THE INFORMATION

A History
A Theory
A Flood

JAMES GLEICK

PANTHEON BOOKS, NEW YORK

THE SENSE OF RANDOMNESS

(In a State of Sin)

"I wonder," she said. "It's getting harder to see the patterns, don't you think?"

—Michael Cunningham (2005)

IN 1958, GREGORY CHAITIN, a precocious eleven-year-old New Yorker, the son of Argentine émigrés, found a magical little book in the library and carried it around with him for a while trying to explain it to other children—and then, he had to admit, trying to understand it himself. It was *Gödel's Proof*, by Ernest Nagel and James R. Newman. Expanded from an article in *Scientific American*, it reviewed the renaissance in logic that began with George Boole; the process of "mapping," encoding statements about mathematics in the form of symbols and even integers; and the idea of metamathematics, systematized language *about* mathematics and therefore *beyond* mathematics. This was heady stuff for the boy, who followed the authors through their simplified but rigorous exposition of Gödel's "astounding and melancholy" demonstration that formal mathematics can never be free of self-contradiction.

The vast bulk of mathematics as practiced at this time cared not at all for Gödel's proof. Startling though incompleteness surely was, it seemed incidental somehow—contributing nothing to the useful work of mathematicians, who went on making discoveries and proving theorems. But philosophically minded souls remained deeply disturbed by it, and these were the sorts of people Chaitin liked to read. One was John von Neumann—who had been there at the start, in Königsberg, 1930, and then in the United States took the central role in the development of computation and computing theory. For von Neumann, Gödel's proof was a point of no return:

> It was a very serious conceptual crisis, dealing with rigor and the proper way to carry out a correct mathematical proof. In view of the earlier notions of the absolute rigor of mathematics, it is surprising that such a thing could have happened, and even more surprising that it could have happened in these latter days when miracles are not supposed to take place. Yet it did happen.

Why? Chaitin asked. He wondered if at some level Gödel's incompleteness could be connected to that new principle of quantum physics, uncertainty, which smelled similar somehow. Later, the adult Chaitin had a chance to put this question to the oracular John Archibald Wheeler. Was Gödel incompleteness related to Heisenberg uncertainty? Wheeler answered by saying he had once posed that very question to Gödel himself, in his office at the Institute for Advanced Study—Gödel with his legs wrapped in a blanket, an electric heater glowing warm against the wintry drafts. Gödel refused to answer. In this way, Wheeler refused to answer Chaitin.

When Chaitin came upon Turing's proof of uncomputability, he thought this must be the key. He also found Shannon and Weaver's book, *The Mathematical Theory of Communication*, and was struck by its upside-down seeming reformulation of entropy: an entropy of bits, measuring information on the one hand and disorder on the other. The common element was randomness, Chaitin suddenly thought. Shannon linked randomness, perversely, to information. Physicists had found randomness inside the atom—the kind of randomness that Einstein deplored by complaining about God and dice. All these heroes of science were talking about or around randomness.

It is a simple word, *random*, and everyone knows what it means. Everyone, that is, and no one. Philosophers and mathematicians struggled endlessly. Wheeler said this much, at least: "Probability, like time, is a concept invented by humans, and humans have to bear the responsibility for the obscurities that attend it." The toss of a fair coin is random, though every detail of the coin's trajectory may be determined à la Newton. Whether the population of France is an even or odd number at any given instant is random, but the population of France itself is surely *not* random: it is a definite fact, even if not knowable. John Maynard Keynes tackled randomness in terms of its opposites, and he chose three: knowledge, causality, and design. What is known in advance, determined by a cause, or organized according to plan cannot be random.

"Chance is only the measure of our ignorance," Henri Poincaré famously said. "Fortuitous phenomena are by definition those whose laws we do not know." Immediately he recanted: "Is this definition very satisfactory? When the first Chaldean shepherds watched the movements of the stars, they did not yet know the laws of astronomy, but would they have dreamed of saying that the stars move at random?" For Poincaré, who understood chaos long before it became a science, examples of randomness included such phenomena as the scattering of raindrops, their causes physically determined but so numerous and complex as to be unpredictable. In physics—or wherever natural processes seem unpredictable—apparent randomness may be noise or may arise from deeply complex dynamics.

Ignorance is subjective. It is a quality of the observer. Presumably randomness—if it exists at all—should be a quality of the thing itself. Leaving humans out of the picture, one would like to say that an event, a choice, a distribution, a game, or, most simply, a number is random.

The notion of a random number is full of difficulties. Can there be such thing as a *particular* random number; a *certain* random number? This number is arguably random:

1009732533765201358634673548768095909117392927494 5 . . .

Then again, it is special. It begins a book published in 1955 with the title *A Million Random Digits*. The RAND Corporation generated the digits by means of what it described as an electronic roulette wheel: a pulse generator, emitting 100,000 pulses per second, gated through a five-place binary counter, then passed through a binary-to-decimal converter, fed into an IBM punch, and printed by an IBM model 856 Cardatype. The process took years. When the first batch of digits was tested, statisticians discovered significant biases: digits, or groups of digits, or patterns of digits that appeared too frequently or not frequently enough. Finally, however, the tables were published. "Because of the very nature of the tables," the editors said wryly, "it did not seem necessary to proofread every page of the final manuscript in order to catch random errors of the Cardatype."

The book had a market because scientists had a working need for random numbers in bulk, to use in designing statistically fair experiments and building realistic models of complex systems. The new method of Monte Carlo simulation employed random sampling to model phenomena that could not be solved analytically; Monte Carlo simulation was invented and named by von Neumann's team at the atomic-bomb project, desperately trying to generate random numbers to help them calculate neutron diffusion. Von Neumann realized that a mechanical computer, with its deterministic algorithms and finite storage capacity, could never generate truly random numbers. He would have to settle for *pseudorandom* numbers: deterministically generated numbers that behaved as if random. They were random enough for practical purposes. "Any one who considers arithmetical methods of producing random digits is, of course, in a state of sin," said von Neumann.

