PHYSICAL REVIEW A VOLUME 40, NUMBER 8 OCTOBER 15, 1989

Algorithmic randomness and physical entropy

W. H. Zurek
Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545
(Received 17 June 1988; revised manuscript received 31 March 1989)

Algorithmic randomness provides a rigorous, entropylike measure of disorder of an individual, microscopic, definite state of a physical system. It is defined by the size (in binary digits) of the shortest message specifying the microstate uniquely up to the assumed resolution. Equivalently, algorithmic randomness can be expressed as the number of bits in the smallest program for a universal computer that can reproduce the state in question (for instance, by plotting it with the assumed accuracy). In contrast to the traditional definitions of entropy, algorithmic randomness can be used to measure disorder without any recourse to probabilities. Algorithmic randomness is typically very difficult to calculate exactly but relatively easy to estimate. In large systems, probabilistic ensemble definitions of entropy (e.g., coarse-grained entropy of Gibbs and Boltzmann's entropy $H=\ln W$, as well as Shannon's information-theoretic entropy) provide accurate estimates of the algorithmic entropy of an individual system or its average value for an ensemble. One is thus able to rederive much of thermodynamics and statistical mechanics in a setting very different from the usual. *Physical entropy*, I suggest, is a sum of (i) the missing information measured by Shannon's formula and (ii) of the algorithmic information content—algorithmic randomness—present in the available data about the system. This definition of entropy is essential in describing the operation of thermodynamic engines from the viewpoint of information gathering and using systems. These Maxwell demon-type entities are capable of acquiring and processing information and therefore can "decide" on the basis of the results of their measurements and computations the best strategy for extracting energy from their surroundings. From their internal point of view the outcome of each measurement is definite. The limits on the thermodynamic efficiency arise not from the ensemble considerations, but rather reflect basic laws of computation. Thus inclusion of algorithmic randomness in the definition of physical entropy allows one to formulate thermodynamics from the Maxwell demon's point of view.

I. ENTROPY AND RANDOMNESS

Until recently it was impossible to define the entropy of a single microscopic state of a system. Rather, one had to consider an equilibrium ensemble of identical systems, and calculate how many microscopic configurations have their macroscopic properties identical with the single microstate, entropy that was being sought. Once the number of such macroscopically indistinguishable microscopic configurations W was known, one could employ the celebrated Boltzmann formula[1]

$$H=\ln W \ , \tag{1.1}$$

or its more general version due to Gibbs.[2] In quantum mechanics the analog of Gibb's formula was introduced by von Neumann:[3]

$$H=-\mathrm{Tr}\rho\ln\rho \ . \tag{1.2}$$

Here ρ represents the density matrix of the considered quantum system.

The unsatisfactory feature of this definition of entropy is well known and immediately apparent. When the system follows the dynamical evolution prescribed by the appropriate Hamiltonian, the quantity H defined by Eqs. (1.1) and (1.2) is constant, as the number of microscopic states remains the same (Liouville's theorem). In particular, a single initial microstate will give rise, at any time t, to just one future microstate. Yet, to prove the second law, one must argue that the number of states in which the system can be found actually increases. As the thesis of the above statement for all reversible, Hamiltonian evolutions is incorrect, one might have expected that the search for a dynamical explanation of irreversibility will lead nowhere. The success of statistical mechanics has shown that in spite of its self-contradictory features, the probabilistic description of deterministic systems has captured an important element of truth. Boltzmann's equation and his H theorem as well as Gibb's idea of coarse graining serve as the most prominent examples. Both of these ideas achieve their objective by similar means: They selectively discard part of the information about the investigated system, which justifies introduction of the probabilistic description.[4,5]

Yet the success is not complete. In spite of the numerous quantitative confirmations of thermodynamic formalism, the issues of the "arrow of time" and, more generally, of the nature of entropy continue to be debated more than a century after their inception.[4–10] Several additional developments have added fuel to this discussion. The realization that irreversibility, entropy, and information play a key role in the discussion of the measurement process in quantum theory is the first of them.[3,11] Development of the rigorous theory of com-

munication, with the formal apparatus full of analogies to Boltzmann-Gibbs entropy, is the second.[12,13] More recently, the development of black-hole thermodynamics[14] as well as the discussions of the arrow of time in the cosmological context[15] force one to reexamine the nature of entropy in a setting very different from the one for which it was originally invented.

So far all of these discussions have relied on a probabilistic formalism in which the entropy of a definite, completely known microstate is always zero, and where the membership of this microstate in an ensemble defines the set of probabilities—the density function used to calculate H. Indeed, the proof that Shannon's missing information measure,

$$H = -\sum_i p_i \log_2 p_i \ , \tag{1.3}$$

which we shall also call Boltzmann-Gibbs-Shannon (BGS) entropy, is the unique reasonable additive measure of ignorance[12,13] appears to have dissuaded many from looking for alternatives, although under some specific circumstances such alternatives have proved quite useful.[9]

The aim of this paper is to investigate a new definition of entropy, which does not use this ensemble strategy and which can be applied to the individual microstates of the system. This entropy is based on the rigorous definition of randomness introduced independently by Solomonoff,[16] Kolmogorov,[17] and Chaitin[18] and further elaborated by Kolmogorov, Chaitin, and others.[19–29] Definition of entropy as algorithmic information content capitalizes on an intuitive notion of what is a random number, or a random configuration. Random means difficult to describe or to reproduce. Consider, for example, two binary strings:

0101010101010101010 ,

10011010010110110010 .

The first string can be described as "ten 01's." The second more random string has no apparent, simple description. This does not always mean that no simple description exists. For example, there exists an infinite series beginning with 01110101000001001111 that also "looks random" at first sight but has a simple description: digits of the binary representation of $\sqrt{2}$. Algorithmic entropy of the binary string s can be defined as the shortest possible description that suffices to reproduce s. One can make this definition rigorous by considering the length (expressed in the number of digits) of the shortest algorithm—e.g., computer program for a universal Turing machine—that produces the output s. I shall describe this definition and discuss some properties of algorithmic entropy in Sec. II.

