Randomness

Paul Vitányi[1]
CWI and Universiteit van Amsterdam

These draft excerpts of the chapter *"Randomness" in 20th Century Mathematics* in preparation for the 'Matematica, Logica, Informatica' Volume 12 of the **Storia del XX Secolo**, to be published by the *Instituto della Enciclopedia Italiana*, are dedicated to Cor Baayen. Here we present in a single essay a combination and completion of the several aspects of the problem of randomness of individual objects which of necessity occur scattered in our text [3].

CONTENTS

1 Introduction **627**

2 Randomness as Unpredictability **631**

3 Randomness in Terms of Expectations **636**

4 Randomness as Incompressibility **639**

1 INTRODUCTION

P.S. Laplace (1749 – 1827) has pointed out the following reason why intuitively a regular outcome of a random event is unlikely.

> "We arrange in our thought all possible events in various classes; and we regard as *extraordinary* those classes which include a very small number. In the game of heads and tails, if head comes up a hundred times in a row then this appears to us extraordinary, because the almost infinite number of combinations that can arise in a hundred throws are divided in regular sequences, or those in which we observe a rule that is easy to grasp, and in irregular sequences, that are incomparably more numerous".

[1]Partially supported by the European Union through NeuroCOLT ESPRIT Working Group Nr. 8556, and by NWO through NFI Project ALADDIN under Contract number NF 62-376. Address: CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Email: paulv@cwi.nl

If by 'regularity' we mean that the complexity is significantly less than maximal, then the number of all regular events is small (because by simple counting the number of different objects of low complexity is small). Therefore, the event that anyone of them occurs has small probability (in the uniform distribution). Yet, the classical calculus of probabilities tells us that 100 heads are just as probable as any other sequence of heads and tails, even though our intuition tells us that it is less 'random' than some others. Listen to the redoubtable Dr. Samuel Johnson:

> "Dr. Beattie observed, as something remarkable which had happened to him, that he chanced to see both the No. 1 and the No. 1000, of the hackney-coaches, the first and the last; 'Why, Sir', said Johnson, 'there is an equal chance for one's seeing those two numbers as any other two.' He was clearly right; yet the seeing of two extremes, each of which is in some degree more conspicuous than the rest, could not but strike one in a stronger manner than the sight of any other two numbers." [Boswell's *Life of Johnson*]

Laplace distinguishes between the object itself and a cause of the object.

> "The regular combinations occur more rarely only because they are less numerous. If we seek a cause wherever we perceive symmetry, it is not that we regard the symmetrical event as less possible than the others, but, since this event ought to be the effect of a regular cause or that of chance, the first of these suppositions is more probable than the second. On a table we see letters arranged in this order `C o r B a a y e n`, and we judge that this arrangement is not the result of chance, not because it is less possible than others, for if this word were not employed in any language we would not suspect it came from any particular cause, but this word being in use among us, it is incomparably more probable that some person has thus arranged the aforesaid letters than that this arrangement is due to chance." [Slightly paraphrasing Laplace]

Let us try to turn Laplace's argument into a formal one. First we introduce some notation. If x is a finite binary sequence, then $l(x)$ denotes the *length* (number of occurrences of binary digits) in x. For example, $l(010) = 3$.

Occam's Razor

Suppose we observe a binary string x of length $l(x) = n$ and want to know whether we must attribute the occurrence of x to pure chance or to a cause. To put things in a mathematical framework, we define *chance* to mean that the literal x is produced by independent tosses of a fair coin. More subtle is the interpretation of *cause* as meaning that the computer on our desk computes x from a program provided by independent tosses of a fair coin. The chance of generating x literally is about 2^{-n}. But the chance of generating x in the form of a short program x^*, the cause from which our computer computes x, is at least $2^{-l(x^*)}$. In other words, if x is regular, then $l(x^*) \ll n$, and it is about $2^{n-l(x^*)}$ times more likely that x arose as the result of computation from some simple cause (like a short program x^*) than literally by a random process.

This approach will lead to an objective and absolute version of the classic maxim of William of Ockham (1290? – 1349?), known as Occam's razor: "if there are alternative explanations for a phenomenon, then, all other things being equal, we should select the simplest one". One identifies 'simplicity of an object' with 'an object having a short effective description'. In other words, *a priori* we consider objects with short descriptions more likely than objects with only long descriptions. That is, objects with low complexity have high probability while objects with high complexity have low probability.

This principle is intimately related with problems in both probability theory and information theory. These problems as outlined below can be interpreted as saying that the related disciplines are not 'tight' enough; they leave things unspecified which our intuition tells us should be dealt with.