Randomness might be defined in terms of order—its absence, that is. This orderly little number sequence can hardly be called "random":

00000

Yet it makes a cameo appearance in the middle of the famous million random digits. In terms of probability, that is to be expected: "00000"

s likely to occur as any of the other 99,999 possible five-digit strings.
ewhere in the million random digits we find:

)10101

is, too, appears patterned.
To pick out fragments of pattern in this jungle of digits requires work
an intelligent observer. Given a long enough random string, every
sible short-enough substring will appear somewhere. One of them
be the combination to the bank vault. Another will be the encoded
nplete works of Shakespeare. But they will not do any good, because
one can find them.
Perhaps we may say that numbers like 00000 and 010101 can be
dom in a particular context. If a person flips a fair coin (one of the
plest mechanical random-number generators) long enough, at some
nt the coin is bound to come up heads ten times in a row. When
t happens, the random-number seeker will typically discard the result
l go for a coffee break. This is one of the ways humans do poorly at
erating random numbers, even with mechanical assistance. Research-
have established that human intuition is useless both in predicting
domness and in recognizing it. Humans drift toward pattern willy-
y. The New York Public Library bought *A Million Random Digits* and
lved it under Psychology. In 2010 it was still available from Amazon
eighty-one dollars.

number is (we now understand) information. When we modern peo-
, Shannon's heirs, think about information in its purest form, we may
agine a string of 0s and 1s, a binary number. Here are two binary
ngs, fifty digits long:

A: 01010101010101010101010101010101010101010101010101
B: 10001010111110101110100110101000011000100111101111

If Alice (A) and Bob (B) both say they generated their strings by flipping a coin, no one will ever believe Alice. The strings are surely not equally random. Classical probability theory offers no solid reason for claiming that B is more random than A, because a random process *could* produce either string. Probability is about ensembles, not individual events. Probability theory treats events statistically. It does not like questions in the form "How likely was that to happen?" If it happened, it happened.

To Claude Shannon, these strings would look like messages. He would ask, *How much information* does each string contain? On their face, they both contain fifty bits. A telegraph operator charging by the digit would measure the length of the messages and give Alice and Bob the same bill. Then again, the two messages seem to differ profoundly. Message A immediately becomes boring: once you see the pattern, further repetitions provide no new information. In message B, every bit is as valuable as every other. Shannon's first formulation of information theory treated messages statistically, as choices from the ensemble of all possible messages—in the case of A and B, 2^{50} of them. But Shannon also considered redundancy within a message: the pattern, the regularity, the order that makes a message compressible. The more regularity in a message, the more predictable it is. The more predictable, the more redundant. The more redundant a message is, the less information it contains.

The telegraph operator sending message A has a shortcut: he can transmit something like "Repeat '01' twenty-five times." For longer messages with easy patterns, the savings in keystrokes becomes enormous. Once the pattern is clear, the extra characters are free. The operator for message B must soldier on the hard way, sending every character, because every character is a complete surprise; every character costs one bit. This pair of questions—*how random* and *how much information*—turn out to be one and the same. They have a single answer.

Chaitin was not thinking about telegraphs. The device he could not get out of his head was the Turing machine—that impossibly elegant abstraction, marching back and forth along its infinite paper tape, reading and writing symbols. Free from all the real world's messiness, free

from creaking wheel-work and finical electricity, free from any need for speed, the Turing machine was the ideal computer. Von Neumann, too, had kept coming back to Turing machines. They were the ever-handy lab mice of computer theory. Turing's *U* had a transcendent power: a universal Turing machine can simulate any other digital computer, so computer scientists can disregard the messy details of any particular make or model. This is liberating.

Claude Shannon, having moved from Bell Labs to MIT, reanalyzed the Turing machine in 1956. He stripped it down to the smallest possible skeleton, proving that the universal computer could be constructed with just two internal states, or with just two symbols, 0 and 1, or blank and nonblank. He wrote his proof in words more pragmatic than mathematical: he described exactly how the two-state Turing machine would step left and right, "bouncing" back and forth to keep track of the larger numbers of states in a more complex computer. It was all very intricate and specific, redolent of Babbage. For example:

> When the reading head moves, the state information must be transferred to the next cell of the tape to be visited using only two internal states in machine B. If the next state in machine A is to be (say) state 17 (according to some arbitrary numbering system) this is transferred in machine B by "bouncing" the reading head back and forth between the old cell and the new one 17 times (actually 18 trips to the new cell and 17 back to the old one).

The "bouncing operation" carries the information from cell to cell, and the cells act as "transmitters" and "controllers."

Turing had titled his great paper "On Computable Numbers," but of course the real focus was on *un*computable numbers. Could uncomputable numbers and random numbers be related? In 1965 Chaitin was an undergraduate at the City College of New York, writing up a discovery he hoped to submit to a journal; it would be his first publication. He began, "In this paper the Turing machine is regarded as a general purpose computer and some practical questions are asked about programming it." Chaitin, as a high-school student in the Columbia Science Honors Program, had the opportunity to practice programming in machine language on giant IBM mainframes, using decks of punched cards—one card for each line of a program. He would leave his card deck in the computer center and come back the next day for the program's output. He could run Turing machines in his head, too: *write 0, write 1, write blank, shift tape left, shift tape right.* . . . The universal computer gave him a nice way to distinguish between numbers like Alice and Bob's A and B. He could write a program to make a Turing machine print out "010101 . . ." a million times, and he could write down the length of that program—quite short. But given a million random digits—no pattern, no regularity, nothing special at all—there could be no shortcut. The computer program would have to incorporate the entire number. To make the IBM mainframe print out those million digits, he would have to put the whole million digits into the punched cards. To make the Turing machine do it, he would still need the million digits for input.

Here is another number (in decimal this time):

C: 3.14159265358979323846264338327950288419716939937 51 . . .

This looks random. Statistically each digit appears with the expected frequency (one in ten); likewise each pair of digits (one in a hundred), each triplet, and so on. A statistician would say it appears to be "normal," as far as anyone can tell. The next digit is always a surprise. The works of Shakespeare will be in there, eventually. But someone might recognize this as a familiar number, π. So it is not random after all.

But why do we say π is not random? Chaitin proposed a clear answer: a number is not random if it is computable—if a definable computer program will generate it. Thus computability is a measure of randomness.

For Turing computability was a yes-or-no quality—a given number either is or is not. But we would like to say that some numbers are more

:andom than others—they are less patterned, less orderly. Chaitin said :he patterns and the order express computability. Algorithms generate patterns. So we can gauge computability by looking at *the size of the algo-rithm*. Given a number—represented as a string of any length—we ask, what is the length of the shortest program that will generate it? Using the anguage of a Turing machine, that question can have a definite answer, neasured in bits.