Second III extends the definition of the algorithmic entropy of a binary string to an idealized physical system—Boltzmann gas. I describe there (and in Appendix A) how a proper algorithmic implementation of particle indistinguishability allows one to avoid Gibb's mixing paradox and leads to the correct expression (that is, the Sackur-Tetrode equation) for the entropy of an ideal gas. Moreover, this concrete example provides one with the opportunity to point out some of the difficulties encountered in an attempt to define algorithmic information content of specific physical systems. These difficulties are further considered and partially settled in Appendix B.

A comparison of algorithmic randomness, regarded as entropy, with Boltzmann and Gibbs-Shannon ensemble entropies is presented in Sec. IV. In contrast to the $-\mathrm{Tr}\rho \ln\rho$ entropies, algorithmic entropy can change even in the course of reversible dynamical evolutions. As mentioned briefly in the first part of Sec. IV and considered in more detail in Appendix C, this behavior allows one to propose an algorithmic version of the second law obeyed even by integrable dynamically evolving systems. By contrast, in the usual equilibrium thermodynamic ensembles algorithmic and statistical approaches are likely to provide almost identical answers for the value of entropy. Therefore, even if one were to claim that it is the algorithmic information content alone that measures disorder of the state of the system, the use of the less direct but more convenient probabilistic prescriptions for entropy could be rigorously justified.

While Secs. III and IV do contain some new applications and novel ways of looking at the algorithmic concepts, they can be regarded as an extended introduction to Sec. V, which is concerned with formulating laws of thermodynamics from an internal viewpoint of an "information gathering and using system" (IGUS), an observer-like entity that can perform measurements, process the acquired information, and use the results to take actions aimed at optimizing thermodynamic efficiency of the engine it controls. Section V proposes that both randomness and missing information play a role in defining *physical entropy,* that is, the quantity that limits the amount of the internal energy of a physical system that can be converted into useful work by an IGUS. I demonstrate there that the physical entropy should be regarded as a sum of two separate contributions, one measuring the randomness of the already known aspects of the state of the system, the other expressing the remaining ignorance of the observer about the actual state. This recognition of the dual nature of physical entropy allows one to discuss the thermodynamics of engines operated by a modern day equivalent of a Maxwell demon—a universal Turing machine capable of performing measurements—from the point of view of such an entity, where the measurement outcomes are definite, without a reference to the statistical ensemble describing all conceivable measurement outcomes. Some of the crucial aspects of the laws of thermodynamics can be regarded from that internal point of view as a consequence of the laws of computation.

A discussion of the consequences of adopting the new definition of physical entropy proposed in Sec. V and including both of its complementary contributions—randomness and ignorance—is conducted in Sec. VI. There we shall also touch on the subject of extending the applicability of algorithmic entropy to quantum systems. Further implications of the new definition of entropy for the second law and for the issues arising in the context of the quantum theory of measurements will be explored in future publications.

The possible relevance of algorithmic information con-

tent to thermodynamics has been anticipated by a brief but insightful discussion in the seminal paper by Bennett.[27] Algorithmic information has also been used to characterize information generated by a dynamically evolving chaotic system. This approach, advocated by Ford,[10] focuses on a sequence of states describing a trajectory of a system.

The following are two notes about the manner in which this paper was written and the way in which it should be read.

(i) Algorithmic information theory is a branch of mathematics, and many of the results employed in the discussion below can be established rigorously. I have opted instead to paraphrase them in an informal manner, revealing the key ideas, but, for the sake of brevity and clarity, without attempting to be rigorous. This comment applies equally well to the theorems proposed below for the first time, and to the results obtained elsewhere, which are just described here. Moreover, in order to illustrate the physical significance of the algorithmic approach as early as possible, I discuss only a comfortable minimum of the results important from the viewpoint of intended physical applications. Readers interested in a more extensive and rigorous overview are directed to Refs. 20–26.

(ii) Those looking for a quick "preview" of the conclusions of this work may find it useful to read the beginning of Sec. V A, as well as Secs. VI and VII after the first perusal of the body of the paper. The purpose of this paper is not to advocate a complete replacement of the ensemble entropy with the algorithmic entropy—as Secs. II, III, and IV may appear to suggest to a casual reader, but rather to conclude that the physically relevant entropy should be defined from the internal viewpoint of the observer, and that it has components of both probabilistic and algorithmic origin. In a sense, the first half of the paper is an extended introduction to the algorithmic ingredient of the new definition of physical entropy. Its physical significance becomes clear only in Secs. V and VI.

II. ALGORITHMIC RANDOMNESS OF BINARY STRINGS

A natural measure of disorder or, equivalently, of the degree of randomness of a state of a system is the size of the smallest prescription required to specify it with some assumed accuracy. Ordered, regular states can be reconstructed from concise algorithms. Random states, by comparison, require many more bits of information to specify. In this very intuitive sense, the required amount of information quantifies the degree of randomness.

