it is just a fixed mathematical statement. However, one should not give up too quickly, as there are certainly *some* probabilistic judgments about mathematical statements that are similar to our judgment about the coin toss. For instance, suppose I choose a random integer n between 1,000,000 and 1,100,000. The prime number theorem tells me that about 7% of the numbers in that interval are prime, so I have no difficulty in saying that the probability that n is prime is about 7%.

This might be regarded as a cheat, however, since I chose n randomly, and if I put randomness in, then it is no surprise if I can get randomness out. So let us instead make it deterministic. I shall define a number n as follows. The number 2^{1200} has around 400 digits. Take the sequence of five consecutive digits of 2^{1200} that start at the 201st digit, add the corresponding five-digit number (which might start with some zeros) to 1,000,000, and let n be the result of this calculation. I would like to argue that it is reasonable to say that the probability that n is prime is about 7% — until, that is, I go ahead and check, which I shall now do. The sequence of five digits starting at the 201st is 06441 (if I didn't miscount), so n =1,006,441, which turns out to be prime, which is quite surprising given that it there was only a 7% chance of this happening.

But what does it mean to say that n had only a 7% chance of being prime? It is definitely prime, so does that not mean that it had a 100% chance of being prime? In this case, the answer seems to be something like the following. Although n is a single fixed number, the way it was chosen was rather arbitrary. In fact, what I did was to devise a simple *pseudorandom generator* — that is, a procedure that was likely to produce for me a sequence of five digits with no discernible pattern. I could just as easily have started the sequence at the 130th digit, or taken the digits in places 100, 200, 300, 400, 500 from the decimal expansion of $\sqrt{2}$, or used any number of other deterministic sequences defined in ways that would lead me to expect that they will not have any properties that would affect the propensity of n to be prime. (I could also have defined it in ways that *would* have had such an effect: for example, had I

taken the last five digits of 2^{1200} then n would have been guaranteed to be even, or if I had taken five consecutive digits from the decimal expansion of 1/7, then there are only six numbers I could have obtained, none of which, it turns out, yield a prime when added to 1,000,000.)

Thus, my judgment that n had a 7% chance of being prime is based on regarding the deterministic choice of n as just one example of a whole class of choices — roughly speaking, numbers obtained from five-digit sequences that experience suggests will be typical. And the role of the calculation I did (with the help of Wolfram Alpha[8]) was very similar to the role you play when you lift your hand and show me the coin. It may be that the fact that n is composite follows from axioms and rules of deduction that I fully accept, but that does not imply that I know that n is composite: to gain such knowledge I need to do the calculation, which takes effort. Moreover, this last point can be made quite formal using the theory of computational complexity, which actually studies the amount of effort needed to determine whether certain mathematical statements are true.

The basic idea here is that a mathematical statement such as "This n is prime" can often be seen as a representative of a much wider class of mathematical statements, and once one has a big class of statements it makes perfect sense to ask what percentage of them are true. But while that kind of analysis shows that the notion of probability can sometimes make good sense, can we apply it to statements such as Goldbach's conjecture, the Riemann hypothesis, or the normality of π?

I believe that it is indeed possible, and that some of our reasons for judging these statements to be likely or unlikely suggest that we do indeed regard them as representative of larger classes of statements.

This is quite easy to see in the case of the normality of π. Let us think about how a typical mathematician's understanding of this problem might develop. As a child, one learns about decimals, and after a while one learns to calculate the decimal expansions of fractions such as 1/2, 1/5 and 1/10, followed by more interesting examples such as 1/3, 1/9 and 1/6, followed by more interesting examples still such as 1/7. In all these cases the decimal expansion either terminates or eventually recurs, so one might come to the

[8] Wolfram Alpha is a website that computes answers to questions from several different domains, including mathematics.

view that non-whole numbers typically have terminating or recurring decimal expansions, which would lead one to expect, wrongly as it turns out, that π has such an expansion and is therefore very far from being normal. Note that π is being seen here not just as an individual number but as an example taken from the set of all numbers. One might realize on reflection that it is certainly not true that all numbers have recurring expansions, since one can just write down an infinite decimal expansion and make sure it doesn't recur. But one could still take the attitude that such numbers were artificial, whereas π arises naturally and is therefore likely to have a more regular decimal expansion.

A little later one's perspective shifts completely: one observes that numbers that arise naturally tend to have very *irregular* decimal expansions unless they are rational numbers. That is, terminating or recurring decimal expansions are very much the exception rather than the rule.[(9)] At that point, why is it natural to expect that π is a normal number? An obvious answer that our budding mathematician might give is, "We know the decimal expansion of π to many millions of decimal places: if there were some kind of bias, one would expect it to have shown up by now." This can be seen as appealing to the following general principle: *if the decimal expansion of an irrational number that arises naturally or is built in a simple way out of numbers that arise naturally does not show any sign of bias in the first million digits, then probably it will never show any sign of bias*. This principle allows for the possibility that there may be exceptions, but there are many numbers to choose from, so π probably isn't an exception.

Suppose we pursue the question a little further, raising the possibility that since π is a very special number, it might conceivably have a property that leads to an unexpected but very small bias in its decimal expansion. Why does this seem so unlikely? Again it is natural to appeal to more general principles: it is hard to imagine what a proof could possibly look like — for one thing, decimal expansions are closely tied to the number 10, which seems to have nothing to do with π — and we know of no such argument for *any* irrational number built in a simple way out of naturally occurring irrational numbers.

The absence of the merest hint of a technique that might be useful for showing that π is not normal is not in itself evidence that π is

[(9)] Of course, for this to happen one must first become familiar with facts such as that π is irrational and that there are a lot of irrational numbers.

normal, but it becomes convincing thanks to an even more general principle that is confirmed by a great deal of mathematical experience.

No-coincidence principle. *If an apparently outrageous coincidence happens in mathematics, then there is a reason for it.*

We have arrived at a possible interpretation of the statement "π is almost certainly normal". Suppose we accept the no-coincidence principle on the grounds that it is amply backed up empirically. From it we can deduce that if there is no good reason for π not to be normal, then π is normal. So now our attention turns to the probability that there is a good reason for π not to be normal. Experience with good reasons suggests that they usually, though not always, kick in reasonably soon, so given that there is no discernible pattern in the digits of π that have so far been calculated, of which there are a huge number, any reason for the non-normality of π would have to be very strange and unusual. So now we wish to assess the probability that there is a good, but strange and unusual, reason for π not being normal. At this point we can perhaps appeal to a more precise variant of no-coincidence principle, namely that, on average at least, the greater the coincidence, the easier it is to find an explanation for it.[(10)] If, for example, there were only finitely many 7s in the decimal expansion of π, that would *a priori* be such an extraordinarily improbable event that we would expect to be able to uncover the reason for it. The fact that we have not been able to find any such argument, for π or for any other number, is therefore quite strong evidence that no such argument exists. Thus, the interpretation of what it means to say that π is almost certainly normal is that while unexpected reasons do exist and therefore strange things do happen, the chances that any such reason applies to any given statement (of a kind that asserts that there is no departure from randomness, and where this is backed up by experimental evidence and a complete failure to find anything resembling a plausible reason that there would be such a departure) are small.