Chaitin's algorithmic definition of randomness also provides an algo-ithmic definition of information: the size of the algorithm measures iow much information a given string contains.

Looking for patterns—seeking the order amid chaos—is what scien-ists do, too. The eighteen-year-old Chaitin felt this was no accident. He nded this first paper by applying algorithmic information theory to the rocess of science itself. "Consider a scientist," he proposed, "who has een observing a closed system that once every second either emits a ray f light or does not."

He summarizes his observations in a sequence of 0s and 1s in which a 0 represents "ray not emitted" and a 1 represents "ray emitted." The sequence may start

0110101110 . . .

and continue for a few thousand more bits. The scientist then examines the sequence in the hope of observing some kind of pattern or law. What does he mean by this? It seems plausible that a sequence of 0s and 1s is patternless if there is no better way to calculate it than just by writing it all out at once from a table giving the whole sequence.

ut if the scientist could discover a way to produce the same sequence ith an algorithm, a computer program significantly shorter than the equence, then he would surely know the events were not random. He ould say that he had hit upon a theory. This is what science always eks: a simple theory that accounts for a large set of facts and allows for prediction of events still to come. It is the famous Occam's razor. "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances," said Newton, "for nature is pleased with simplicity." Newton quantified *mass* and *force*, but *simplicity* had to wait.

Chaitin sent his paper to the *Journal of the Association for Computing Machinery*. They were happy to publish it, but one referee mentioned that he had heard rumors of similar work coming from the Soviet Union. Sure enough, the first issue of a new journal arrived (after a journey of months) in early 1966: *Проблемы Передачи Информации*, *Problems of Information Transmission*. It contained a paper titled "Three Approaches to the Definition of the Concept 'Amount of Information,'" by A. N. Kolmogorov. Chaitin, who did not read Russian, had just time to add a footnote.

Andrei Nikolaevich Kolmogorov was the outstanding mathematician of the Soviet era. He was born in Tambov, three hundred miles southeast of Moscow, in 1903; his unwed mother, one of three sisters Kolmogorova, died in childbirth, and his aunt Vera raised him in a village near the river Volga. In the waning years of tsarist Russia, this independent-minded woman ran a village school and operated a clandestine printing press in her home, sometimes hiding forbidden documents under baby Andrei's cradle.

Moscow University accepted Andrei Nikolaevich as a student of mathematics soon after the revolution of 1917. Within ten years he was proving a collection of influential results that took form in what became the theory of probability. His *Foundations of the Theory of Probability*, published in Russian in 1933 and in English in 1950, remains the modern classic. But his interests ranged widely, to physics and linguistics as well as other fast-growing branches of mathematics. Once he made a foray into genetics but drew back after a dangerous run-in with Stalin's favorite pseudoscientist, Trofim Lysenko. During World War II

Kolmogorov applied his efforts to statistical theory in artillery fire and devised a scheme of stochastic distribution of barrage balloons to protect Moscow from Nazi bombers. Apart from his war work, he studied turbulence and random processes. He was a Hero of Socialist Labor and seven times received the Order of Lenin.

He first saw Claude Shannon's *Mathematical Theory of Communication* rendered into Russian in 1953, purged of its most interesting features by a translator working in Stalin's heavy shadow. The title became *Statistical Theory of Electrical Signal Transmission*. The word *information*, информация, was everywhere replaced with данные, *data*. The word *entropy* was placed in quotation marks to warn the reader against inferring a connection with entropy in physics. The section applying information theory to the statistics of natural language was omitted entirely. The result was technical, neutral, juiceless, and thus unlikely to attract interpretation in the terms of Marxist ideology. These were serious concerns; "cybernetics" was initially defined in the *Short Philosophical Dictionary* (standard reference of ideological orthodoxy) as a "reactionary pseudoscience" and "an ideological weapon of imperialist reaction." Kolmogorov leapt upon Shannon's paper nonetheless; he, at least, was unafraid to use the word *information*. Working with his students in Moscow, he put forth a rigorous mathematical formulation of information theory, with definitions of the fundamental concepts, careful proofs, and new discoveries—some of which, he soon learned to his sorrow, had appeared in Shannon's original paper but had been omitted from the Russian version.

In the Soviet Union, still moderately isolated from the rest of the world's science, Kolmogorov was well placed to carry the banner of information. He was in charge of all mathematics in the *Great Soviet Encyclopedia*, choosing the authors, editing the articles, and writing much of it himself. In 1956 he delivered a long plenary report on the theory of information transmission to the Soviet Academy of Sciences. His colleagues thought this was a bit "addled"—that Shannon's work was "more technology than mathematics," as Kolmogorov recalled it afterward. "It is true," he said, "that Shannon left to his successors the rigorous 'justification' of his ideas in some difficult cases. However, his mathematical intuition was amazingly precise." Kolmogorov was not as enthusiastic about cybernetics. Norbert Wiener felt a kinship with him—they had both done early work on stochastic processes and Brownian motion. On a visit to Moscow, Wiener said, "When I read the works of Academician Kolmogorov, I feel that these are my thoughts as well, this is what I wanted to say. And I know that Academician Kolmogorov has the same feeling when reading my works." But the feeling was evidently not shared. Kolmogorov steered his colleagues toward Shannon instead. "It is easy to understand that as a mathematical discipline cybernetics in Wiener's understanding lacks unity," he said, "and it is difficult to imagine productive work in training a specialist, say a postgraduate student, in cybernetics in this sense." He already had real results to back up his instincts: a useful generalized formulation of Shannon entropy, and an extension of his information measure to processes in both discrete and continuous time.

Prestige in Russia was finally beginning to flow toward any work that promised to aid electronic communication and computing. Such work began almost in a void. Pragmatic electrical engineering barely existed; Soviet telephony was notoriously dismal, a subject for eternally bitter Russian humor. As of 1965, there was still no such thing as direct long-distance dialing. The number of toll calls nationally had yet to surpass the number of telegrams, a milestone that had been reached in the United States before the end of the previous century. Moscow had fewer telephones per capita than any major world city. Nonetheless, Kolmogorov and his students generated enough activity to justify a new quarterly journal, *Problems of Information Transmission*, devoted to information theory, coding theory, theory of networks, and even information in living organisms. The inaugural issue opened with Kolmogorov's "Three Approaches to the Definition of the Concept 'Amount of Information'"—almost a manifesto—which then began its slow journey toward the awareness of mathematicians in the West.