Lacuna of Classical Probability Theory

An adversary claims to have a true random coin and invites us to bet on the outcome. The coin produces a hundred heads in a row. We say that the coin cannot be fair. The adversary, however, appeals to probabity theory which says that each sequence of outcomes of a hundred coin flips is equally likely, $1/2^{100}$, and one sequence had to come up.

Probability theory gives us no basis to challenge an outcome *after* it has happened. We could only exclude unfairness in advance by putting a penalty side-bet on an outcome of 100 heads. But what about 1010...? What about an initial segment of the binary expansion of π?

Regular sequence

$$\Pr(00000000000000000000000000) = \frac{1}{2^{26}}$$

Regular sequence

$$\Pr(01000110110000010100111001) = \frac{1}{2^{26}}$$

Random sequence

$$\Pr(10010011011000111011010000) = \frac{1}{2^{26}}$$

The first sequence is regular, but what is the distinction of the second sequence and the third? The third sequence was generated by flipping a quarter. The second sequence is very regular: $0, 1, 00, 01, \ldots$. The third sequence will pass (pseudo-)randomness tests.

In fact, classical probability theory cannot express the notion of *randomness of an individual sequence.* It can only express expectations of properties of outcomes of random processes, that is, the expectations of properties of the total set of sequences under some distribution.

Only relatively recently, this problem has found a satisfactory resolution by combining notions of computability and statistics to express the complexity of a finite object. This complexity is the length of the shortest binary program from which the object can be effectively reconstructed. It may be called the *algorithmic information content* of the object. This quantity turns out to be an attribute of the object alone, and absolute (in the technical sense of being recursively invariant). It is the *Kolmogorov complexity* of the object.

Lacuna of Information Theory

Claude Shannon's classical information theory assigns a quantity of information to an ensemble of possible messages. All messages in the ensemble being equally probable, this quantity is the number of bits needed to count all possibilities.

This expresses the fact that each message in the ensemble can be communicated using this number of bits. However, it does not say anything about the number of bits needed to convey any individual message in the ensemble. To illustrate this, consider the ensemble consisting of all binary strings of length 9999999999999999.

By Shannon's measure, we require 9999999999999999 bits on the average to encode a string in such an ensemble. However, the string consisting of 9999999999999999 1's can be encoded in about 55 bits by expressing 9999999999 999999 in binary and adding the repeated pattern '1'. A requirement for this to work is that we have agreed on an algorithm that decodes the encoded string. We can compress the string still further when we note that 9999999999999999 equals $3^2 \times 1111111111111111$, and that 1111111111111111 consists of 2^4 1's.

Thus, we have discovered an interesting phenomenon: the description of some strings can be compressed considerably, provided they exhibit enough regularity. This observation, of course, is the basis of all systems to express very large numbers and was exploited early on by Archimedes in his treatise *The Sand Reckoner*, in which he proposes a system to name very large numbers:

> "There are some, King Golon, who think that the number of sand is infinite in multitude [...or] that no number has been named which is great enough to exceed its multitude. [...] But I will try to show you, by geometrical proofs, which you will be able to follow, that, of the numbers named by me [...] some exceed not only the mass of sand equal in magnitude to the earth filled up in the way described, but also that of a mass equal in magnitude to the universe."

However, if regularity is lacking, it becomes more cumbersome to express large numbers. For instance, it seems easier to compress the number 'one billion,' than the number 'one billion seven hundred thirty-five million two hundred sixty-eight thousand and three hundred ninety-four,' even though they are of the same order of magnitude.

The above example shows that we need too many bits to transmit regular objects. The converse problem, too little bits, arises as well since Shannon's theory of information and communication deals with the specific technology problem of data transmission. That is, with the information that needs to be

transmitted in order to select an object from a previously agreed upon set of alternatives; agreed upon by both the sender and the receiver of the message. If we have an ensemble consisting of the *Odyssey* and the sentence "let's go drink a beer" then we can transmit the *Odyssey* using only one bit. Yet Greeks feel that Homer's book has more information contents. Our task is to widen the limited set of alternatives until it is universal. We aim at a notion of 'absolute' information of individual objects, which is the information which by itself describes the object completely.

Formulation of these considerations in an objective manner leads again to the notion of shortest programs and Kolmogorov complexity.

2 Randomness as Unpredictability

What is the proper definition of a random sequence, the 'lacuna in probability theory' we have identified above? Let us consider how mathematicians test randomness of individual sequences. To measure randomness, criteria have been developed which certify this quality. Yet, in recognition that they do not measure 'true' randomness, we call these criteria 'pseudo' randomness tests. For instance, statistical survey of initial segments of the sequence of decimal digits of π have failed to disclose any significant deviations of randomness. But clearly, this sequence is so regular that it can be described by a simple program to compute it, and this program can be expressed in a few bits.