The strategy we shall adopt to measure the randomness of states of a physical system will involve a simple computer. A program for this computer will be considered a good description of the system if the output will contain sufficient information in some standard format to reconstruct—for instance, to plot—the state with the required accuracy. Positions and momenta of all the particles of an ideal gas are an example of such an output. Binary strings—for instance, in the form of numbers specifying these coordinates—are the key ingredient of such prescriptions. They can be also thought of as representing a state of a one-dimensional chain of spins and therefore as a direct description of an elementary example of a physical system. The aim of this section is to discuss a measure of randomness of binary strings known as the algorithmic information content, algorithmic randomness, algorithmic entropy, or, sometimes, as the algorithmic complexity.[16–29]

The algorithmic randomness $K(s)$ of a binary string s is defined as the length, in the number of digits, of the shortest program s^* that will produce output s and halt when used as input of a universal Turing machine T:

$$K(s) \equiv |s^*| \ . \tag{2.1}$$

A computer U is universal—as the Turing machine used in the above definition—if for any other computer C there is a prefix τ_C one can add to any program p so that $\tau_C p$ will execute the same computation on computer U as p alone did on C. Algorithmic randomness of a typical string s is to the leading order given by its length in bits, $|s|$:

$$K(s) \simeq |s| \ . \tag{2.2}$$

In other words, typical strings are random and cannot be generated from more concise programs. Small corrections to this estimate of a typical value $K(s)$ depend on the issues that will be considered below. When s is interpreted as a binary representation of an integer, Eq. (2.2) implies that $K(s) \simeq \log_2(s)$. We shall restrict ourselves to binary representations in the remainder of this paper.

While typical strings possess algorithmic information content comparable to their length, and are therefore random, there exist strings which are obviously algorithmically simple. We have already listed some examples from both categories in the Introduction. It is possible to define whole classes of strings which are either algorithmically random or simple. For instance, the reader can verify that minimal programs used to define algorithmic information content must be algorithmically random. Let us also note that the logarithm relating the information content of a typical string with the size of the integer it represents is suggestive of the "log" appearing in the probabilistic definition of entropy. To enumerate—and, hence, to specify—W distinct, equally probable states of a physical system one obviously needs numbers of order W. Hence the typical algorithmic entropy of such a state, crudely defined as the size of the number giving its "address," can be expected to be similar to the estimate obtained through the Boltzmann formula, Eq. (1.1). We can thus anticipate that the numerical estimates of the algorithmic randomness will be compatible with the more traditional probabilistic, ensemble calculations.

There are several points that must be clarified before algorithmic entropy can be accepted as an unambiguous measure of randomness. The first of them—potential dependence of the size of the program on the "addressee"—has been already largely bypassed by relying on a universal computer U. Use of different universal computers makes at most a difference bounded from above by a finite constant [that is, of order unity, $O(1)$]

in the size of the minimal program.[23]

The next difficulty arises from the somewhat counterintuitive fact: It is often possible to generate all the members of a certain set with a smaller algorithm than the minimal algorithm necessary to print out just one of them. To illustrate this fact by a simple and concrete example (we shall come back to more physically motivated examples later in the paper), we consider a program that can list all finite strings corresponding to natural numbers. A simple "counting loop," a Turing-machine version of the FORTRAN loop,

```
      N=0
1     PRINT N
      N=N+1
      GO TO 1
```

will do the job. Given sufficient time this program will generate every finite string s starting with 1. A slightly more sophisticated program can print out all the binary strings arranged in what is known as *lexicographic order.*

$$\Lambda,0,1,00,01,10,11,000,001,010,011,100,101,110,111,0000,0001,0010,\ldots . \tag{2.3}$$

Above, Λ stands for the empty string. Lexicographic ordering establishes a correspondence between all the strings and the set of natural numbers. We can either allow such strings to be a part of one long output tape—perhaps separated by commas—or we could add an instruction to erase the string (number) N before the string (number) $N+1$ is printed. This way, instead of a tape with all the strings we could have an erasable tape with strings appearing "one at a time." (The disadvantage of this arrangement is that unless we intervene, the numbers will all disappear.)

We have just demonstrated how one can use a concise program to list arbitrarily many strings each of which is supposed, typically, to have randomness given by its length, Eq. (2.2). Clearly, we are running into a contradiction. Unless one specifies in more detail the requirements a program must satisfy, we could encounter an uncomfortable upper limit given by the size of the binary version of the program constructing the lexicographic list, Eq. (2.3), on the entropy of any finite binary string.

One can dispose of the difficulty illustrated above in a few steps. To begin with, one can demand that all the programs halt after a finite number of steps. This excludes the program described above, but it does not settle the issue. One can modify the loop by forcing it to halt after some large but algorithmically simple upper limit NMAX (e.g., NMAX=111111111111...111). Now the apparent upper limit on the algorithmic entropy of all the numbers smaller than NMAX would be somewhat larger, but still far from reasonable. Suitably chosen NMAX can be encoded into a very compact subroutine.

The next remedy is to demand that after the computer halts, the output tape should contain nothing but the output string s. This condition obviously settles the issue raised above. It will be indispensable in the discussion of the algorithmic version of the second law in Appendix C.

Another requirement often imposed on the minimal programs is the demand that they be *prefix-free* or *self-delimiting.* Self-delimiting programs[21,28,29] carry within them information about their size: They allow the computer to "decide" when to stop reading the input tape. They are also known as "instantaneous" codes, since they can be executed alone, without any additional "prefixes" or "end markers." An additional benefit of using self-delimiting codes is the simplicity of employing them as subroutines. What is most important from the point of view of this paper is that the use of the self-delimiting codes allows one to formulate algorithmic information theory in a manner more closely analogous to the Shannon information theory.[21] We shall find this especially useful in Sec. IV, in the course of the discussion of the relationship between ensemble entropies and algorithmic randomness, where the connection between unique decodability of a sequence of bits regarded as a message and the self-delimiting property shall be further explored.

With the above considerations in mind we can now define the *joint algorithmic entropy* of two strings s,t as the size of the smallest self-delimiting program that makes U calculate both of them. Joint entropy satisfies the familiar inequality

$$K(s,t)\leq K(s)+K(t)+O(1) \tag{2.4}$$

only when the admissible programs are self-delimiting. The price one pays for having Eq. (2.4) comes in the form of a modification of the estimate of the typical value of algorithmic information of a single string:

$$K(s)\simeq |s|+O(\log_2|s|)+O(1)\simeq |s|+K(|s|)+O(1)\ . \tag{2.5}$$

The origin of this logarithmic correction to the first guess given by Eq. (2.2) is simple: The *self-delimiting* program used to reproduce s must contain the information about the number of digits on the input tape. To encode this, additional $\log_2|s|$ bits will be typically necessary.