"At each given moment there is only a fine layer between the 'trivial' and the impossible," Kolmogorov mused in his diary. "Mathematical discoveries are made in this layer." In the new, quantitative view of information he saw a way to attack a problem that had eluded probability theory, the problem of randomness. How much information is contained in a given "finite object"? An object could be a number (a series of digits) or a message or a set of data.

He described three approaches: the combinatorial, the probabilistic, and the algorithmic. The first and second were Shannon's, with refinements. They focused on the probability of one object among an ensemble of objects—one particular message, say, chosen from a set of possible messages. How would this work, Kolmogorov wondered, when the object was not just a symbol in an alphabet or a lantern in a church window but something big and complicated—a genetic organism, or a work of art? How would one measure the amount of information in Tolstoy's *War and Peace*? "Is it possible to include this novel in a reasonable way in the set of 'all possible novels' and further to postulate the existence of a certain probability distribution in this set?" he asked. Or could one measure the amount of genetic information in, say, the cuckoo bird by considering a probability distribution in the set of all possible species?

His third approach to measuring information—the algorithmic—avoided the difficulties of starting with ensembles of possible objects. It focused on the object itself.* Kolmogorov introduced a new word for the thing he was trying to measure: *complexity*. As he defined this term, the complexity of a number, or message, or set of data is the inverse of simplicity and order and, once again, it corresponds to information. The simpler an object is, the less information it conveys. The more complexity, the more information. And, just as Gregory Chaitin did, Kolmogorov put this idea on a solid mathematical footing by calculating complexity in terms of algorithms. The complexity of an object is the size of the smallest computer program needed to generate it. An object that can be produced by a short algorithm has little complexity. On the other hand, an object needing an algorithm every bit as long as the object itself has maximal complexity.

A simple object can be generated—or computed, or described—with just a few bits. A complex object requires an algorithm of many bits. Put this way, it seemed obvious. But until now it had not been understood mathematically. Kolmogorov put it this way:

> The intuitive difference between "simple" and "complicated" objects has apparently been perceived a long time ago. On the way to its formalization, an obvious difficulty arises: something that can be described simply in one language may not have a simple description in another and it is not clear what method of description should be chosen.

That difficulty is solved by using computer language. It does not matter which computer language, because they are all equivalent, reducible to the language of a universal Turing machine. The Kolmogorov complexity of an object is the size, in bits, of the shortest algorithm needed to generate it. This is also the amount of information. And it is also the degree of randomness—Kolmogorov declared "a new conception of the notion 'random' corresponding to the natural assumption that randomness is the absence of regularity." The three are fundamentally equivalent: information, randomness, and complexity—three powerful abstractions, bound all along like secret lovers.

For Kolmogorov, these ideas belonged not only to probability theory but also to physics. To measure the complexity of an orderly crystal or a helter-skelter box of gas, one could measure the shortest algorithm needed to describe the state of the crystal or gas. Once again entropy

*"Our definition of the quantity of information has the advantage that it refers to individual objects and not to objects treated as members of a set of objects with a probability distribution given on it. The probabilistic definition can be convincingly applied to the information contained, for example, in a stream of congratulatory telegrams. But it would not be clear how to apply it, for example, to an estimate of the quantity of information contained in a novel or in the translation of a novel into another language relative to the original."

was the key. Kolmogorov had a useful background in difficult physical problems to which these new methods could be applied. In 1941 he had produced the first useful, though flawed, understanding of the local structure of turbulent flows—equations to predict the distribution of whorls and eddies. He had also worked on perturbations in planetary orbits, another problem surprisingly intractable for classical Newtonian physics. Now he began laying the groundwork for the renaissance in chaos theory to come in the 1970s: analyzing dynamical systems in terms of entropy and information dimension. It made sense now to say that a dynamical system produces information. If it is unpredictable, it produces a great deal of information.

Kolmogorov knew nothing of Gregory Chaitin, nor did either man know of an American probability theorist named Ray Solomonoff, who had developed some of the same ideas. The world was changing. Time, distance, and language still divided mathematicians in Russia from their Western counterparts, but the gulf narrowed every year. Kolmogorov often said that no one should do mathematics after the age of sixty. He dreamed of spending his last years as a buoy keeper on the Volga, making a watery circuit in a boat with oars and a small sail. When the time came, buoy keepers had switched to motorboats, and for Kolmogorov, this ruined the dream.

Now the paradoxes returned.

Zero is an interesting number. Books have been written about it. One is certainly an interesting number—it is the first and the foremost (not counting zero), the singular and unique. Two is interesting in all kinds of ways: the smallest prime, the definitive even number, the number needed for a successful marriage, the atomic number of helium, the number of candles to light on Finnish Independence Day. *Interesting* is an everyday word, not mathematicians' jargon. It seems safe to say that any small number is interesting. All the two-digit numbers and many of the three-digit numbers have their own Wikipedia entries.

Number theorists name entire classes of interesting numbers: prime numbers, perfect numbers, squares and cubes, Fibonacci numbers, factorials. The number 593 is more interesting than it looks; it happens to be the sum of nine squared and two to the ninth—thus a "Leyland number" (any number that can be expressed as $x^y + y^x$). Wikipedia also devotes an article to the number 9,814,072,356. It is the largest holodigital square—which is to say, the largest square number containing each decimal digit exactly once.

What would be an uninteresting number? Presumably a random number. The English number theorist G. H. Hardy randomly rode in taxi No. 1729 on his way to visit the ailing Srinivasa Ramanujan in 1917 and remarked to his colleague that, as numbers go, 1,729 was "rather a dull one." On the contrary, replied Ramanujan (according to a standard anecdote of mathematicians), it is the smallest number expressible as the sum of two cubes in two different ways.* "Every positive integer is one of Ramanujan's personal friends," remarked J. E. Littlewood. Due to the anecdote, 1,729 is known nowadays as the Hardy-Ramanujan number. Nor is that all; 1,729 also happens to be a Carmichael number, an Euler pseudoprime, and a Zeisel number.