> "Any one who considers arithmetical methods of producing random digits is, of course, in a state of sin. For, as has been pointed out several times, there is no such thing as a random number—there are only methods to produce random numbers, and a strict arithmetical procedure is of course not such a method. (It is true that a problem we suspect of being solvable by random methods may be solvable by some rigorously defined sequence, but this is a deeper mathematical question than we can go into now.)" [von Neumann]

This fact prompts more sophisticated definitions of randomness. In his famous address to the International Mathematical Congress in 1900, D. Hilbert proposed twenty-three mathematical problems as a program to direct the mathematical efforts in the twentieth century. The 6th problem asks for "To treat (in the same manner as geometry) by means of axioms, those physical sciences in which mathematics plays an important part; in the first rank are the theory of probability ..". Thus, Hilbert views probability theory as a physical applied theory. This raises the question about the properties one can expect from typical outcomes of physical random sources, which *a priori* has no relation whatsoever with an axiomatic mathematical theory of probabilities. That is, a mathematical system has no direct relation with physical reality. To obtain a mathematical system that is an appropriate model of physical phenomena one needs to identify and codify essential properties of the phenomena under consideration by empirical observations.

Notably Richard von Mises (1883-1953) proposed notions that approach the very essence of true randomness of physical phenomena. This is related with

the construction of a formal mathematical theory of probability, to form a basis for real applications, in the early part of this century. While von Mises' objective was to justify the applications to the real phenomena, A.N. Kolmogorov's (1903-1987) classic 1933 treatment constructs a purely axiomatic theory of probability on the basis of set theoretic axioms.

> "This theory was so successful, that the problem of finding the basis of real applications of the results of the mathematical theory of probability became rather secondary to many investigators. ... [however] the basis for the applicability of the results of the mathematical theory of probability to real 'random phenomena' must depend in some form on the *frequency concept of probability*, the unavoidable nature of which has been established by von Mises in a spirited manner." [Kolmogorov]

The point made is that the axioms of probability theory are designed so that abstract probabilities can be computed, but nothing is said about what probability really means, or how the concept can be applied meaningfully to the actual world. Von Mises analyzed this issue in detail, and suggested that a proper definition of probability depends on obtaining a proper definition of a random sequence. This makes him a 'frequentist'—a supporter of the frequency theory.

The frequency theory to interpret probability says, roughly, that if we perform an experiment many times, then the ratio of favorable outcomes to the total number n of experiments will, *with certainty*, tend to a limit, p say, as $n \to \infty$. This tells us something about the *meaning* of probability, namely, the measure of the positive outcomes is p. But suppose we throw a coin 1000 times and wish to know what to expect. Is 1000 enough for convergence to happen? The statement above does not say. So we have to add something about the rate of convergence. But we cannot assert a *certainty* about a particular number of n throws, such as 'the proportion of heads will be $p \pm \epsilon$ for large enough n (with ϵ depending on n)'. We can at best say 'the proportion will lie between $p \pm \epsilon$ with at least such and such probability (depending on ϵ and n_0) whenever $n > n_0$'. But now we defined probability in an obviously circular fashion.

In 1919 von Mises proposed to eliminate the problem by simply dividing all infinite sequences into special random sequences (called *collectives*), having relative frequency limits, which are the proper subject of the calculus of probabilities and other sequences. He postulates the existence of random sequences (thereby circumventing circularity) as certified by abundant empirical evidence, in the manner of physical laws and derives mathematical laws of probability as a consequence. In his view a naturally occurring sequence can be nonrandom or unlawful in the sense that it is not a proper collective.

> Von Mises views the theory of probabilities insofar as they are numerically representable as a physical theory of definitely observable phenomena, repetitive or mass events, for instance, as found in games of chance, population statistics, Brownian motion. 'Probability' is a primitive notion of the theory comparable to those of 'energy' or 'mass' in other physical theories.

> Whereas energy or mass exist in fields or material objects, probabilities exist only in the similarly mathematical idealization of collectives (random sequences). All problems of the theory of probability consist of deriving, according to certain rules, new collectives from given ones and calculating the distributions of these new collectives. The exact formulation of the properties of the collectives is secondary and must be based on empirical evidence. These properties are the existence of a limiting relative frequency and randomness.
>
> The property of randomness is a generalization of the abundant experience in gambling houses, namely, the impossibility of a successful gambling system. Including this principle in the foundation of probability, von Mises argues, we proceed in the same way as the physicists did in the case of the energy principle. Here too, the experience of hunters of fortune is complemented by solid experience of insurance companies and so forth.
>
> A fundamentally different approach is to justify *a posteriori* the application of a purely mathematically constructed theory of probability, such as the theory resulting from the Kolmogorov axioms. Suppose, we can show that the appropriately defined random sequences form a set of measure one, and without exception satisfy all laws of a given axiomatic theory of probability. Then it appears practically justifiable to assume that as a result of an (infinite) experiment only random sequences appear.