[(10)] I first heard something like this principle from Don Zagier, who argued that many problems in number theory are hard precisely because they are asking you to prove that something happens that is *exactly what you would expect*. If a phenomenon comes into this category, it is not clear that it demands an explanation, which raises the likelihood that there is no easy proof of it. Goldbach's conjecture is a case in point — as already discussed, it would be a big surprise if some large even integer were *not* a sum of two primes.

The example of $e^{\pi\sqrt{163}}$ discussed earlier can also be analysed in this way. To somebody who does not know the relevant number theory and has not heard that $e^{\pi\sqrt{163}}$ is extremely close to an integer, it is just a fairly arbitrary number one can build, not interestingly different from $\pi^{\sqrt{2}}$ or e^{π^2} or $e^{\pi+\sqrt{163}}$. And if you build numbers that are not obviously rational out of numbers like π, e and square roots of positive integers, then they will tend to have fractional parts that are pretty uniformly distributed in the interval $[0,1)$, so although some of them will be fairly close to integers just by chance, the probability of stumbling on one that just happens to be within a million millionth of an integer will be astronomically small.

One is therefore justified in being very surprised that $e^{\pi\sqrt{163}}$ is indeed within a million millionth of an integer: again the point is that we are seeing $e^{\pi\sqrt{163}}$ as a member of a more general class. Of course, that does not mean that an event with a probability of order of magnitude 10^{-12} has just occurred, because we should at least entertain the possibility that some of the numbers in the class are close to integers for special reasons. For example, this is true of the number $\phi^{100}/\sqrt{5}$, where ϕ is the golden ratio $\frac{\sqrt{5}+1}{2}$. The 100th Fibonacci number is given by the formula $\frac{\phi^{100}-(-\phi)^{-100}}{\sqrt{5}}$. Since $\phi>1$, ϕ^{-100} is tiny, so $\phi^{100}/\sqrt{5}$ is extremely close to the 100th Fibonacci number, which is an integer. Thus, some simple expressions do turn out, for good reasons, to define irrational numbers that are very close to integers, and this affects one's judgment of the probability that it will happen in any particular case. That contrasts with the case of the normality of π, where we completely lack arguments that might be used to prove non-normality, and therefore feel very confident that π is normal. Note also that additional mathematical knowledge can affect one's judgment of the likelihood of an argument of a certain kind existing. If, for instance, one is told that 163 is the largest positive integer k such that the field $\mathbb{Q}(\sqrt{-k})$ has class number 1, then one has the information that the number 163 has a very special number-theoretic property. This could be completely irrelevant, but it at least increases the probability that an argument exists that would show that $e^{\pi\sqrt{163}}$ is very close to an integer. And as it happens, it turns out to be highly relevant.

We can interpret the assertion that Goldbach's conjecture is almost certainly true in a similar way, but only after a certain amount of mathematical thought, because while the failure of some

very large even integer n to be the sum of two primes would be a remarkably improbable event if we assume that the primes are "distributed randomly", the primes are in fact created in a rather systematic way — they are what is left over when we remove from the positive integers the number 1, then all multiples of 2 (not including 2 itself), then all multiples of 3, then all multiples of 4 (though they have already been removed), and so on. The sets we are removing have a lot of structure — they are arithmetic progressions — and it is not obvious that we can't exploit that structure to show that the set that is left over has interesting structure as well, over and above simple facts such as that only one prime is even. However, we now have a great deal of experimental evidence that the primes are indeed about as random as they can be, and we also have enough theoretical understanding of this quasirandomness to have been able to prove that every odd integer from 9 onwards is a sum of three odd primes, as well as several other results in a similar vein. This provides strong evidence that there is unlikely to exist a surprising reason for some large even integer *not* to be a sum of two primes, evidence that is all the more compelling given the tables that show a very close agreement between how many ways there are of writing large even numbers as a sum of two primes, and how many ways we would expect there to be.

Again, thanks to the no-coincidence principle, the focus has shifted from the probability that Goldbach's conjecture is true to the probability of the existence of a certain kind of argument. That we judge to be unlikely for various reasons: not only have we failed, after centuries of thinking about the primes, to find any arguments that would lead us to expect departure from randomness (if "randomness" is suitably interpreted), but the fact that we have also not observed any such departure empirically, either with Goldbach's conjecture or with a large number of other conjectures that assert that the primes behave roughly as one would expect them to, makes it unlikely that any such arguments exist, since if they did exist then we would probably have observed their consequences.

This kind of analysis can be used to understand both our confidence in the Riemann hypothesis, such as it is, and the fact that that confidence is not total. For there to be a counterexample to the Riemann hypothesis, there would have to be a very unexpected bias in the distribution of the prime numbers, so there would need to be a reason, and that reason would have to be somewhat strange, since the bias has not shown up despite a great deal of computational

evidence being collected. However, as one of the main collectors of that computational evidence has pointed out, while it is difficult to imagine an explanation for the existence of a counterexample with what would have to be a very large imaginary part, it is not impossible, since there are reasons to believe that the typical behaviour of the Riemann zeta function manifests itself extremely slowly. That reduces our confidence that no argument of the required kind exists.

When it comes to the problem of whether P=NP, the no-coincidence principle does not at first seem to help in quite the same way. What might a reason for P not equalling NP be like? An obvious answer is that it might be an efficient algorithm for solving an NP-complete problem. But such an algorithm does not seem as though it would be an explanation for an outrageous coincidence: it would just be a straightforward counterexample to the assertion that no such algorithm exists.

Nevertheless, a slight modification of the no-coincidence principle does apply here reasonably well. Consider again the NP problem where one is presented with a number n (which one should think of as having a large number of digits), told that it is the product of two primes, and asked to work out what those two primes are. As mentioned before, it is relatively easy to program a computer to check whether a proposed answer is correct: if somebody suggests the two numbers p and q, then all one has to do is multiply p and q together using a standard procedure such as long multiplication. If P=NP, then this fact is enough to guarantee that one can actually *find* p and q, which on the face of it seems unlikely. (If you think it is not unlikely, see if you can find the two prime factors of 35873023. In case it helps, I obtained the numbers in question from a Wikipedia page entitled "List of prime numbers".)