But even the mind of Ramanujan was finite, as is Wikipedia, as is the aggregate sum of human knowledge, so the list of interesting numbers must end somewhere. Surely there must be a number about which there is nothing special to say. Wherever it is, there stands a paradox: the number we may describe, interestingly, as "the smallest uninteresting number."

This is none other than Berry's paradox reborn, the one described by Bertrand Russell in *Principia Mathematica*. Berry and Russell had devilishly asked, What is the least integer not nameable in fewer than nineteen syllables? Whatever this number is, it can be named in eighteen syllables: *the least integer not nameable in fewer than nineteen syllables*. Explanations for why a number is interesting are ways of naming the number: "the square of eleven," for example, or "the number of stars in the American

* $1729 = 1^3 + 12^3 = 9^3 + 10^3$

flag." Some of these names do not seem particularly helpful, and some are rather fuzzy. Some are pure mathematical facts: whether, for example, a number is expressible as the sum of two cubes in two different ways. But some are facts about the world, or about language, or about human beings, and they may be accidental and ephemeral—for example, whether a number corresponds to a subway stop or a date in history.

Chaitin and Kolmogorov revived Berry's paradox in inventing algorithmic information theory. An algorithm names a number. "The paradox originally talks about English, but that's much too vague," Chaitin says. "I pick a computer-programming language instead." Naturally he picks the language of a universal Turing machine.

> And then what does it mean, how do you name an integer? Well, you name an integer by giving a way to calculate it. A program names an integer if its output is that integer—you know, it outputs that integer, just one, and then it stops.

Asking whether a number is interesting is the inverse of asking whether it is random. If the number n can be computed by an algorithm that is relatively short, then n is interesting. If not, it is random. The algorithm PRINT I AND THEN PRINT 100 ZEROES generates an interesting number (a googol). Similarly, FIND THE FIRST PRIME NUMBER, ADD THE NEXT PRIME NUMBER, AND REPEAT A MILLION TIMES generates a number that is interesting: the sum of the first million primes. It would take a Turing machine a long time to compute that particular number, but a finite time nonetheless. The number is computable.

But if the most concise algorithm for n is "PRINT [n]"—an algorithm incorporating the entire number, with no shorthand—then we may say that there is nothing interesting about n. In Kolmogorov's terms, this number is random—maximally complex. It will have to be patternless, because any pattern would provide a way to devise a shorthand algorithm. "If there is a small, concise computer program that calculates the number, that means it has some quality or characteristic that enables you to pick it out and to compress it into a smaller algorithmic description," Chaitin says. "So that's unusual; that's an interesting number."

But *is* it unusual? Looking generally at all the numbers, how can a mathematician know whether the interesting ones are rare or common? For that matter, looking at any one number, can a mathematician ever know for sure whether a smaller algorithm might be found? For Chaitin, these were the critical questions.

He answered the first with a counting argument. The vast majority of numbers have to be uninteresting because there cannot possibly be enough concise computer programs to go around. Count them. Given 1,000 bits (say), one has 2^{1000} numbers; but not nearly that many useful computer programs can be written in 1,000 bits. "There are a lot of positive integers," Chaitin says. "If the programs have to be smaller, then there just aren't enough of them to name all those different positive integers." So most n's of any given length are random.

The next question was far more troubling. Knowing that most numbers are random, and given any particular number n, can mathematicians prove it to be random? They cannot tell by looking at it. They can often prove the opposite, that n is interesting: in that case they just have to find a short algorithm that generates n. (Technically, it must be shorter than $\log_2 n$ bits, the number needed to write n in binary.) Proving the negative is a different story. "Even though most positive integers are uninteresting," Chaitin declared, "you can never be sure. . . . You can only prove it in a small number of cases." One could imagine trying to do it by brute force, writing down every possible algorithm and testing them one by one. But a computer will have to perform the tests—an algorithm testing other algorithms—and soon, Chaitin demonstrated, a new version of Berry's paradox appears. Instead of "the smallest uninteresting number," one inevitably encounters a statement in the form of "the smallest number that we can prove cannot be named in fewer than n syllables." (We are not really talking about syllables any more, of

course, but Turing-machine states.)* It is another recursive, self-looping twist. This was Chaitin's version of Gödel's incompleteness. Complexity, defined in terms of program size, is generally uncomputable. Given an arbitrary string of a million digits, a mathematician knows that it is almost certainly random, complex, and patternless—but cannot be absolutely sure.

Chaitin did this work in Buenos Aires. When he was still a teenager, before he could graduate from City College, his parents moved back to their home in Argentina, and he got a job there with IBM World Trade. He continued to nurse his obsession with Gödel and incompleteness and to send papers to the American Mathematical Society and the Association for Computing Machinery. Eight years later, Chaitin returned to the United States to visit IBM's research center in Yorktown Heights, New York, and placed a telephone call to his hero, then nearing seventy at the Institute for Advanced Study in Princeton. Gödel answered, and Chaitin introduced himself and said he had a new approach to incompleteness, based on Berry's paradox instead of the liar paradox.

"It doesn't make any difference which paradox you use," said Gödel.

"Yes, but . . ." Chaitin said he was on the trail of a new "information-theoretic" view of incompleteness and asked if he could call on Gödel in Princeton. He was staying in the YMCA in White Plains and would take the train, changing in New York City. Gödel agreed, but when the day came, he canceled. It was snowing, and he was fearful for his health. Chaitin never did meet him. Gödel, increasingly unstable, afraid of poisoning, died in the winter of 1978 of self-starvation.

Chaitin spent the rest of his career at the IBM Watson Research Center, one of the last great scientists to be so well supported in work of no plausible use to his corporate patron. He sometimes said he was "hiding" in a physics department; he felt that more conventional mathematicians dismissed him as "a closet physicist" anyway. His work treated mathematics as a sort of empirical science—not a Platonic pipeline to absolute truth, but a research program subject to the world's contingencies and uncertainties. "In spite of incompleteness and uncomputability and even algorithmic randomness," he said, "mathematicians don't want to give up absolute certainty. Why? Well, absolute certainty is like God."

In quantum physics and later in chaos, scientists found the limits to their knowledge. They explored the fruitful uncertainty that at first so vexed Einstein, who did not want to believe that God plays dice with the universe. Algorithmic information theory applies the same limitations to the universe of whole numbers—an ideal, mental universe. As Chaitin put it, "God not only plays dice in quantum mechanics and nonlinear dynamics, but even in elementary number theory."