Another formula modified by the requirement of self-delimiting programs is the relation for the conditional entropy of string s given t, $K(s|t)$. Defined as the size of the program needed to calculate s from the input t, it is connected with the joint information through the equation

$$K(s,t)=K(t)+K(s|t,K(t))+O(1)\ . \tag{2.6}$$

This differs somewhat from the "classical" Shannon information theory, where the conditional information satisfies $K(s,t)=K(t)+K(s|t)$.

Mutual information is a quantity that has obvious intuitive significance. It is defined both in the algorithmic and classical information theory through

$$K(s:t)=K(s)+K(t)-K(s,t)\ . \tag{2.7}$$

Its meaning is clear: It provides a measure of the in-

dependence of two strings. That is, it indicates how many more bits of information one needs to calculate s and t separately rather than jointly. Mutual information is symmetrical. Pairs of strings which have mutual information of zero are called algorithmically independent.

Most of the binary strings of a certain fixed length are algorithmically random $[K(s) \approx |s|]$ and independent $[K(s:t) \approx 0]$. However, there are obvious examples of strings that appear random but are in fact algorithmically simple $[K(s) \ll |s|]$. Binary representations of easily calculable numbers such as $\sqrt{2}$, e, π, etc., are algorithmically simple. Some of the algorithmically simple numbers are easy to spot. Consider, for example, a string that has a low density of randomly distributed 1's. A simple strategy designed to minimize program size is to encode sizes of the consecutive intervals of 0's. Readers are invited to demonstrate that this strategy will, in the limit of a very long string, lead to the estimate of algorithmic entropy equal to the Shannon entropy calculated on the basis of probability of 0's and 1's.

Chaitin has demonstrated that the impossibility of proving the randomness of a "random-looking" long string is a natural consequence of the algorithmic definition of randomness, and can be regarded as a manifestation of Gödel's theorem.[21,22] The proof can be paraphrased as follows: Suppose that a short algorithm can be used to demonstrate randomness of a string l much longer than this algorithm. One could then use such program as a subroutine of a somewhat larger program designed to generate such l. To accomplish this, one would include the algorithm in a loop generating proofs of randomness of integers in order of increasing proof size. If a proof of randomness of some integer considerably greater than the program were accomplished, the program would print that integer and halt. Thus having a short program prove that l was random would make it algorithmically nonrandom. This contradiction can be resolved only if we conclude that either the axioms used in the proof of randomness were inconsistent, or that the search for proofs of randomness would continue forever without any successful case in the domain of strings much longer than the program. Therefore it may be not just difficult to estimate the algorithmic entropy of a specific binary string; it will also often be impossible to distinguish random and nonrandom configurations. In Sec. III we shall show that in spite of this difficulty associated with calculating $K(s)$ exactly, the algorithmic viewpoint can lead to useful insights into the entropy of Boltzmann gas.

III. ALGORITHMIC ENTROPY OF A PHYSICAL SYSTEM: BOLTZMANN GAS

Consider a container with N particles of "gas" inside. These "atoms" of gas can be thought of as miniature "hard spheres" with no internal degrees of freedom, and with dimensions much smaller than the size of the container. In short, we are dealing with a "Boltzmann gas."[7]

Discussion in this section shall focus on the task of describing the system by means of the most compact "message." We shall imagine that the microstate of the system is known, and our task is to communicate it to someone else or to record it in some reproducible fashion. For definiteness, we assume in accord with Sec. II that the addressee is a certain specific Turing machine U. As the proof that the message was "understood" we shall accept the ability of the addressee to reproduce, by some reasonable means, the state of the system up to the specified accuracy.

To describe the microstate we adopt the following strategy. (i) We divide the volume occupied by the particles into a lattice of cells small compared with the separations of the particles so that the probability of finding two particles in the same cell is negligibly small. (This assumption of one particle per cell is convenient, but not essential.) (ii) We specify which of the cells contain particles and which of them are empty. The size of the cells determines the resolution with which the state of the system is given. More sophisticated strategies, which do not employ fixed grids, but, rather, explore all possible methods of encoding that satisfy some criterion for the accuracy of the description can be also employed. In essence, one is dealing with an issue analogous to those encountered in the domain of coding and information theory.[12,13]

Algorithmic randomness of the microstate is given by the length of the shortest computer program sufficient to "reproduce" its state with the requisite resolution. The form of the reproduction is somewhat arbitrary. For concreteness, one can imagine the following procedure which could be regarded as a substitute for "computer graphics." Suppose that a possible output of the universal Turing machine U is the multidimensional "plotter" with a discrete resolution. A state of the system will be considered "reproduced" if the machine will fill pixels of the plot space corresponding to empty cells with 0 and pixels corresponding to filled cells with 1.

Machines with access to several multidimensional tapes of this sort and other kinds of hardware [e.g., random-access memory (RAM) storage, etc.] are more convenient to use, and correspond more closely to the real-world computers, but can perform only these very same tasks as a one-tape universal Turing machine. Using them will not influence estimates of the algorithmic entropy. This last statement is true in general—it is responsible for the importance attached to Turing machines by the theory of computation—but is particularly easy to see for the "graphical" representation of the Boltzmann gas. A primitive, but for our purposes sufficient way to produce a plot would be to have U print out a tape with consecutive sections separated by commas and corresponding to rows in the phase space of the system partitioned into cells and organized in some definite order.

It is important to draw the distinction between the "binary image" of the gas microstate and the program that can generate it. An image provides a direct description of the state of the system. The program generating it contains the same information, but represented in a different, typically more concise manner. There are, generally, at least several programs that can generate the same plot—the same binary image. The length of the

shortest one will define the algorithmic randomness of the gas microstate.