But why should ease of checking have anything to do with ease of finding? If I'm looking for my telephone, I'll recognise it very easily when I see it, but that doesn't mean it is easy to find. Similarly, my ability to do long multiplication seems to have almost nothing to do with factorizing.[(11)] So although the statement that P=NP is not obviously a claim that an outrageous coincidence occurs, it is still claiming something pretty outrageous that would demand an explanation. Therefore, we can maintain the spirit of the no-coincidence principle by generalizing it slightly as follows.

[(11)] As I mentioned earlier, it is not impossible that there is some extremely clever method of factorizing efficiently. But if P=NP it would be possible to factorize efficiently by virtue of the fact that we can multiply efficiently — that is what seems unlikely.

No-miracle principle. *If an apparent miracle happens in mathematics, then there is a reason for it.*

The difference between this and the no-coincidence principle is that we allow miracles that do not have natural probabilistic descriptions: it would be miraculous if the ability to carry out long multiplication guaranteed the ability to factorize efficiently, but the word "coincidence" does not seem appropriate for that particular miracle.

That said, with a little more effort it is possible to use the no-coincidence principle for the P versus NP problem as well. Indeed, one of Scott Aaronson's arguments has that flavour: if P did equal NP then it would be quite a coincidence that we find it easy to satisfy 87.5% of the clauses in a 3-SAT instance but extremely hard to satisfy 87.6% of them. There is also a concept of *pseudorandom generators*, which are deterministic methods of generating sequences that appear to all intents and purposes to be random. (Earlier in this article I defined a baby pseudorandom generator when I wanted to create a random-looking five-digit sequence and took a string of digits from the middle of the decimal expansion of 2^{100}.) There are several such methods, and the evidence is that they work, in the sense that they have no features that would allow one to detect efficiently how they were created or make any guesses about how they will continue. However, if somebody explains how they generated a pseudorandom sequence, it is easy to check that their method does indeed generate the given sequence, so if P=NP then there is some outrageous coincidence hidden in the behaviour of all these apparently random sequences that allows one to efficiently work out how they were generated (or at least how each one could have been generated).

Whichever version of the principle we decide to use, the conclusion is the same: for P to equal NP something outrageous would have to happen; that would demand an explanation; and the evidence leads us to believe with very high confidence that no such explanation exists. (Our confidence is perhaps not total, because over the years some very surprising efficient algorithms have been discovered, but there are additional reasons to believe that an algorithm for efficiently solving an NP-complete problem would be a step too far. That said, the existence of surprising algorithms is commonly taken as a reason to expect it to be very hard to *prove* that P does not equal NP.) So once again the focus shifts from the statement itself to the existence of a proof of a certain kind.

Let me try to summarize the proposal I am making for how to interpret probabilistic judgments concerning mathematical statements. Perhaps the most important point is the suggestion that they are not really probabilistic judgments concerning the *truth values* of statements, but about the existence of *arguments* (not necessarily rigorous) in support either of those statements or of their negations. In particular, many probabilistic judgments appear to have justifications of the following general kind.

1. If statement S were false, a miracle would have to occur, and there would have to be an explanation for why it occurs.
2. There is probably no explanation of the kind that would be needed to explain such a miracle.

However, so far this replaces one probabilistic judgment of a deterministic event by another. What does it mean to say that there is probably no explanation of a certain kind, when either there is one or there isn't?

Here the proposal would be to interpret the probabilistic judgment as standing for a statement of the form that statements of a certain kind tend not to be backed up by arguments. That is, we look at the miracle not in isolation but as representative of a certain class of miracles. Our judgment is then a genuine probabilistic statement: that a random miracle chosen from that class has only a small chance of being explainable (and therefore only a small chance of actually occurring).

We have seen an example of this: somebody who did not know about $e^{\pi\sqrt{163}}$ would argue that it would take a miracle for such a number to be within 10^{-12} of an integer, and therefore that it could not be so close to an integer without some very good reason, and that such reasons, though they do exist, are few and far between. Therefore, $e^{\pi\sqrt{163}}$ is very unlikely to be within 10^{-12} of an integer. That would be a perfectly reasonable judgment, even though it happens to be mistaken.

Note that the more one knows about a problem, the more features of any given miracle one will be aware of, which affects what class of miracles one regards the particular miracle as belonging to. For instance, if you are asked whether $(\sqrt{2}+1)^{30}$ is likely to be very close to an integer after reading what I wrote earlier about the golden ratio, you will probably regard $(\sqrt{2}+1)^{30}$ not just as some random strange number that can be built out of well known irrational numbers, but as a power of a root of a fairly simple quadratic equation, which might well have a reason to be close to an integer — as indeed it turns out to be.

§ 5. — Goodman-type paradoxes.

Our discussion has taken us quite a long way from *experimental* evidence, so let us briefly consider what part it has to play in the above account. I have suggested that what is meant by the judgment that S is probably true is that the miracle that would be required for S to be false belongs to a class of miracles of which only very few actually occur, and that is because miracles do not occur without explanations, and explanations of the kind that would be needed to explain miracles in the given class are very rare.

If that is a correct interpretation, what is the role of experiment in increasing our confidence in a mathematical statement? More generally, how does that interpretation fit with the pseudo-Bayesian account of how we update our perceptions of probability in the light of new evidence?

The answer to the second question is quite easy to give in connection with specific examples. For instance, let us consider again the fact that our confidence in Goldbach's conjecture increases somewhat when we find that it holds for all even integers up to 10^{10} (say), quite a lot more when we find that the number of ways of writing an even integer n up to 10^{10} as a sum of two primes is not just non-zero but large whenever n is large, and even more when we find that the number of ways of doing it is not just large but a good approximation to what a natural random model of the primes would predict.

The fact that Goldbach's conjecture is true up to 10^{10} means that if it is false, then it has to be true up to some large integer and then false. That would be quite strange, so it counts as a miracle that would need an explanation. However, there are quite a lot of explanations around that have the right form: that is, they explain why some statement of the form $\forall n\, P(n)$ is false but also explain why the smallest counterexample is very large. That is not to say that such explanations are commonplace, but it says that in the absence of any further considerations, the miracle would not be all *that* miraculous, so we do not want to assign too small a probability to Goldbach's conjecture being false.