Among its lessons were these:

- Most numbers are random. Yet very few of them can be *proved* random.
- A chaotic stream of information may yet hide a simple algorithm. Working backward from the chaos to the algorithm may be impossible.
- Kolmogorov-Chaitin (KC) complexity is to mathematics what entropy is to thermodynamics: the antidote to perfection. Just as we can have no perpetual-motion machines, there can be no complete formal axiomatic systems.
- Some mathematical facts are true for no reason. They are accidental, lacking a cause or deeper meaning.

Joseph Ford, a physicist studying the behavior of unpredictable dynamical systems in the 1980s, said that Chaitin had "charmingly captured the essence of the matter" by showing the path from Gödel's incompleteness to chaos. This was the "deeper meaning of chaos," Ford declared:

* More precisely, it looked like this: "The finite binary sequence S with the first proof that S cannot be described by a Turing machine with n states or less" is a $(\log_2 n+c_p)$–state description of S.

> Chaotic orbits exist but they are Gödel's children, so complex, so overladen with information that humans can never comprehend them. But chaos is ubiquitous in nature; therefore the universe is filled with countless mysteries that man can never understand.

Yet one still tries to take their measure.

How much information . . . ?

When an object (a number or a bitstream or a dynamical system) can be expressed a different way in fewer bits, it is compressible. A frugal telegraph operator prefers to send the compressed version. Because the spirit of frugal telegraph operators kept the lights on at Bell Labs, it was natural for Claude Shannon to explore data compression, both theory and practice. Compression was fundamental to his vision: his war work on cryptography analyzed the disguising of information at one end and the recovery of the information at the other; data compression likewise encodes the information, with a different motivation—the efficient use of bandwidth. Satellite television channels, pocket music players, efficient cameras and telephones and countless other modern appurtenances depend on coding algorithms to compress numbers—sequences of bits—and those algorithms trace their lineage to Shannon's original 1948 paper.

The first of these, now called Shannon-Fano coding, came from his colleague Robert M. Fano. It began with the simple idea of assigning short codes to frequent symbols, as in Morse code. They knew their method was not optimal, however: it could not be relied on to produce the shortest possible messages. Within three years it was surpassed by work of a graduate student of Fano's at MIT, David Huffman. In the decades since, versions of the Huffman coding algorithm have squeezed many, many bytes.

Ray Solomonoff, a child of Russian immigrants who studied at the University of Chicago, encountered Shannon's work in the early 1950s and began thinking about what he called the Information Packing Problem: how much information could one "pack" into a given number of bits, or conversely, given some information, how could one pack it into the fewest possible bits. He had majored in physics, studied mathematical biology and probability and logic on the side, and gotten to know Marvin Minsky and John McCarthy, pioneers in what would soon be called artificial intelligence. Then he read Noam Chomsky's offbeat and original paper "Three Models for the Description of Language," applying the new information-theoretic ideas to the formalization of structure in language. All this was bouncing around in Solomonoff's mind; he was not sure where it led, but he found himself focusing on the problem of *induction.* How do people create theories to account for their experience of the world? They have to make generalizations, find patterns in data that are always influenced by randomness and noise. Could one enable a machine to do that? In other words, could a computer be made to learn from experience?

He worked out an elaborate answer and published it in 1964. It was idiosyncratic, and hardly anyone noticed until the 1970s, when both Chaitin and Kolmogorov discovered that Solomonoff had anticipated the essential features of what by then was called algorithmic information theory. In effect, Solomonoff, too, had been figuring out how a computer might look at sequences of data—number sequences or bit strings—and measure their randomness and their hidden patterns. When humans or computers learn from experience, they are using induction: recognizing regularities amid irregular streams of information. From this point of view, the laws of science represent data compression in action. A theoretical physicist acts like a very clever coding algorithm. "The laws of science that have been discovered can be viewed as summaries of large amounts of empirical data about the universe," wrote Solomonoff. "In the present context, each such law can be transformed into a method of compactly coding the empirical data that gave rise to that law." A good scientific theory is economical. This was yet another way of saying so.

Solomonoff, Kolmogorov, and Chaitin tackled three different problems and came up with the same answer. Solomonoff was interested in inductive inference: given a sequence of observations, how can one make the best predictions about what will come next? Kolmogorov was looking for a mathematical definition of randomness: what does it mean to say that one sequence is more random than another, when they have the same probability of emerging from a series of coin flips? And Chaitin was trying to find a deep path into Gödel incompleteness by way of Turing and Shannon—as he said later, "putting Shannon's information theory and Turing's computability theory into a cocktail shaker and shaking vigorously." They all arrived at minimal program size. And they all ended up talking about complexity.

The following bitstream (or number) is not very complex, because it is rational:

D: 142857142857142857142857142857142857142857142857 . . .

It may be rephrased concisely as "PRINT 142857 AND REPEAT," or even more concisely as "1/7." If it is a message, the compression saves keystrokes. If it is an incoming stream of data, the observer may recognize a pattern, grow more and more confident, and settle on *one-seventh* as a theory for the data.

In contrast, this sequence contains a late surprise:

E: 10101010101010101010101010101010101010101010101013

The telegraph operator (or theorist, or compression algorithm) must pay attention to the whole message. Nonetheless, the extra information is minimal; the message can still be compressed, wherever pattern exists. We may say it contains a redundant part and an arbitrary part.

It was Shannon who first showed that anything nonrandom in a message allows compression:

F: 10110101111011011010111010111011110100111011010011110110

Heavy on ones, light on zeroes, this might be emitted by the flip of a biased coin. Huffman coding and other such algorithms exploit statistical regularities to compress the data. Photographs are compressible because of their subjects' natural structure: light pixels and dark pixels come in clusters; statistically, nearby pixels are likely to be similar; distant pixels are not. Video is even more compressible, because the differences between one frame and the next are relatively slight, except when the subject is in fast and turbulent motion. Natural language is compressible because of redundancies and regularities of the kind Shannon analyzed. Only a wholly random sequence remains incompressible: nothing but one surprise after another.