The use of the analogy between the actual microstate of the many-body system—e.g., gas or fluid—and the corresponding binary image (spin lattice) is, of course, not new. It was extensively and successfully exploited in the investigation of phase transitions. Our treatment will focus on a different aspect of binary images. Nevertheless, previous successful applications are not unrelated, and further motivate our approach.

Evaluation of the algorithmic entropy of different microscopic configurations of a gas contained in a D-dimensional cube can be carried out with the help of a special-purpose computer L following a straightforward, but not necessarily optimal algorithm. L would first read in the key data (the number of dimensions D, the number of particles N). Subsequently, $N\times 2D$ phase-space coordinates of the individual particles would be generated in a double loop. For a typical (that is, algorithmically random) configuration individual coordinates would have to be generated by $N\times 2D$ separate "subroutines," which contain appropriate data. The first $2D$ numbers can be then used to print 1 in the appropriate empty square of the tape corresponding to the position and momentum of the first particle—that is, at the location

$$(x_1^{(1)}/\Delta_x, x_2^{(1)}/\Delta_x, \ldots, x_D^{(1)}/\Delta_x; p_1^{(1)}/\Delta_p, \ldots, p_D^{(1)}/\Delta_p) .$$

The process is repeated N times, after which the machine fills in the remaining blank spaces with 0's, and halts. For instance, when $D=2$, a binary image—a primitive plot—could begin by printing out a string of pixels of the plot filled in with the appropriate symbols corresponding to the first row: $x_1/\Delta_x=0$, $x_2/\Delta_x=0$, $p_1/\Delta_p=0$, which are kept constant, and the variable p_2/Δ_p changing from its minimal to its maximal value. The row with $x_1/\Delta_x=x_2/\Delta_x=0$, $p_1/\Delta_p=1$ follows, and so on, until all of the volume of the phase space of the gas in the volume of the container is mapped out. The size of the program for the special-purpose machine L can now be reported as the estimate (and, almost certainly, an overestimate) of the algorithmic randomness of the state of the gas. The additional information that needs to be supplied to a universal computer U can be clearly encoded in the form of a compact binary string.

When the configuration of the system is nonrandom (for example, when the particles are placed on a regular lattice), the size of the program can be decreased by generating their coordinates by means of a simple subroutine. Hence we can expect to shorten programs generating plots of less random configurations of Boltzmann gas.

The measure of randomness suggested above is very different from the one based on ensembles. In particular, it bypasses the concept of probability, and can be used to define the entropy of a completely specified microscopic state of an individual dynamical system. The question therefore arises as to whether the value of the algorithmic entropy is related to the value of the statistical or thermodynamic entropy. This issue will be considered throughout the course of this paper. In Appendix A we implement the strategy outlined above to estimate algorithmic randomness of a typical configuration of the classical Boltzmann gas and explore one of its aspects—indistinguishability—in more detail. It is demonstrated there that the algorithmic randomness of a typical microstate of the gas of N indistinguishable particles in D dimensions confined to the volume V is

$$K=N\left[\log_2\frac{V}{N\Delta V}+\frac{D}{2}\log_2\frac{mkT}{(\Delta_p)^2}\right]+O(1) . \qquad (3.1)$$

Here k is the Boltzmann constant, T the temperature, while $\Delta V=(\Delta_x)^D$ and Δ_p define the resolution in the configuration and momentum halves of the phase space.

Equation (3.1) is known as the Sackur-Tetrode equation for the entropy of the ideal gas. A shortcut derivation of the Sackur-Tetrode equation would proceed as follows. There are $\Omega=\mathcal{C}^N/N!$ distinct ways of distributing N indistinguishable particles of gas among the available $\mathcal{C}\approx(V/\Delta V)(\sqrt{mkT}/\Delta_p)^D$ cells in the phase space. One way to specify a certain configuration it to give its "address" among the Ω possibilities. The typical value of the address numeral is then $\sim\Omega$. Hence its binary specification requires $\sim\log_2\Omega$ bits. The Sackur-Tetrode equation (3.1) follows.

This second argument yields the correct answer. Moreover, it establishes an intimate connection with Boltzmann's approach to the entropy of a microcanonical ensemble of ideal gas, as it is directly equivalent to the "counting of the complexions." Furthermore, the approach based on the size of the address can be easily adopted for an arbitrary microcanonical ensemble. Yet it follows a philosophy that is quite different from the explicitly descriptive strategy introduced earlier in this section and implemented in Appendix A. Algorithmic randomness is, of course, given by the shortest program that generates the description of the state in question as an output. However, the fact that these two very different programming strategies result in an approximately identical answer is encouraging, as it implies that, in spite of undecidability, finding a reasonable estimate of algorithmic randomness of a typical state may not be that difficult.

The value of the statistical entropy of a classical system depends on the "resolution"—the volume of the cells in the phase space corresponding to distinct microstates—as well as on the structure of the "grid" these cells form. Both of these issues, the dependence on the resolution as well as the subjectivity associated with the "coarse graining," have their algorithmic counterparts. For the purpose of this paper it would have been sufficient to adopt the attitude that both of these features of the grid are determined by the observer: The measurements performed by the observer define a certain "natural grid," which in turn specifies the value of the algorithmic randomness in the state of the system. One might, therefore, expect that the algorithmic randomness of a physical system is as highly subjective as statistical entropy. As discussed in more detail in Appendix B, this remark applies without qualifications only to the dependence on the resolution. Quantum mechanics must be invoked to fix the volume of the cells in the phase space and thus to assure that the entropy is finite. There is, however, a strategy

that removes some of the subjectivity associated with the shape of the grid: In the intuitive definition of the algorithmic randomness as the size of the message it is natural to demand that description of the coarse-graining must be included as a part of the message. As demonstrated in more detail in Appendix B, this strategy relegates much of subjectivity to the differences between the distinct Turing machines. This, in turn, allows such residual subjectivity to be quantified by employing algorithmic methods.