Von Mises' notion of infinite random sequence of 0's and 1's (collective) essentially appeals to the idea that no gambler, making a fixed number of wagers of 'heads', at fixed odds [say p versus $1-p$] and in fixed amounts, on the flips of a coin [with bias p versus $1-p$], can have profit in the long run from betting according to a system instead of betting at random. Says Church: "this definition [below] ... while clear as to general intent, is too inexact in form to serve satisfactorily as the basis of a mathematical theory."

DEFINITION 1 An infinite sequence $a_1, a_2, \ldots$ of 0's and 1's is a random sequence in the special meaning of *collective* if the following two conditions are satisfied.

1. Let f_n is the number of 1's among the first n terms of the sequence. Then

$$\lim_{n\to\infty} \frac{f_n}{n} = p,$$

   for some p, $0 < p < 1$.

2. A *place-selection rule* is a partial function ϕ, from the finite binary sequences to 0 and 1. It takes the values 0 and 1, for the purpose of selecting one after the other those indices n for which $\phi(a_1 a_2 \ldots a_{n-1}) = 1$. We require (1), with the same limit p, also for every infinite subsequence

$$a_{n_1} a_{n_2} \ldots$$

   obtained from the sequence by some *admissible* place-selection rule. (We have not yet formally stated which place-selection rules are admissible.)

The existence of a relative frequency limit is a strong assumption. Empirical evidence from long runs of dice throws, in gambling houses, or with death statistics in insurance mathematics, suggests that the relative frequencies are *apparently convergent*. But clearly, no empirical evidence can be given for the existence of a definite limit for the relative frequency. However long the test run, in practice it will always be finite, and whatever the apparent behavior in the observed initial segment of the run, it is always possible that the relative frequencies keep oscillating forever if we continue.

The second condition ensures that no strategy using an admissible place-selection rule can select a subsequence which allows different odds for gambling than a subsequence which is selected by flipping a fair coin. For example, let a casino use a coin with probability $p = 1/4$ of coming up heads and pay-off heads equal 4 times pay-off tails. This 'Law of Excluded Gambling Strategy' says that a gambler betting in fixed amounts cannot make more profit in the long run betting according to a system than from betting at random.

> "In everyday language we call random those phenomena where we cannot find a regularity allowing us to predict precisely their results. Generally speaking, there is no ground to believe that random phenomena should possess any definite probability. Therefore, we should distinguish between randomness proper (as absence of any regularity) and stochastic randomness (which is the subject of probability theory). There emerges the problem of finding reasons for the applicability of the mathematical theory of probability to the real world." [Kolmogorov]

Intuitively, we can distinguish between sequences that are irregular and do not satisfy the regularity implicit in stochastic randomness, and sequences that are irregular but do satisfy the regularities associated with stochastic randomness. Formally, we will distinguish the second type from the first type by whether or not a certain complexity measure of the initial segments goes to a definite limit. The complexity measure referred to is the length of the shortest description of the prefix (in the precise sense of Kolmogorov complexity) divided by its length. It will turn out that almost all infinite strings are irregular of the second type and satisfy all regularities of stochastic randomness.

> "In applying probability theory we do not confine ourselves to negating regularity, but from the hypothesis of randomness of the observed phenomena we draw definite positive conclusions." [Kolmogorov]

Considering the sequence as fair coin tosses with $p = 1/2$, the second condition in Definition 1 says there is no *strategy* ϕ (*principle of excluded gambling system*) which assures a player betting at fixed odds and in fixed amounts, on the tosses of the coin, to make infinite gain. That is, no advantage is gained in the long run by following some system, such as betting 'head' after each run of seven consecutive tails, or (more plausibly) by placing the nth bet 'head' after the appearance of $n + 7$ tails in succession. According to von Mises, the above conditions are sufficiently familiar and a uncontroverted empirical generalization to serve as the basis of an applicable calculus of probabilities.

EXAMPLE 1 It turns out that the naive mathematical approach to a concrete formulation, admitting simply *all* partial functions, comes to grief as follows.