Random sequences are "normal"—a term of art meaning that on average, in the long run, each digit appears exactly as often as the others, one time in ten; and each pair of digits, from 00 to 99, appears one time in a hundred; and each triplet likewise, and so on. No string of any particular length is more likely to appear than any other string of that length. Normality is one of those simple-seeming ideas that, when mathematicians look closely, turn out to be covered with thorns. Even though a truly random sequence must be normal, the reverse is not necessarily the case. A number can be statistically normal yet not random at all. David Champernowne, a young Cambridge friend of Turing's, invented (or discovered) such a creature in 1933—a construction made of all the integers, chained together in order:

G: 123456789101112131415161718192021222324252627282 93 . . .

It is easy to see that each digit, and each combination of digits, occurs equally often in the long run. Yet the sequence could not be less random. It is rigidly structured and completely predictable. If you know where you are, you know what comes next.

Even apart from freaks like Champernowne's, it turns out that normal numbers are difficult to recognize. In the universe of numbers, normality is the rule; mathematicians know for sure that almost all numbers are normal. The rational numbers are not normal, and there are infinitely many rational numbers, but they are infinitely outnumbered by normal numbers. Yet, having settled the great and general question, mathematicians can almost never prove that any particular number is normal. This in itself is one of the more remarkable oddities of mathematics.

Even π retains some mysteries:

C: 3.14159265358979323846264338327950288419716939937 51 . . .

The world's computers have spent many cycles analyzing the first trillion or so known decimal digits of this cosmic message, and as far as anyone can tell, they appear normal. No statistical features have been discovered—no biases or correlations, local or remote. It is a quintessentially nonrandom number that seems to behave randomly. Given the *n*th digit, there is no shortcut for guessing the *n*th plus one. Once again, the next bit is always a surprise.

How much information, then, is represented by this string of digits? Is it information rich, like a random number? Or information poor, like an ordered sequence?

The telegraph operator could, of course, save many keystrokes—infinitely many, in the long run—by simply sending the message "π." But this is a cheat. It presumes knowledge previously shared by the sender and the receiver. The sender has to recognize this special sequence to begin with, and then the receiver has to know what π is, and how to look up its decimal expansion, or else how to compute it. In effect, they need to share a code book.

This does not mean, however, that π contains a lot of information. The essential message can be sent in fewer keystrokes. The telegraph operator has several strategies available. For example, he could say, "Take 4, subtract 4/3, add 4/5, subtract 4/7, and so on." The telegraph operator sends an algorithm, that is. This infinite series of fractions converges slowly upon π, so the recipient has a lot of work to do, but the message itself is economical: the total information content is the same no matter how many decimal digits are required.

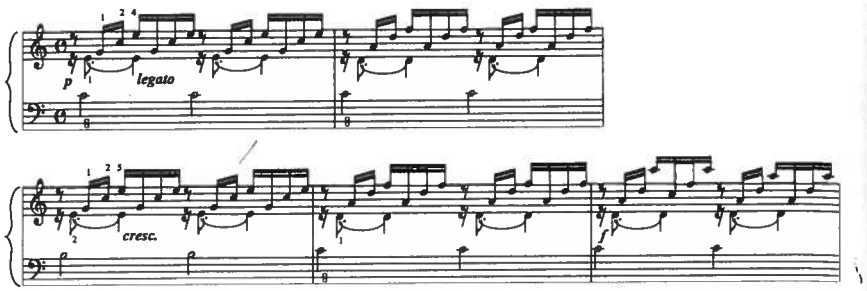
The issue of shared knowledge at the far ends of the line brings complications. Sometimes people like to frame this sort of problem—the problem of information content in messages—in terms of communicating with an alien life-form in a faraway galaxy. What could we tell them? What would we want to say? The laws of mathematics being universal, we tend to think that π would be one message any intelligent race would recognize. Only, they could hardly be expected to know the Greek letter. Nor would they be likely to recognize the decimal digits "3.1415926535 . . ." unless they happened to have ten fingers.

The sender of a message can never fully know his recipient's mental code book. Two lights in a window might mean nothing or might mean "The British come by sea." Every poem is a message, different for every reader. There is a way to make the fuzziness of this line of thinking go away. Chaitin expressed it this way:

> It is preferable to consider communication not with a distant friend but with a digital computer. The friend might have the wit to make inferences about numbers or to construct a series from partial information or from vague instructions. The computer does not have that capacity, and for our purposes that deficiency is an advantage. Instructions given the computer must be complete and explicit, and they must enable it to proceed step by step.

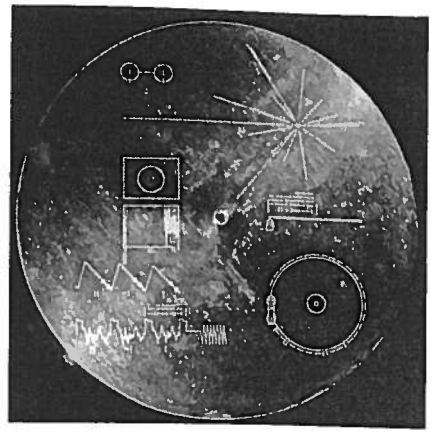
In other words: the message is an algorithm. The recipient is a machine; it has no creativity, no uncertainty, and no knowledge, except whatever "knowledge" is inherent in the machine's structure. By the 1960s, digital computers were already getting their instructions in a form measured in bits, so it was natural to think about how much information was contained in any algorithm.

A different sort of message would be this:



Even to the eye this sequence of notes seems nonrandom. It happens that the message they represent is already making its way through interstellar space, 10 billion miles from its origin, at a tiny fraction of light speed. The message is not encoded in this print-based notation, nor in any digital form, but as microscopic waves in a single long groove winding in a spiral engraved on a disc twelve inches in diameter and one-fiftieth of an inch in thickness. The disc might have been vinyl, but in this case it was copper, plated with gold. This analog means of capturing, preserving, and reproducing sound was invented in 1877 by Thomas Edison, who called it phonography. It remained the most popular audio technology a hundred years later—though not for much longer—and in 1977 a committee led by the astronomer Carl Sagan created a particular phonograph record and stowed copies in a pair of spacecraft named *Voyager 1* and *Voyager 2*, each the size of a small automobile, launched that summer from Cape Canaveral, Florida.