If we now find that the large even integers we look at can all be written in many ways as a sum of two primes, then we start to get the impression that writing an even integer as a sum of two primes tends to be not just possible but actually quite easy. It now becomes

significantly harder to imagine what an explanation for the conjecture being false could look like. There would presumably have to be something special about some particular very large even integer n that caused a "conspiracy" between the primes — a deal that each time p is prime, $n - p$ will make sure it has a non-trivial factor. There is no hint of any such conspiracy occurring, or even getting close to occurring, for any of the even integers we have looked at so far, and it is very difficult to think of a property of an extremely large n that would not hold for any smaller n and would explain why the primes suddenly aligned themselves for that n.

Finally, if we go on to observe that the number of ways of writing a large integer n as a sum of two primes is well predicted by a random model, then what is demanded of a putative explanation for the failure of Goldbach's conjecture is even stronger: we would need to know why the number of ways of doing it behaves exactly as a random model would predict for all the even integers we have looked at, but then for some very large even integer not only fails to behave as predicted but fails in a very radical way. Since the random model makes good predictions not just about Goldbach's conjecture but about many other statements to do with the primes, the evidence strongly suggests that no such explanations exist (since if they existed, we would expect to have observed their consequences by now). This contrasts with the situation we were in when all we knew was that Goldbach's conjecture itself was true up to 10^{10}, when we were considering a different class of explanations — broadly speaking, explanations of why some natural property is true for all positive integers up to some very large n but not true for all positive integers — of which there are some notable examples.

Let us now consider how a famous paradox of Goodman plays out in the context of computational evidence for mathematical statements. Goodman pointed out that the traditional problem of induction — why should we believe that a certain statement will continue to be true just because it has been observed to be true up to now? — is not adequately formulated, because there are many statements for which we do not have such a belief. To make his point, he defined a predicate "grue", which means "green up to time t and blue thereafter". If we set t to be the beginning of 2050, say, then all experimental confirmation of the statement "all emeralds are green" also confirms the statement "all emeralds are grue". And yet we do not believe the latter statement — we believe that at

the beginning of 2050, emeralds will remain green. It is clear that the reason for this has something to do with the fact that "grue" is a very unnatural predicate, but it is quite hard to specify which predicates should count as natural.

Here is a statement that relates in a Goodman-like way to Goldbach's conjecture.

Conjecture. *No power of* $2^{1,000,000,000}$ *is a sum of two primes, but all other even integers greater than 2 are sums of two primes.*

All the computational evidence we have obtained is consistent with this modified conjecture — indeed, the largest known prime is far smaller than $2^{1,000,000,000}$. So why do we not regard the evidence as supporting the modified conjecture?

If one uses the framework suggested in the previous section, then the answer is surprisingly easy: for the modified conjecture to be false, a miracle would not have to occur. Indeed, all that would be needed for the modified conjecture to be false is for the primes to behave exactly as expected, since that would guarantee that $2^{1,000,000,000}$ could be written in many ways as the sum of two primes. In the case of Goldbach's conjecture itself, a miracle needs to occur for it to be false, and the experimental evidence, which has failed to find even a hint of such a miracle, strongly supports the hypothesis that there is no reason for a miracle to occur, which in turn (thanks to the no-miracle principle) strongly supports the hypothesis that no miracle in fact occurs. For the modified conjecture, a miracle doesn't need to occur, so the experimental evidence no longer plays this role.

A very interesting contribution to the large literature about Goodman's paradox was made by Rosemarie Rheinwald in 1993. She points out that one's starting beliefs, or what she calls our *epistemic situation*, have a critical effect on whether a predicate seems natural or not. She illustrates this with the following beautiful example. Apparently, there are two kinds of hares, field hares and Alpine hares. Field hares are always brown. As for Alpine hares, they themselves come in two kinds: one kind is always white, while the other kind is brown in the summer and white in the winter. She invents a Goodman-like predicate "su-wi-brote" that allows one to express this situation by saying, "All hares are su-wi-brote". She then argues that "su-wi-brote" is a *projectible* predicate — this means that it is suitable for the purposes of inductive inference — but only to somebody with the right set of background beliefs. To

someone like me before I read her article, the fact that all hares I had observed up to now (not just in real life but in books, photographs, etc.) were su-wi-brote would not count as evidence in favour of the hypothesis that all hares are su-wi-brote, since they had also been brown and I had not heard of Alpine hares. But if somebody who *had* heard of Alpine hares and knew a bit about them were to obtain appropriate confirming instances of the hypothesis "All hares are su-wi-brote" — and for the confirmation to be appropriate it would of course be necessary to look at field hares and Alpine hares, and to look at the latter both in summer and winter — then their confidence in the statement "All hares are su-wi-brote" would be increased.

We have already seen examples of how background beliefs have an important effect on one's perception of how likely mathematical statements are to be true. They can also have an effect on which predicates are projectible, as the following example, which is well known to mathematicians, illustrates nicely.

The example concerns the following question. Suppose you draw n points round the circumference of a circle, and you join each pair of points with a line segment. Assuming that the points are in general position (in particular, no three of the lines meet at a point that isn't one of the points on the circumference), how many regions is the circle divided into by the lines?

If one draws one point, then there are no lines and the circle is "divided" into one region. If one draws two points, then there is one line, which divides the circle into two regions. With three points, the lines form a triangle, and there are therefore four regions — three outside the triangle and one inside. With four points, the lines form a quadrilateral with its two diagonals, making eight regions — four outside the quadrilateral and four inside. With five points there are sixteen regions, as you will readily see if you draw a picture. (There is a pentagon with five regions outside it and a five-pointed star inside. Inside the pentagon but outside the star there are five triangular regions. The star itself consists of an inner pentagonal region and five triangles attached to it. This makes 5+5+5+1=16 regions.)

Writing down the numbers we have obtained so far, we obtain the sequence 1, 2, 4, 8, 16. It would seem that the only natural hypothesis we can form is that the number of regions doubles each time one adds a point, so that with n points there are 2^{n-1} regions.

However, as is often pointed out, *any* continuation of *any* short sequence can be justified with the help of the fact that if d terms $a_1, \dots, a_d$ of a sequence are given, then there is a polynomial P of degree at most $d-1$ that $P(1) = a_1, P(2) = a_2, \dots, P(d) = a_d$. For example, suppose I wish to create a cubic polynomial P such that $P(1) = 1, P(2) = 1, P(3) = 2$ and $P(4) = 3$. I first note that the polynomial $(x-2)(x-3)(x-4)$ is a cubic polynomial that takes the value -6 at 1 and 0 at 2, 3 and 4, and in a similar way I can create cubic polynomials that vanish at any three of the numbers $1, 2, 3, 4$ and not the fourth. Then a suitable combination of these cubic polynomials can be used to give me one that takes the values I wish. In this case, we will end up with the polynomial

$$P(x) = -\frac{(x-2)(x-3)(x-4)}{6} + \frac{(x-1)(x-3)(x-4)}{2} - \frac{(x-1)(x-2)(x-4)}{2} + \frac{(x-1)(x-2)(x-3)}{6},$$

the origin of which I could disguise if I simplified it by multiplying out all the brackets and collecting terms.