Let $a = a_1 a_2 \ldots$ be any collective. Define ϕ_1 as $\phi_1(a_1 \ldots a_{i-1}) = 1$ if $a_i = 1$, and undefined otherwise. But then $p = 1$. Defining ϕ_0 by $\phi_0(a_1 \ldots a_{i-1}) = b_i$, with b_i the complement of a_i, for all i, we obtain by the second condition of Definition 1 that $p = 0$. Consequently, if we allow functions like ϕ_1 and ϕ_0 as strategy, then von Mises' definition cannot be satisfied at all. $\diamond$

In the thirties, Abraham Wald proposed to restrict the *a priori* admissible ϕ to any fixed countable set of functions. Then collectives do exist. But which countable set? In 1940, Alonzo Church proposed to choose a set of functions representing 'computable' strategies. According to Church's Thesis, this is precisely the set of *recursive functions*. With recursive ϕ, not only is the definition completely rigorous, and random infinite sequences do exist, but moreover they are abundant since the infinite random sequences with $p = 1/2$ form a set of measure one. From the existence of random sequences with probability $1/2$, the existence of random sequences associated with other probabilities can be derived. Let us call sequences satisfying Definition 1 with recursive ϕ *Mises-Wald-Church random*. That is, the involved *Mises-Wald-Church place-selection rules* consist of the partial recursive functions.

Appeal to a theorem by Wald yields as a corollary that the set of Mises-Wald-Church random sequences associated with any fixed probability has the cardinality of the continuum. Moreover, each Mises-Wald-Church random sequence qualifies as a normal number. (A number is *normal* if each digit of the base, and each block of digits of any length, occurs with equal asymptotic frequency.) Note however, that not every normal number is Mises-Wald-Church random. This follows, for instance, from Champernowne's sequence (or number),

$$0.12345678910111213141516171819 20 \ldots$$

due to D.G. Champernowne, which is normal in the scale of 10 and where the ith digit is easily calculated from i. The definition of a Mises-Wald-Church random sequence implies that its consecutive digits cannot be effectively computed. Thus, an existence proof for Mises-Wald-Church random sequences is necessarily nonconstructive. Unfortunately, the von Mises-Wald-Church definition is not yet good enough, as was shown by J. Ville in 1939. There exist sequences that satisfy the Mises-Wald-Church definition of randomness, with limiting relative frequency of ones of $1/2$, but nonetheless have the property that

$$\frac{f_n}{n} \geq \frac{1}{2} \text{ for all } n.$$

The probability of such a sequence of outcomes in random flips of a fair coin is zero. Intuition: if you bet '1' all the time against such a sequence of outcomes, then your accumulated gain is always positive! Similarly, other properties of randomness in probability theory such as the Law of the Iterated Logarithm do not follow from the Mises-Wald-Church definition.

3 Randomness in Terms of Expectations

For a better understanding of the problem revealed by Ville, and its subsequent solution by P. Martin-Löf in 1966, we look at some aspects of the methodology of probability theory. Consider the sample space of all one-way infinite binary sequences generated by fair coin tosses. Intuitively, we call a sequence 'random' iff it is 'typical'. It is not 'typical', say 'special', if it has a particular distinguishing property. An example of such a property is that an infinite sequence contains only finitely many ones. There are infinitely many such sequences. But the probability that such a sequence occurs as the outcome of fair coin tosses is zero. 'Typical' infinite sequences will have the converse property, namely, they contain infinitely many ones.

In fact, one would like to say that 'typical' infinite sequences will have *all converse properties* of the properties which can be enjoyed by 'special' infinite sequences. This is formalized as follows. If a particular property, such as containing infinitely many occurrences of ones (or zeros), the Law of Large Numbers, or the Law of the Iterated Logarithm, has been shown to have probability one, then one calls this a *Law of Randomness.*

An infinite sequence is 'typical' or 'random' if it satisfies all Laws of Randomness. That is, a *particular* 'random' infinite sequence posesses all properties which are expected to hold with probability one for the ensemble of *all* infinite sequences. This is the substance of so-called pseudo-randomness tests. For example, to test whether the sequence of digits corresponding to the decimal expansion of $\pi = 3.1415\ldots$ is random one tests whether the initial segment satisfies some properties which hold with probability one for the ensemble of all sequences.

Example 2 One such property is so-called normality. E. Borel (1909) has called an infinite sequence of decimal digits *normal* in the scale of ten if, for each k, the frequency of occurrences (possibly overlapping) of each block y of length $k \geq 1$ in the initial segment of length n goes to limit 10^{-k} for n grows unbounded, [1]. It is known that normality is not sufficient for randomness, since Champernowne's sequence

$$123456789101112\ldots$$

is normal in the scale of ten. On the other hand, it is universally agreed that a random infinite sequence must be normal. (If not, then some blocks occur more frequent than others, which can be used to obtain better than fair odds for prediction.)