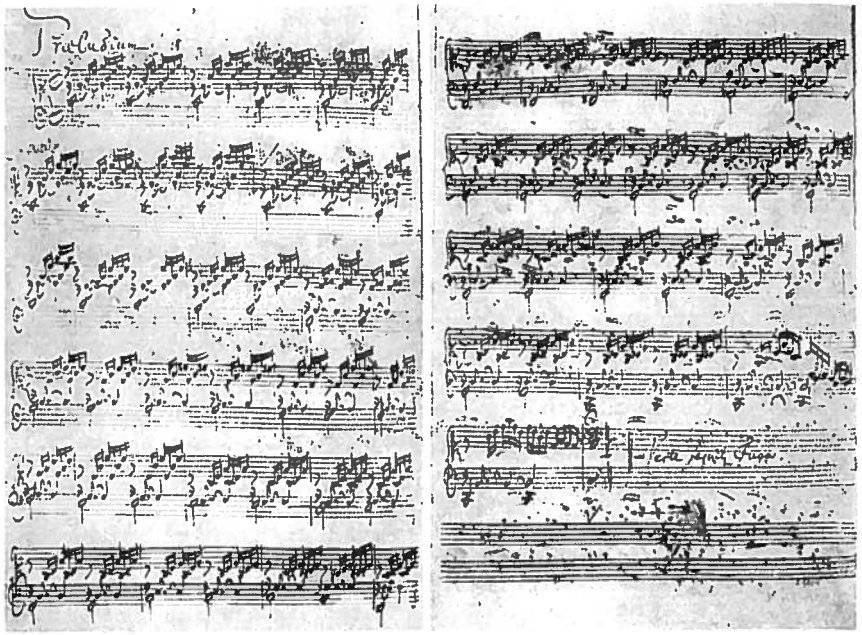
So it is a message in an interstellar bottle. The message has no meaning, apart from its patterns, which is to say that it is abstract art: the first prelude of Johann Sebastian Bach's *Well-Tempered Clavier*, as played on the piano by Glenn Gould. More generally, perhaps the meaning is "There is intelligent life here." Besides the Bach prelude, the record includes music samples from several different cultures and a selection of earthly sounds: wind, surf, and thunder; spoken greetings in fifty-five languages; the voices of crickets, frogs, and whales; a ship's horn, the clatter of a horse-drawn cart, and a tapping in Morse code. Along with the phonograph record are a cartridge and needle and a brief pictographic instruction manual. The committee did not bother with a phonograph player or a source of electrical power. Maybe the aliens will find a way to convert those analog metallic grooves into waves in whatever fluid serves as their atmosphere—or into some other suitable input for their alien senses.



THE "GOLDEN RECORD" STOWED ABOARD THE VOYAGER SPACECRAFT

Would they recognize the intricate patterned structure of the Bach prelude (say), as distinct from the less interesting, more random chatter of crickets? Would the sheet music convey a clearer message—the written notes containing, after all, the essence of Bach's creation? And, more generally, what kind of knowledge would be needed at the far end of the line—what kind of code book—to decipher the message? An appreciation of counterpoint and voice leading? A sense of the tonal context and performance practices of the European Baroque? The sounds—the notes—come in groups; they form shapes, called melodies; they obey the rules of an implicit grammar. Does the music carry its own logic with it, independent of geography and history? On earth, meanwhile, within a few years, even before the *Voyagers* had sailed past the solar system's edge, music was seldom recorded in analog form anymore. Better to store the sounds of the *Well-Tempered Clavier* as bits: the waveforms discretized without loss as per the Shannon sampling theorem, and the information preserved in dozens of plausible media.

In terms of bits, a Bach prelude might not seem to have much information at all. As penned by Bach on two manuscript pages, this one amounts to six hundred notes, characters in a small alphabet. As Glenn Gould played it on a piano in 1964—adding the performer's layers of nuance and variation to the bare instructions—it lasts a minute and thirty-six seconds. The sound of that performance, recorded onto a CD,



microscopic pits burned by a laser onto a slim disc of polycarbonate plastic, comprises 135 million bits. But this bitstream can be compressed considerably with no loss of information. Alternatively, the prelude fits on a small player-piano roll (descendant of Jacquard's loom, predecessor of punched-card computing); encoded electronically with the MIDI protocol, it uses a few thousands bits. Even the basic six-hundred-character message has tremendous redundancy: unvarying tempo, uniform timbre, just a brief melodic pattern, a word, repeated over and over with slight variations till the final bars. It is famously, deceptively simple. The very repetition creates expectations and breaks them. Hardly anything happens, and everything is a surprise. "Immortal broken chords of radiantly white harmonies," said Wanda Landowska. It is simple the way a Rembrandt drawing is simple. It does a lot with a little. Is it then rich in information? Certain music could be considered information poor. At one extreme John Cage's composition titled *4'33"* contains no "notes" at all: just four minutes and thirty-three seconds of near silence, as the piece absorbs the ambient sounds around the still pianist—the listeners' shifting in their seats, rustling clothes, breathing, sighing.

How much information in the Bach C-major Prelude? As a set of patterns, in time and frequency, it can be analyzed, traced, and understood, but only up to a point. In music, as in poetry, as in any art, perfect understanding is meant to remain elusive. If one could find the bottom it would be a bore.

In a way, then, the use of minimal program size to define complexity seems perfect—a fitting apogee for Shannon information theory. In another way it remains deeply unsatisfying. This is particularly so when turning to the big questions—one might say, the human questions—of art, of biology, of intelligence.

According to this measure, a million zeroes and a million coin tosses lie at opposite ends of the spectrum. The empty string is as simple as can be; the random string is maximally complex. The zeroes convey no information; coin tosses produce the most information possible. Yet these extremes have something in common. They are dull. They have no value. If either one were a message from another galaxy, we would attribute no intelligence to the sender. If they were music, they would be equally worthless.

Everything we care about lies somewhere in the middle, where pattern and randomness interlace.

Chaitin and a colleague, Charles H. Bennett, sometimes discussed these matters at IBM's research center in Yorktown Heights, New York. Over a period of years, Bennett developed a new measure of value, which he called "logical depth." Bennett's idea of depth is connected to complexity but orthogonal to it. It is meant to capture the usefulness of a message, whatever usefulness might mean in any particular domain. "From the earliest days of information theory it has been appreciated that information per se is not a good measure of message value," he wrote, finally publishing his scheme in 1988.