IV. ENSEMBLE ESTIMATES OF ALGORITHMIC RANDOMNESS

The purpose of this section is to investigate the relation between the algorithmic and statistical measures of disorder. Physics has traditionally employed entropy in two distinct roles: (i) as a measure of irreversibility and (ii) as an equilibrium thermodynamic potential.

The use of entropy in the formulation of the second law and the conflict between the irreversibility of thermodynamics and reversibility of the underlying dynamics remains a leading theme of statistical mechanics. Indeed, the observation that in the closed system $\mathrm{Tr}\rho \ln\rho$ is a constant of motion has been (and remains) one of the key reasons for dissatisfaction with the traditional definitions of entropy. One might be concerned that algorithmic definitions of entropy will automatically inherit this troublesome feature. In particular, one might argue that, since the time-evolved state of the system is obtained from the initial state by the action of the Hamiltonian, the randomness of all the states along the trajectory should not exceed the sum of the algorithmic information content of the initial state and of the Hamiltonian. Therefore one would be forced to conclude that all the descendant states of an algorithmically simple state must be algorithmically simple. I demonstrate in Appendix C that this conclusion is incorrect: The typical algorithmic randomness of a microstate— even if it is descended from a simple initial condition—is consistent with Boltzmann's entropy of the corresponding microcanonical ensemble.

One way of understanding this is to recognize that the labeling strategy employed in the derivation of Eq. (3.1) in Sec. III is automatically accomplished in the course of a dynamical evolution of an ergodic system: There the ordered sequence of microstates is generated by the Hamiltonian. Time—the duration of the evolution from some initial state—can be regarded as a label of a microstate. In a perfectly ergodic system all of the microstates are traversed before the system returns (having completed its Poincaré cycle $\mathcal{P}$) to the initial configuration. When this initial state is chosen to be algorithmically simple, the leading contribution to the algorithmic randomness of a typical descendant is $\approx \log_2 \mathcal{P}$, which leads one to the correct answer. As described in more detail in Appendix C, the requirement of a specific output—corresponding to a description of a single, definite microstate—at the completion of the computation is crucial in arriving at the correct algorithmic estimate of the randomness of the time-evolved state. Indeed, for reasons analogous to these discussed in connection with Eq. (2.2), a program that would list all the microstates along the trajectory would be more concise than the program that generates a description of a definite, typical microstate.

Irreversibility is of obvious interest as a separate issue. Therefore we shall leave a more detailed investigation of the algorithmic approach to irreversibility for the sequel of this paper. By contrast, the relation between algorithmic randomness and Boltzmann-Gibbs-Shannon entropy considered in the remainder of this section is essential for further discussion of physical entropy and is the true focus of this section.

A result of Sec. III—the algorithmic derivation of the Sackur-Tetrode equation for the entropy of ideal gas—was a forerunner of the conclusion we shall reach: We shall demonstrate that, in general, probabilistic considerations result in accurate estimates of the average algorithmic entropy of a member of an ensemble. In spite of this conclusion, we shall go on in the next section to suggest that the physical entropy, the true measure of the amount of energy extractable from the system, contains both algorithmic and missing information contributions.

A. Entropy of thermodynamic ensembles

Consider a statistical ensemble $\mathcal{E}$, a collection of microstates $\{s_k\}$ that occurs with probabilities p_k. The statistical entropy of this ensemble is given by the formula

$$H(\mathcal{E}) = \sum_{s_k \in \mathcal{E}} p_k \log_2 \frac{1}{p_k} \ . \tag{4.1}$$

In statistical mechanics Eq. (4.1) plays a key role in deriving thermodynamic characteristics of a system from microphysics. In this subsection we shall discuss ensembles that are defined concisely. That is, we shall assume that there exists a concise algorithm ε which, given sufficient time, will generate as its output a description of every relevant microstate of $\mathcal{E}$ and compute its probability with some (arbitrarily high, but finite) predetermined accuracy. As $\mathcal{E}$ may contain infinitely many microstates, we shall not demand of ε to halt. Instead, we shall require that the microstates listed as the output be "weakly sorted" according to their probabilities in the following sense: For every arbitrarily small but finite $\delta > 0$ there should exist a finite time (number of steps) $\mathcal{N}_\delta$ after which descriptions of all the states of $\mathcal{E}$ with the probabilities $p_k > \delta$ must be listed. The length of the smallest program ε^* which satisfies these criteria will be regarded as the algorithmic information content of the ensemble $\mathcal{E}$. The inequality

$$K(\mathcal{E}) \equiv |\varepsilon^*| \lll H(\mathcal{E}) \tag{4.2}$$

will define, for the purpose of this paper, *thermodynamic ensembles*. In the context of statistical mechanics, thermodynamic ensembles are determined by macroscopic—which we take to mean "simply describable"—constraints and the probabilities are easily computable functions of the microstates. This motivates our focus on thermodynamic ensembles. They will allow us to show that the concept of entropy in equilibrium thermodynamics can be based on the algorithmic foundation. However, as we shall discuss at some length in Sec. V, these are not the only ensembles possible. The three kinds of en-

sembles widely used in statistical mechanics—microcanonical, canonical, and grand canonical—are clearly "thermodynamic" for sufficiently large systems in the sense of the defining Eq. (4.2).