For a particular binary sequence $\omega = \omega_1\omega_2\ldots$ let $f_n = \omega_1 + \omega_2 + \cdots + \omega_n$. Of course, we cannot effectively test an infinite sequence. Therefore, a so-called pseudo-randomness test examines increasingly long initial segments of the individual sequence under consideration.

We can define a pseudo randomness test for the normality property with $k = 1$ to test a candidate infinite sequence for increasing n whether the deviations

from one half 0's and 1's become too large. For example, by checking for each successive n whether

$$|f_n - \frac{n}{2}| > \sqrt{\frac{n \log \log n}{2}}.$$

(The Law of the Iterated Logarithm states that this inequality should not hold for infinitely many n). If within n trials in this process we find that the inequality holds k times, then we assume the original infinite sequence to be random with confidence at most, say, $\sum_{i=1}^{n} 1/2^i - \sum_{i=1}^{k} 1/2^i$. (The sequence is random if the confidence is greater than zero for n goes to infinity, and not random otherwise.)

Clearly, the number of pseudo-randomness tests we can devise is infinite. Namely, just for the normality property alone there is a similar pseudo-randomness test for each $k \geq 1$. ◇

But now we are in trouble. Each individual infinite sequence induces its very own pseudo-randomness test which tests whether a candidate infinite sequence is in fact that individual sequence. Each infinite sequence forms a singleton set in the sample space of all infinite sequences. *All* complements of singleton sets in the sample space have probability one. The intersection of all complements of singleton sets is clearly empty. Therefore, the intersection of all sets of probability one is empty. Thus, there are no random infinite sequences!

Martin-Löf, using ideas related to Kolmogorov complexity, succeeded in defining random infinite sequences in a manner which is free of such difficulties. His starting point is to observe that all laws which are proven in probability theory to hold with probability one are effective. That is, we can effectively test whether a particular infinite sequence does not satisfy a particular Law of Randomness by effectively testing whether the law is violated on increasingly long initial segments of the sequence.

The natural formalization is to identify the effective test with a partial recursive function. This suggests that one ought to consider not the intersection of all sets of measure one, but only the intersection of all sets of measure one with recursively enumerable complements. (Such a complement set is expressed as the union of a recursively enumerable set of cylinders). It turns out that this intersection has again measure one. Hence, almost all infinite sequences satisfy all effective Laws of Randomness with probability one. This notion of infinite random sequences turns out to be related to infinite sequences of which all finite initial segments have high Kolmogorov complexity.

> The notion of randomness satisfied by both the Mises-Wald-Church collectives and the Martin-Löf random infinite sequences is roughly that *effective tests* cannot detect regularity. This does not mean that a sequence may not exhibit regularities which cannot be effectively tested. Collectives generated by Nature, as postulated by von Mises, may very well always satisfy stricter criteria of randomness. Why should collectives generated by quantum mechanic phenomena care about mathematical notions of computability? Again, satisfaction of all effectively testable prerequisites for randomness is some form of regularity. Maybe nature is

more lawless than adhering strictly to regularities imposed by the statistics of randomness.

Until now the discussion has centered on infinite random sequences where the randomness is defined in terms of limits of relative frequencies. However,

> "The frequency concept based on the notion of *limiting frequency* as the number of trials increases to infinity, does not contribute anything to substantiate the application of the results of probability theory to real practical problems where we always have to deal with a finite number of trials." [Kolmogorov]

The practical objection against both the relevance of considering infinite sequences of trials and the existence of a relative frequency limit is concisely put in J.M. Keynes' famous phrase "in the long run we shall all be dead." It seems more appealing to try to define randomness for finite strings first, and only then define random infinite strings in terms of randomness of initial segments.

The approach of von Mises to define randomness of infinite sequences in terms of *unpredictability* of continuations of finite initial sequences under certain laws (like recursive functions) did not lead to satisfying results. The Martin-Löf approach does lead to satisfying results, and is to a great extent equivalent with the Kolmogorov complexity approach. Although certainly inspired by the random sequence debate, the introduction of Kolmogorov complexity marks a definite shift of point of departure. Namely, to define randomness of sequences by the fact that no program from which an initial segment of the sequence can be computed is significantly shorter than the initial segment itself, rather than that no program can predict the next elements of the sequence. Thus, we change the focus from the 'unpredictability' criterion to the 'incompressibility' criterion, and since this will turn out to be equivalent with Martin-Löf's approach, the 'incompressibility' criterion is both necessary and sufficient.