Returning to the sequence 1, 2, 4, 8, 16, we therefore see two things: first, given any continuation of this sequence, there is a polynomial of degree at most 5 that will give that continuation, and secondly, there is a polynomial of degree at most 4 that agrees with the sequence so far. In the light of that, should we revise our assessment that 32 is the obvious continuation of the sequence?

If one is given the sequence 1, 2, 4, 8, 16 in isolation and asked to continue it, then 32 is unquestionably the most natural answer: it is unusual for each term of a sequence to be obtained from the previous one by a process as simple as doubling, and this has now happened four times. (I suppose I must acknowledge that it is also the case that we have *troubled* the previous number four times, where that means doubling it if the number is at most 10 and trebling it if it is greater than 10, but troubling is quite clearly a more complicated operation to define than doubling.) By contrast, it is not at all unusual for a sequence of five numbers to be given by the values of a quartic polynomial, since that is true of *all* sequences of five numbers.

Interestingly, one can say the same if one has a little more information about where the sequence comes from. Suppose one knows that it comes from a parameter associated with a natural sequence of combinatorial structures. Experience shows that something like the following principle is quite reliable.

Nice-formula principle. *Given a sequence that arises naturally in mathematics, either it is given by a nice formula, or there is no hope of expressing it by any exact formula but it can at least be approximated by a nice formula.*[12]

I would not wish to claim that the above principle holds universally — indeed, I can think of examples of combinatorial problems that give rise to bizarre formulae — but exceptions to it appear to be quite rare. (Note that this probabilistic statement is quite straightforward to interpret: it is simply saying that the proportion of naturally occurring sequences that are exceptions to the rule is small.) Nevertheless, if one knows that a sequence that begins 1,2,4,8,16 has arisen naturally in a mathematical context, there is a very good chance that it is the sequence of powers of 2.[13]

Now let us change our mathematical epistemic situation by reflecting a little more on how the number of regions the circle is divided into behaves as the number of points increases. Each region contains on its boundary either one of the n points on the circle or a point where two of the connecting lines intersect. Each of the n points on the circle belongs to n regions, and since each intersection of two of the lines can be specified by four of the points (the end points of the two lines in question), the number of intersections of two of the lines is at most $\binom{n}{4}$. Moreover, each such intersection belongs to four regions. Therefore, the number of regions is certainly less than $n^2 + 4\binom{n}{4}$, which is easily checked to be at most n^4 for all positive integers n.

This shows that the formula 2^{n-1} cannot be correct — it grows exponentially quickly, which, as any mathematician knows, is a far quicker rate of growth than any polynomial, and in particular far quicker than n^4. Indeed, we can say more: if there is a polynomial formula for the number of regions, then that polynomial will have to be of degree 4 at most, since polynomials of higher degree

[12] An example of the second situation is the sequence of primes: there is no nice formula for the nth prime, and almost certainly no sensible formula at all, but it follows from the prime number theorem that the nth prime is approximately equal to $n \log n$. For the first situation, I am taking the word "formula" in a fairly broad sense, and would include sequences given by simple recurrence relations or as the coefficients of the power series of some nice function, for instance.

[13] I implicitly appealed to the nice-formula principle when discussing the Riemann hypothesis earlier. If one forms the sequence of real parts of the zeros of the zeta function, in increasing order of their imaginary parts, then it must equal $1/2$ for all the many terms we know so far, and must equal $1/2$ at least 40% of the time. The nice-formula principle then tells us that it should be a constant sequence.

grow faster than polynomials of lower degree and we know that the growth rate is not faster than that of n^4.[14]

With this new perspective, we find ourselves asking the following question: what is the nicest formula that would yield the values 1,2,4,8,16 but grow at most as fast as a quartic polynomial? And the answer is probably that one should actually take a polynomial. There is only one that can work, and if it gives the correct formula, then the number of regions we obtain with six points round the circle should be 31. And that turns out to be the case.[15]

The main point I am making here is that the hypothesis that the sequence 1,2,4,8,16 is given by the unique quartic polynomial that takes those values is a bit like the hypothesis that all hares are suwi-brote: at first glance it seems very unnatural, but if one knows a bit more, then that perception changes substantially.

This is also another nice example to test out the analysis of probabilistic judgments. It now seems highly probable that the quartic polynomial is indeed the correct formula for the number of regions, but why?

Before we did the experiment of drawing a sixth point and finding that the number of regions was 31, our two likely hypotheses, given the nice-formula principle, were these.

1. The number of regions is given by a nice formula.
2. The number of regions is not given by a nice formula but can be approximated by a nice formula.

The fact that the obvious nice formula 2^{n-1} was doomed to fail meant that the nicest formula by some way was given by the one quartic polynomial that agrees with the sequence 1,2,4,8,16. Therefore, if the first hypothesis is true, then probably the number of regions is given by this formula. If the second hypothesis is true, then there is some formula that will be a good approximation to the number of regions when n is large, and may or may not be a good approximation when n is small. With a bit of experience, we might also judge that the chances of the number of regions having a polynomial dependence on n are reasonably high — something

[14] A less conclusive but still quite convincing argument that 2^{n-1} is unlikely to be the correct formula is that it fails when $n = 0$: if you draw no points, then the number of regions will be 1 and not 1/2. This is less conclusive because sometimes zero behaves differently from other numbers, but it is a disturbing observation nevertheless.

[15] The polynomial in question can be written as $\binom{n-1}{0}+\binom{n-1}{1}+\binom{n-1}{2}+\binom{n-1}{3}+\binom{n-1}{4}$. The advantage of writing it this way is that it makes it clear why it gives powers of 2 for small n and not for large n. Note also that if we substitute in $n = 0$ we obtain the answer 1, so the troubling anomaly at zero is no longer present.

like 75%. (This judgment is based on having seen a multitude of problems, some of which lead to nice exact formulae and others of which don't, and feeling that this one could go either way but is probably the right sort of problem for a polynomial dependence.)

If the first hypothesis is correct, then the next term of the sequence will be 31. If it is incorrect, then the next term might be 31 just by chance, but that is fairly unlikely — there are probably around half a dozen numbers it might be. If we take this probability to be 1/6, then Bayes's formula tells us that the probability that the first hypothesis is true given that the number of regions is indeed 31 when there are six points is

$$\frac{3/4}{\frac{3}{4}+\frac{1}{6}\times\frac{1}{4}}=\frac{18}{19}.$$

Of course, the above calculation depended very much on the prior: if we had started out with a greater faith in the formula turning out to be nice, then our confidence after the experiment would be that much higher.