Here we shall demonstrate that the average algorithmic entropy of member of a thermodynamic ensemble

$$\langle K \rangle_{\mathcal{E}} = \sum_{s_k \in \mathcal{E}} p_k K(s_k) \tag{4.3}$$

is closely approximated by the corresponding Shannon entropy, Eq. (4.1):

$$\langle K \rangle_{\mathcal{E}} \cong \sum_{s_k \in \mathcal{E}} p_k \log_2 \frac{1}{p_k} \ . \tag{4.4}$$

In the context of the algorithmic information theory this connection between $\langle K \rangle_{\mathcal{E}}$ and $H(\mathcal{E})$ was first established by Kolmogorov, Levin, and Zvonkin (see Ref. 24) and was extended and strengthened—using advantages of the definition of algorithmic randomness by means of self-delimiting programs—by Chaitin (see especially Theorem 3.4, as well as 3.2 and 3.5 in Ref. 21). Bennett[27] has pointed out the physical significance of this result for the thermodynamic case defined by Eq. (4.2). Below, I shall present a new, physically motivated proof of the key inequalities and—in Sec. V—apply it to the situations involving measurements that are sufficiently detailed to invalidate Eq. (4.2).

The line of argument I shall follow to investigate the relationship between $H(\mathcal{E})$, $\langle K \rangle_{\mathcal{E}}$ and $K(\mathcal{E})$ is borrowed from the coding theory (see, e.g., Hamming[13]). One of the fundamental problems of coding theory is the search for efficient ways of representing symbols $s_1, s_2, \ldots, s_k, \ldots$, put out by a source of information with the respective probabilities $p_1, p_2, \ldots, p_k, \ldots$, in terms of an alphabet consisting of another set of symbols (e.g., 0,1) for the purpose of convenient storage or transmission. We shall be particularly interested in the case when the code words $\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_k, \ldots$, corresponding to the source symbols are composed from a binary alphabet.

A natural measure of the efficiency of encoding is the average length (the number of binary symbols per source state) of the encoded message

$$\mathcal{L} = \langle |\tilde{s}| \rangle = \sum_i p_i l_i \ , \tag{4.5}$$

where

$$l_i = |\tilde{s}_i| \ . \tag{4.6}$$

The encoded message should be uniquely decodable. That is, a sequence of 0's and 1's, which constitutes an encoded message $\tilde{s}_i \tilde{s}_j \tilde{s}_k \ldots$, should correspond to a single, unique sequence of source symbols. The best way to accomplish this is to use the so-called "instantaneous" or "prefix-free" coding. In the instantaneous code no code word $\tilde{s}_k$ is the "prefix" (the first part) of any other code word. For instance, the encoding

$$s_1 \to 0, \quad s_2 \to 1, \quad s_3 \to 00, \quad s_4 \to 11$$

is not uniquely decodable. An encoded message 00111 can be decoded as $s_1 s_1 s_2 s_4$ or $s_3 s_4 s_2$ or in a few more ways. By contrast,

$$s_1 \to 0, \quad s_2 \to 01, \quad s_3 \to 011, \quad s_4 \to 111$$

is uniquely decodable but not instantaneous. The whole message 00111 has only one decoding ($s_1 s_1 s_4$). However, it must be received *in toto* before unique decoding is possible (e.g., the first three digits "read" $s_1 s_2$).

This necessity of requiring the whole message to decode it is avoided by an instantaneous code

$$s_1 \to 0, \quad s_2 \to 10, \quad s_3 \to 110, \quad s_4 \to 111 \ .$$

With this encoding symbols can be decoded as soon as they are received.

The analogy that will be pursued in this section regards an ensemble $\mathcal{E}$ as a "source" of individual microstates which play a role of source symbols. These microstates occur in the ensemble with frequencies proportional to their probabilities p_k. The most efficient encoding will minimize the average length

$$\langle K \rangle_{\mathcal{E}} = \sum_k p_k K(s_k) \tag{4.7}$$

of the programs which can be used to record (or communicate) individual microstates.

The prefix condition corresponds to the requirement for the program to be self-delimiting. A self-delimiting program must carry within it the information about its size. Hence it will be able to initiate the computation without having to be prompted by some additional symbol (e.g., ","). In this sense it is, therefore, "instantaneous."

The real advantage of self-delimiting codes follows from the relation between their sizes and the probabilities that they will be generated by some random process, such as the flipping of an unbiased coin.[21,26] The probability that a specific program of length l will be generated is obviously 2^{-l}. The sum of all such weights—the total probability that a valid, halting program will be obtained by flipping a coin—cannot be more than 1. This in turn implies that sizes of self-delimiting programs as well as the lengths of the encoded words $l_i = |\tilde{s}_i|$ are constrained by the *Kraft inequality*. For the encoding to be uniquely decodable l_i must satisfy

$$\sum_i 2^{-l_i} \leq 1 \ . \tag{4.8}$$

The converse of this theorem is also true; that is, whenever the sizes of the words do satisfy the Kraft inequality, an instantaneous code $\tilde{s}_i$ with word sizes l_i exists.

Shannon-Fano coding is a specific, efficient implementation of the coding which satisfies the Kraft inequality. The length of the words is chosen to satisfy

$$l_k = |\tilde{s}_k| = \lceil -\log_2 p_k \rceil \ , \tag{4.9}$$

where $\lceil a \rceil$ is defined as the smallest natural number equal or greater than a. The length l_k of a word $\tilde{s}_k$ is therefore bounded from both above and below:

$$-\log_2 p_k \leq |\tilde{s}_k| < -\log_2 p_k + 1 \ . \tag{4.10}$$

Taking the average over the whole ensemble one arrives at

$$-\sum_k p_k \log_2 p_k \leq \sum_i p_i l_k < -\sum_k p_k \log_2 p_k + 1 \ . \tag{4.11}$$

The lower bound on the average size of the message follows directly from the Kraft inequality: The concavity of the log function yields immediately that

$$\sum_i p_k \left[\log_2 \frac{1}{p_k} - l_k \right] \leq \log_2 \left[\sum_k \frac{p_k 2^{-l_k}}{p_k} \right] \leq 0 \ . \tag{4.12}$$

The upper bound exists by virtue of the specific (Shannon-Fano) encoding strategy.