Finite sequences which cannot be effectively described in a significant shorter description than their literal representation are called random. Our aim is to characterize random infinite sequences as sequences of which all initial finite segments are random in this sense. Martin-Löf's related approach characterizes random infinite sequences as sequences of which all initial finite segments pass all effective randomness tests.

> Initially, before the idea of complexity, Kolmogorov proposed a close analogy to von Mises' notions in the finite domain. Consider a generalization of place-selection rules insofar as the selection of a_i can depend on a_j with $j > i$ [A.N. Kolmogorov, *Sankhyā*, Series A, 25(1963), 369-376]. Let Φ be a finite set of such generalized place-selection rules. Kolmogorov suggested that an arbitrary finite binary sequence a of length $n \geq m$ can be called (m, ϵ)-random with respect to Φ, if there exists some p such that the relative frequency of the 1's in the subsequences $a_{i_1} \ldots a_{i_r}$ with $r \geq m$, selected by applying some ϕ in Φ to a, all lie within ϵ of p. (We discard ϕ that yield subsequences shorter than m.) Stated differently, the relative frequency in this finite subsequence is approximately (to within ϵ) invariant under any of the methods of subsequence selection that yield

subsequences of length at least m. Kolmogorov has shown that if the cardinality of Φ satisfies:

$$d(\Phi) \leq \frac{1}{2} e^{2m\epsilon^2(1-\epsilon)},$$

then, for any p and any $n \geq m$ there is some sequence a of length n which is (m, ϵ)-random with respect to Φ.

4 Randomness as Incompressibility

We are to admit no more causes of natural things (as we are told by *Newton*) than such as are both true and sufficient to explain their appearances. This central theme is basic to the pursuit of science, and goes back to the principle known as Occam's razor: "if presented with a choice between indifferent alternatives, then one ought to select the simplest one". Unconsciously or explicitly, informal applications of this principle in science and mathematics abound. The conglomerate of different research threads drawing on an objective and absolute form of this approach appears to be part of an emergent applied science ranking with information theory and probability theory.

Intuitively, the amount of information in a finite string is the size (number of binary digits or *bits*) of the shortest program that, without additional data, computes the string and terminates. A similar definition can be given for infinite strings, but in this case the program produces element after element forever. Thus, a long sequence of 1's such as

$$\overbrace{11111\ldots1}^{10{,}000 \text{ times}}$$

contains little information because a program of size about $\log 10,000$ bits outputs it:

```
for i := 1 to 10,000
    print 1
```

Likewise, the transcendental number $\pi = 3.1415\ldots$, an infinite sequence of seemingly 'random' decimal digits, contains but a few bits of information. (There is a short program that produces the consecutive digits of π forever.) Such a definition would appear to make the amount of information in a string (or other object) depend on the particular programming language used.

Fortunately, it can be shown that all reasonable choices of programming languages lead to quantification of the amount of 'absolute' information in individual objects that is invariant up to an additive constant. We call this quantity the 'Kolmogorov complexity' of the object. If an object is regular, then it has a shorter description than itself. We call such an object 'compressible'.

More precisely, suppose we want to describe a given object by a finite binary string. We do not care whether the object has many descriptions; however, each description should describe but one object. From among all descriptions

of an object we can take the length of the shortest description as a measure of the object's complexity. It is natural to call an object 'simple' if it has at least one short description, and to call it 'complex' if all of its descriptions are long.

But now we are in danger of falling in the trap so eloquently described in the Richard-Berry paradox, where we define a natural number as "the least natural number that cannot be described in less than twenty words". If this number does exist, we have just described it in thirteen words, contradicting its definitional statement. If such a number does not exist, then all natural numbers can be described in less than twenty words. We need to look very carefully at the notion of 'description'.

Assume that each description describes at most one object. That is, there is a specification method D which associates at most one object x with a description y. This means that D is a function from the set of descriptions, say Y, into the set of objects, say X. It seems also reasonable to require that, for each object x in X, there is a description y in Y such that $D(y) = x$. (Each object has a description.) To make descriptions useful we like them to be finite. This means that there are only countably many descriptions. Since there is a description for each object, there are also only countably many describable objects. How do we measure the complexity of descriptions?

Taking our cue from the theory of computation, we express descriptions as finite sequences of 0's and 1's. In communication technology, if the specification method D is known to both a sender and a receiver, then a message x can be transmitted from sender to receiver by transmitting the sequence of 0's and 1's of a description y with $D(y) = x$. The cost of this transmission is measured by the number of occurrences of 0's and 1's in y, that is, by the length of y. The least cost of transmission of x is given by the length of a shortest y such that $D(y) = x$. We choose this least cost of transmission as the 'descriptional' complexity of x under specification method D.