If we work out the next value of the polynomial, we find that it is 57, and this agrees with the number of regions when there are seven points round the circle. A Bayesian calculation similar to the above, but with 1/6 replaced by a smaller fraction (because now the range of plausible numbers if there is not an exact formula is larger) increases the probability to one that is quite a bit closer to 1 — something like 199/200.

It is worth reiterating that this sort of reasoning is of more than merely philosophical interest, and plays an extremely important role in the choices mathematicians make when deciding what to think about next when they are searching for proofs. This can be seen very directly if one looks at a famous database of integer sequences created by Neil Sloane. Sloane's database is a huge collection of beginnings of sequences obtained from all over mathematics, and it has led to many discoveries in the following way. A mathematician thinks about a problem that leads naturally to the definition of an integer sequence, calculates the first few terms of that sequence, searches for a sequence in Sloane's database that agrees with this sequence as far as the calculations have been carried out, and then guesses that the two sequences are equal. This often turns out to be the case even if the sequences have been obtained in very different ways. There follows the fascinating process of trying to explain *why* the sequences are the same — the no-coincidence principle gives one considerable confidence that such an explanation exists.

The two conclusions we can draw about Goodman-like mathematical statements after the discussion above are these.

1. Some such statements are not backed up by experimental evidence because they do not require a miracle to be false, so the analysis of why we judge some statements to be likely, and of the role that experimental evidence can play in that judgment, does not apply.
2. What looks like a natural hypothesis can change dramatically when one knows more about a question.

This does not fully solve Goodman's paradox in a mathematical context, however. For instance, I have not given a Goodman-proof account of what a "miracle" is or of what a "nice formula" is.

These two questions are closely related. For instance, we do not say that it is a miracle that of all the million possibilities that it could be, the sequence of six consecutive digits of π that starts at the 20th digit after the decimal point should turn out to be 626433, but had it turned out to be 000000 then we would have been surprised. Had it been 333333 or 123456 we would have still been surprised but perhaps slightly less so, and if it had been 345678 or 246810 we would have been slightly less surprised still. And the reason for that seems to be tied to how easy it is to describe the sequences.[16] [17]

So a rough definition of "miracle" is something like this. Before we do an experiment (which might take the form of a hand calculation) we have a probability distribution in mind of what the outcome is likely to be. If a certain property is, according to that probability distribution, very unlikely to occur *and also has a very simple description*, then that counts as a miracle. Of course, it is only a miracle from our prior perspective: the no-miracle principle tells us that there is probably an explanation for what we have just observed, and if we find one then we will no longer perceive the observation as miraculous.

To give an example, if we found a very large even number that could not be expressed as a sum of two primes, that would be an

[16] The reflex for many mathematicians would be to mention Kolmogorov complexity at this point, but while in many situations it would be very surprising if a random-looking sequence could in fact be generated by a short program, the miracles that occur in normal mathematics tend to involve sequences with patterns that are simpler than is guaranteed by their being the output of a short program.

[17] Only after writing this paragraph did I learn that starting at the 768th digit of π there is a sequence of six consecutive 9s. This is genuinely surprising, but not surprising enough to demand an explanation.

extremely unlikely event according to a very natural random model of the primes, and it is also an easy event to describe, so we would count it as a miracle. The event "Can be written as a sum of two primes in under half the number of ways one would expect" would count as a miracle if it occurred — not quite as big a miracle, but still (to an expert) utterly astonishing. But the event "Can be written in r ways for some positive integer r that has four 5s among its last six digits" would not be a miracle — there are just too many descriptions of about that level of complexity, given the probability of the event that we are witnessing.

Thus, it remains to discuss concepts such as "nice formula" or "simple description" or "natural statement".

§ 6. — Naturalness and levels of abstraction.

If I am using LaTeX to typeset a document and want to emphasize a word, then I have two options. One is to enclose the word in curly brackets and write `\textit` outside it. This has the effect of putting the word into italics. The other is to use the command `\emph` instead, which stands for "emphasize". Here is a sample sentence, typeset once using `\textit` and once using `\emph`.

1. We have also *troubled* the previous number four times.

2. We have also *troubled* the previous number four times.

Which command I chose made no difference to the output, but there is nevertheless an important difference between the two commands, and in many contexts it is considered better style to use `\emph`. To see why, let us look at an example where the choice of command *does* make a difference. Again, I shall use `\textit` first and then `\emph`.

Definition. *A positive integer n is said to be troubled if $n \leqslant 10$ and it is doubled or $n > 10$ and it is trebled.*

Definition. *A positive integer n is said to be* troubled *if $n \leqslant 10$ and it is doubled or $n > 10$ and it is trebled.*

The advantage of `\emph` is now obvious: inside the particular LaTeX definition environment I chose, the non-mathematical text is italicized, so if I use `\textit` in order to stipulate that the word "troubled" should be italicized, the result is that it is not distinguished from its surroundings and is therefore not emphasized.

But if I use the command `\emph`, then LaTeX knows that I want to emphasize the word, so it chooses a contrasting font *whatever environment it is in*. The command `\textit` is something like a rigid designator — it causes text to be in *that* font, regardless of the environment — whereas `\emph` is more like a definite description — it causes text to be in "the font that contrasts appropriately with the surrounding font", which varies from environment to environment.

A great advantage of "nonrigid" commands is that they make it much easier to change the look of a document. Suppose, for example, that a journal's house style was to use boldface to emphasize words in definition environments. It could have a style file in which the `\emph` command was defined so as to ensure that that happened, and there would be no need to make any changes to the source file of the document itself.

Let me now jump to a different example. In 1988, Douglas Hofstadter, Melanie Mitchell and others developed a program called Copycat, with an extremely interesting architecture, which was designed to solve simple analogy problems concerning letter strings, of which a representative example is the problem `abc:abd::ijk:?`, which can be read as "`abc` is to `abd` as `ijk` is to what?" The class of problems of this kind is surprisingly rich, which makes solving them an interesting computational challenge. But here I just want to discuss what makes one solution more satisfying than another.

For that purpose, let us look at the following example: `abc` is to `abd` as `abbcccdddd` is to what? Here are several possible answers to that question, together with brief justifications for each.