Obviously, this descriptional complexity of x depends crucially on D. The general principle involved is that the syntactic framework of the description language determines the succinctness of description.

In order to objectively compare descriptional complexities of objects, to be able to say "x is more complex than z", the descriptional complexity of x should depend on x alone. This complexity can be viewed as related to a universal description method which is *a priori* assumed by all senders and receivers. This complexity is optimal if no other description method assigns a lower complexity to any object.

We are not really interested in optimality with respect to all description methods. For specifications to be useful at all it is necessary that the mapping from y to $D(y)$ can be executed in an effective manner. That is, it can at least in principle be performed by humans or machines. This notion has been formalized as 'partial recursive functions'. According to generally accepted mathematical viewpoints it coincides with the intuitive notion of effective computation.

The set of partial recursive functions contains an optimal function which

minimizes description length of every other such function. We denote this function by D_0. Namely, for any other recursive function D, for all objects x, there is a description y of x under D_0 which is shorter than any description z of x under D. (That is, shorter up to an additive constant which is independent of x.) Complexity with respect to D_0 minorizes the complexities with respect to all partial recursive functions.

We identify the length of the description of x with respect to a fixed specification function D_0 with the 'algorithmic (descriptional or Kolmogorov) complexity' of x. The optimality of D_0 in the sense above means that the complexity of an object x is invariant (up to an additive constant independent of x) under transition from one optimal specification function to another. Its complexity is an objective attribute of the described object alone: it is an intrinsic property of that object, and it does not depend on the description formalism. This complexity can be viewed as 'absolute information content': the amount of information which needs to be transmitted between all senders and receivers when they communicate the message in absence of any other *a priori* knowledge which restricts the domain of the message.

Broadly speaking, this means that all description syntaxes which are powerful enough to express the partial recursive functions are approximately equally succinct. The remarkable usefulness and inherent rightness of the theory of Kolmogorov complexity stems from this independence of the description method. Thus, we have outlined the program for a general theory of algorithmic complexity. The four major innovations are as follows.

1. In restricting ourselves to formally effective descriptions our definition covers every form of description that is intuitively acceptable as being effective according to general viewpoints in mathematics and logics.

2. The restriction to effective descriptions entails that there is a universal description method that minorizes the description length or complexity with respect to any other effective description method. This would not be the case if we considered, say, all noneffective description methods. Significantly, this implies Item 3.

3. The description length or complexity of an object is an intrinsic attribute of the object independent of the particular description method or formalizations thereof.

4. The disturbing Richard-Berry paradox above does not disappear, but resurfaces in the form of an alternative approach to proving Kurt Gödel's famous result that not every true mathematical statement is provable in mathematics.

Randomness of Individual Sequences Resolved

The notion of randomness of an infinite sequence in the sense of Martin-Löf, as posessing all effectively testable properties of randomness (one of which

is unpredictability), turns out to be identical with the notion of an infinite sequence having maximal Kolmogorov complexity of all finite initial segments. This equivalence of a single notion being defined by two completely different approaches is a truly remarkable fact. (To be precise, the so-called prefix Kolmogorov complexity of each initial segment of the infinite binary sequence must not decrease more than a fixed constant, depending only on the infinite sequence, below the length of that initial segment, [3].) This property sharply distinguishes the random infinite binary sequences from the nonrandom ones. The set of random infinite binary sequences has uniform measure one. That means that as the outcome from independent flips of a fair coin they occur with probability one.

For finite binary sequences the distinction between randomness and nonrandomness cannot be abrupt, but must be a matter of degree. For example, it would not be reasonable if one string is random but becomes nonrandom if we flip the first nonzero bit. In this context too it has been shown that finite binary sequences which are random in Martin-Löf's sense correspond to those sequences which have Kolmogorov complexity at least their own length. Space limitations forbid a complete treatment of these matters here. Fortunately, it can be found elsewhere, [3].

References

1. D.E. Knuth, *Seminumerical Algorithms*, Addison-Wesley, 1981.
2. A.N. Kolmogorov, Three approaches to the definition of the concept 'quantity of information', *Problems in Information Transmission*, **1:1**(1965), 1-7.
3. M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag, New York, 1993.
4. P. Martin-Löf, On the definition of random sequences, *Information and Control*, (1966).
5. R. von Mises, *Probability, Statistics and Truth*, MacMillan, 1939. Reprint: Dover, 1981.
6. C.E. Shannon, A mathematical theory of communication, *Bell System Tech. J.*, **27**(1948), 379-423, 623-656.