1. `abd` (replace the string by `abd`)
2. `abdcccdddd` (replace the third letter in the string by `d`)
3. `abbcccdddd` (replace the last letter by a `d`)
4. `abcccdddd` (replace the third letter by its successor)
5. `abbddddddd` (replace all `c`'s by `d`'s)
6. `abbddddddd` (replace all `c`'s by their successors)
7. `abbcccddde` (replace the last letter by its successor)
8. `abbccceeee` (replace the letters in the last group by their successors)
9. `abbccceeeee` (replace the key parameters that define the last group by their successors)

It is clear that the last answer is by far the most satisfying, and that amongst the others some are more satisfying than others, with some of the less satisfying ones being quite laughable. For instance, a natural reaction to the first proposed answer is "Why would one replace every string by `abd`?"

A good measure of what makes one answer better than another appears to be nonrigidity. In each case, the justification takes the form of a description of some process that can be applied to at least some letter strings. Some of these processes involve constants, such as particular letters of the alphabet or particular small numbers. To the extent that they do so, they are rigid. Others are defined in ways that vary in a simple manner according to the sequence, which makes them nonrigid. For example, the second answer makes use of the letter `d` and the number 3, so it is a highly rigid process, and that is why it seems ridiculous, and similar reasoning applies, to differing extents, to all of the first six proposals.

The analysis of the last three answers is slightly subtler, but still involves rigidity. If we look at the sequence `abbcccdddd`, we very easily spot that it can be described as one `a` followed by two `b`'s followed by three `c`'s followed by four `d`'s. And while we do not have a convenient language to express the thought concisely, we definitely also notice that the letters and numbers involved in that description were not arbitrary: we went through the first four letters of the alphabet in order and the first four natural numbers in order. We can even go up a level of abstraction and say that we performed the same process on the natural numbers as we did on the letters (thereby avoiding having to say twice that we chose the first four in order).

This description is greatly preferable to describing the sequence by simply listing it, because it is significantly less rigid. Instead of specifying a string of ten letters one by one (as we would have to do for a string such as `rjkepgjwrt`), we can regard the generation of the sequence as itself the solution of some smaller-scale analogy problems, such as `a:bb::bb:?`, together with an instruction to stop after four steps. (We would also think of `a` not as "the letter `a`" but as "the first letter of the alphabet", and so on.)

Once we think about the sequence in this way, a natural response to answer (7) is "Why do you want to do something to the last letter in the sequence?" The operation "pick the last term" is a bit too rigid, but it can be made less so if we replace it by "pick the last group", which yields answer (8). Similarly, the operation "replace

all letters in the last group by their successors" can be made less rigid still if we go for "take the letters generated in the last stage of the generating process and replace them by the letters in what would have been the next stage of the generating process".

The suggestion I would like to make in this section is that the more abstract (that is, non-rigid) a statement is in mathematics, the more natural it is perceived as being, and the more likely it is that we will be convinced by confirming evidence. For example, the variant of Goldbach's conjecture that I suggested earlier, that every even number is a sum of two primes as long as it is not a power of $2^{1,000,000,000}$, is not a natural statement because it depends on the constant $2^{1,000,000,000}$ (though the fact that we can express it more concisely as $2^{10^{3^2}}$ makes it more natural than if we replaced it by a number that was much harder to specify).

For a more interesting example, consider the following two statements.

Statement 1. *Let G be a group. If $x^2 = e$ for every element x of G, then G is Abelian.*

Statement 2. *Let G be a group. If $x^5 = e$ for every element x of G, then G is Abelian.*

As it happens, the first of these statements is true and the second is false. However, I would like to argue also that the first statement is more natural than the second.

At first sight they seem very similar — all I have done is replace one constant, 2, by another. However, the first statement can be reformulated in a more abstract way as follows.

Statement 1 (Equivalent version)**.** *Let G be a group. If every element of G is equal to its own inverse, then G is Abelian.*

Once the statement is formulated like this, we see that 2 was not really a constant: rather, it was the only number that bore a particular relation to the notion of a group. (This is similar to our earlier recognition that the string `abbcccdddd` can be specified in a more abstract way.)

The fact that 5 does not have this property has important consequences for what my judgments about the second statement would have been had I not known whether it was true or false. I would have expected that *if* the second statement was true, then it would be a special case of a more general statement such as that for any

prime p, a group G must be Abelian if x^p is always equal to the identity. That would be a more abstract statement and therefore more natural.

Similarly, the fact that 2 does have this property has a big influence on how I would set about proving Statement 1 if I did not already know the proof. I would try to resist thinking of 2 as "that number" and instead focus on its properties, so if I spotted that the equation $x^2 = e$ implies the equation $x = x^{-1}$, I would seize on that observation as being likely to help.[(18)]

For a third example, consider the following two theorems from linear algebra, and particularly their proofs.

Theorem. *Let V be a finite-dimensional vector space. Then V is isomorphic to its dual space V^*.*

Proof. Let $v_1, \dots, v_n$ be a basis of V. Then define a basis $v_1^*, \dots, v_n^*$ of V^* as follows: if v is a vector in V, then it can be written uniquely in the form $\sum_{j=1}^n \lambda_j v_j$. Let $v_i^*(v)$ be defined to be λ_i. (Thus, v_i^* picks out the ith coordinate of v with respect to the given basis.) It is easy to check that $v_1^*, \dots, v_n^*$ is a basis of V^*, and then the map $\sum_i \lambda_i v_i \mapsto \sum_i \lambda_i v_i^*$ defines an isomorphism between the two spaces. □

Theorem. *Let V be a vector space. Then V embeds isomorphically into its second dual space V^{**}.*

Proof. Let $\mathbb{F}$ be the field of scalars of V. For each $v \in V$ we can define a linear functional $\tau_v : V^* \to \mathbb{F}$ by $\tau_v(v^*) = v^*(v)$. If $\tau_v(v^*) = 0$ for all v^*, then $v^*(v) = 0$ for all v^*, so $v^* = 0$. Also, the map $v \mapsto \tau_v$ is easily checked to be linear. Therefore, that map is an isomorphic embedding. □

There is a very important distinction between the two proofs, which leads mathematicians to describe the isomorphism given in the first proof as unnatural and the isomorphic embedding given in the second proof as natural. Indeed, in contexts such as this one, the word "natural" has a formal meaning (the category-theoretic concept of a *natural transformation*) that captures the difference. The clearest sign of the difference is the first sentence of the first proof. One is asked to take a basis, but is not told how to do so. Indeed,

(18) And indeed it does. It is a standard fact in elementary group theory that $ab = b^{-1}a^{-1}$ and if every element equals its own inverse then we deduce immediately that $ab = ba$.

given an arbitrary finite-dimensional vector space there is no single best method of picking a basis. (For example, consider the 2-dimensional vector space of all triples $(x, y, z) \in \mathbb{R}^3$ such that $x + y + z = 0$. It is easy to find a basis, but not easy to argue of any particular basis that it is the most natural and obvious basis to pick.) Thus, the basis feels like a rigid choice, because it does not relate to the space in a clear way. This is a slightly different kind of rigidity: it is not saying that the basis is always the same regardless of the space, but rather that each basis one chooses exists only in the space for which it was chosen. Mathematicians would say that the choice is *non-canonical*. By contrast, the choice of τ_v given v is entirely canonical — it is the obvious linear functional to build out of v.

Returning to the example of division of the circle, we can see why the sequences that played a role in that discussion were natural. We knew from experiment that we had a sequence that began 1,2,4,8,16. The sequence of powers of 2 can be thought of as the solution to the easy analogy problem `1:2::2:4::4:8::8:16::16:?` (and the reason we regard the operation performed at each step as doubling rather than troubling is that the definition of troubling involves additional constants such as 3 and 10), while the correct sequence 1,2,4,8,16,31,57,... can be defined as the sequence of values of the lowest-degree polynomial consistent with the values 1,2,4,8,16. (Another way to pick out the sequence of powers of 2 would be to say that $a_n = P(a_{n-1})$ for some polynomial P, and moreover P is the polynomial of smallest degree that is consistent with the values 1,2,4,8,16, which happens to be the polynomial $2x$.)

Often the naturalness of a mathematical statement is not immediately apparent. The Riemann hypothesis provides an example of this. The precise statement can be written as follows.

Conjecture (Riemann hypothesis)**.** *If $\zeta(s) = 0$, then either s is a negative even integer or the real part of s is equal to $1/2$.*

At first sight, the fact that this statement is suggesting that there are two different kinds of zeros makes it seem unnatural. However, as Riemann showed, the statement has another formulation that lacks this defect. The modern way to express it is to define a function ξ by the formula

$$\xi(s) = \frac{1}{2}s(s-1)\pi^{-s/2}\Gamma(s/2)\zeta(s),$$

which I shall not try to explain here. Riemann showed that the xi-function satisfies the functional equation $\xi(s) = \xi(1-s)$. This property is not enjoyed by the Riemann zeta function, which makes the ξ function in some ways more natural (in a different sense of "natural" from the one I have been discussing). Better still, the zeros of the ξ function are the zeros of the ζ function with the zeros at negative integers excluded, so the Riemann hypothesis is equivalent to the following statement.

Conjecture (Reformulation of Riemann hypothesis)**.** *If $\xi(s) = 0$, then the real part of s is equal to $1/2$.*

But even this statement involves the constant $1/2$. Is that a sign of unnaturalness? For all sorts of reasons it isn't. One is that the vertical line through $1/2$ in the complex plane is the only one that is mapped to itself by the symmetry $s \mapsto 1-s$ of the ξ function — this gives a way of specifying the number $1/2$ in terms of the function rather than simply as the number $1/2$. Another is that, as I mentioned earlier, the Riemann hypothesis is equivalent to a statement about the distribution of the primes, which says, roughly speaking, that the error term in the prime number theorem is roughly what the natural random model would predict. The appearance of the number $1/2$ in this model is closely related to probabilistic phenomena such as that the expected end-to-end distance of a random walk of length n is around $n^{1/2}$. So again we can think of it not as "that number" but as "the exponent that comes up when you add together a number of independent random variables that satisfy some commonly occurring conditions".

As one final example, let me return to the question I left open at the end of my discussion in §2 of the behaviour of the Ramsey number $R(k,k)$. There I had argued that if $R(k,k)$ was approximately C^k for some very large k, then we would expect $R(k,k)$ to be approximately C^k for *all* very large k. I then invented an artificial function $D(k)$ given by the formula $(2+\cos(\pi \log_2 k)/10)^k$. What was unnatural about it?

One answer is that it involved various constants such as π and 10. One could argue that π relates so closely to the cosine function that it should not be regarded as problematic. (In fact, the *absence* of π here could be argued to be less natural than its presence.) But that still leaves 10, and the cosine function itself is also somewhat rigid because it does not relate in any obvious way to the graph-theoretic problem at hand, so has to be thought of as "that function".

Recall that we know that $R(k,k)$ lies between $(\sqrt{2})^k$ and 4^k, or equivalently that $R(k,k)^{1/k}$ lies between $\sqrt{2}$ and 4. The simplest functions that are bounded above and below are constant functions, so the most natural hypothesis, given that small values show that $R(k,k)^{1/k}$ is not in fact constant, is that $R(k,k)$ at least *converges* to a constant. (Applying the nice-functions principle, I then conclude that this hypothesis is very likely to be true.) One could go further and argue that the most natural candidates for the value of C are $\sqrt{2}$ and 4, which would be saying that at least one of the arguments we have found so far cannot be substantially improved. The next most natural candidate is probably 2, since C has to be bigger than 1 and 2 is the simplest such number. (There are also reasons more connected with the problem itself to think that 2 might conceivably be correct.)

§ 7. — Conclusion.

The argument I have put forward can be summarized as follows.

1. Mathematicians update their beliefs using something like Bayes's formula, but with vague notions such as "probable" or "extremely unlikely" replacing actual probabilities.
2. The prior distributions and conditional probabilities are best thought of not as probabilities that certain statements are true, but as probabilities that certain kinds of arguments exist, which we estimate based on our experience with finding or not finding arguments of a similar kind in related contexts.
3. Those probabilities relate to our confidence in the original statements because of the no-miracle principle: miracles in mathematics have explanations, so if we have no evidence of a certain type of miracle existing and cannot conceive of what an explanation might be like if it did exist, then we become confident that it does not exist.
4. Miracles are situations where an event E with a natural description occurs, but according to our prior distribution E is very unlikely to occur. (Thus, whether or not E is miraculous depends heavily on our state of knowledge. In particular, as soon as an explanation is discovered it ceases to be a miracle.)
5. A description of an event is more natural the more abstract and less "rigid" it is. (This too can change as our state of knowledge changes, as additional knowledge can help us reformulate descriptions in more abstract ways.)

There is plenty more one could say about all of these points, but I shall instead end with one more question. I stated the no-miracle principle and the nice-formula principle, which play a very useful guiding role for mathematicians when they are doing research, but gave no justification for either. That is because I find them rather mysterious. Why should it be that mathematics is so full of interesting patterns and that these patterns so often have explanations, some of which are remarkably deep and lead to whole new areas of study? There seems to be something extraordinarily "productive" about the particular set of axioms and deduction rules that we allow ourselves. With a nod to Eugene Wigner, one might call this the unreasonable coherence of mathematics, a miraculous-seeming phenomenon that itself demands an explanation.

Timothy Gowers,
Collège de France and the University of Cambridge.

⋆

⋆ ⋆