Shannon-Fano coding is not optimal: A recursive strategy known as Huffman coding can be shown to be optimal. To implement Huffman coding one starts with the two least probable states and assigns them code words that differ in only one (last) digit. The combination of these two states is thereafter treated as a single new state. Within this redefined list of states, the two least probable are found and assigned the last yet unassigned digit. The procedure is repeated until the list of redefined states contains only two states, which are assigned digits 0 and 1. In this way each state is assigned an instantaneous sequence of binary digits, its code word.[13]

Huffman coding must, of course, satisfy the left-hand side of inequality (4.11), since it follows directly from the Kraft inequality, which in turn follows from the requirement of unique decodability. Furthermore, since it is even more concise than the Shannon-Fano coding, it must at least obey the same upper bound on the average message size. Indeed, the best upper bound in terms of the Shannon entropy of the source is—even for Huffman coding—given by the right-hand side of Eq. (4.11).

The conclusion of the above discussion is, therefore, that the average length of the messages of our instantaneous code can be made very close to the entropy of the source. Below we shall derive a version of the inequality (4.11) valid for self-delimiting programs.

Our strategy is based on the definition of the ensemble: We have assumed the existence of the concise program ε which, given enough time, will generate the output consisting of "weakly sorted" descriptions of microstates s_k belonging to $\mathcal{E}$ and will compute their respective probabilities p_k with the requisite (finite, but otherwise arbitrary) accuracy. This program can be now used as a subroutine which, in addition to (i) ensemble description ε, contains also (ii) a sorting routine σ which (a) arranges descriptions in the order of decreasing computed probabilities (so that s_i appears before s_j if and only if $p_i \geq p_j$) and (b) for microstates that have, within errors, the same computed probabilities, some other definite order (e.g., lexicographic) is adopted; and (iii) a program c which uses the ordered list containing probabilities p_k to assign each microstate s_k a definite binary "code word" $\bar{s}_k$ using either the Shannon-Fano or the Huffman procedure.

Shannon-Fano coding can be illustrated in a particularly straightforward manner (see Fig. 1). To set up a definite correspondence between states s_k and sequences of 0's and 1's we consider the sum

$$\chi_k = \sum_{i=1}^{k} 2^{-\lfloor \log_2(1/p_i) \rfloor} = \sum_{i=1}^{k} 2^{-l_i} \ . \tag{4.13a}$$

For definite s_k the first l_k binary digits of χ_k constitute an instantaneous encoding of s_k. (We have, of course, assumed that the states are presorted in the order of decreasing probabilities.) The consecutive digits $\alpha_1, \alpha_2, \ldots,$ of the code word can be then read off directly from the binary expansion of χ_k

$$\chi_k = \alpha_1^{(k)} 2^{-1} + \alpha_2^{(k)} 2^{-2} + \cdots + \alpha_{l_k}^{(k)} 2^{-l_k} \ . \tag{4.13b}$$

It is perhaps worth emphasizing that even when the initial digits are 0 they should not be omitted in constructing the code word $\bar{s}_k \equiv \alpha_1^{(k)} \alpha_2(k) \ldots \alpha_{l_k}^{(k)}$.

Let us now return to the derivation of the analog of inequality (4.11) for the average algorithmic randomness of an ensemble $\mathcal{E}$. The complete program capable of uniquely reproducing any microstate s_k of the ensemble must, by virtue of the Kraft inequality, satisfy

$$H(\mathcal{E}) \leq \sum_i p_i K(s_i) \ . \tag{4.14}$$

The upper bound of the average algorithmic randomness follows from the inequality

$$K(s_i) < |\bar{s}_i| + K(\mathcal{E}) + K(\sigma, c) + O(1) \ . \tag{4.15}$$

Here $|\bar{s}_i|$ is the length of the code words, $K(\mathcal{E})$ is the length of the minimal program to generate descriptions and probabilities of the ensemble microstates, $K(\sigma,c)$ is the joint algorithmic complexity of the sorting and decoding algorithms, and $O(1)$ is the usual constant associated with the choice of the universal computer.

For Shannon-Fano coding $|\bar{s}_i| \leq -\log_2 p_i + 1$. Moreover, $K(\sigma,c)$ does not depend on $\mathcal{E}$ and therefore can be incorporated in the constant term $O(1)$ along with the extra "$+1$." Multiplying the inequality $K(s_i) < -\log_2 p_i + K(\mathcal{E}) + O(1)$ by p_i and summing over the whole ensemble, we arrive at

$$\sum_i p_i K(s_i) < H(\mathcal{E}) + K(\mathcal{E}) + O(1) \ . \tag{4.16}$$

We have therefore established the following.

Theorem 4.1. The ensemble average of the algorithmic randomness of microstates belonging to the ensemble $\mathcal{E}$ is bounded from below by the statistical entropy of that ensemble and from above by the sum of the Shannon entropy and algorithmic information content of $\mathcal{E}$:

$$H(\mathcal{E}) \leq \langle K(s_i) \rangle_{\mathcal{E}} < H(\mathcal{E}) + K(\mathcal{E}) + O(1) \ . \tag{4.17}$$

For thermodynamic ensembles $K(\mathcal{E}) \lll H(\mathcal{E})$, which implies that the relative difference between $K(\mathcal{E})$ and $H(\mathcal{E})$ is negligible.

An important issue, omitted in the discussion above, concerns the accuracy with which the probabilities of the ensemble microstates have to be computed in order for Theorem 4.1 to apply. Problems would emerge if one

were required to attain infinite accuracy in calculating probabilities of states. Fortunately, it is not difficult to see that the inequality (4.17) will be satisfied as long as the generated probability distribution q_k is a faithful approximation of the actual distribution given by p_k. One way of phrasing this requirement is to demand that the expression

$$\Delta = \sum_k p_k \log_2 \frac{q_k}{p_k} \tag{4.18}$$

be of order unity.[27]

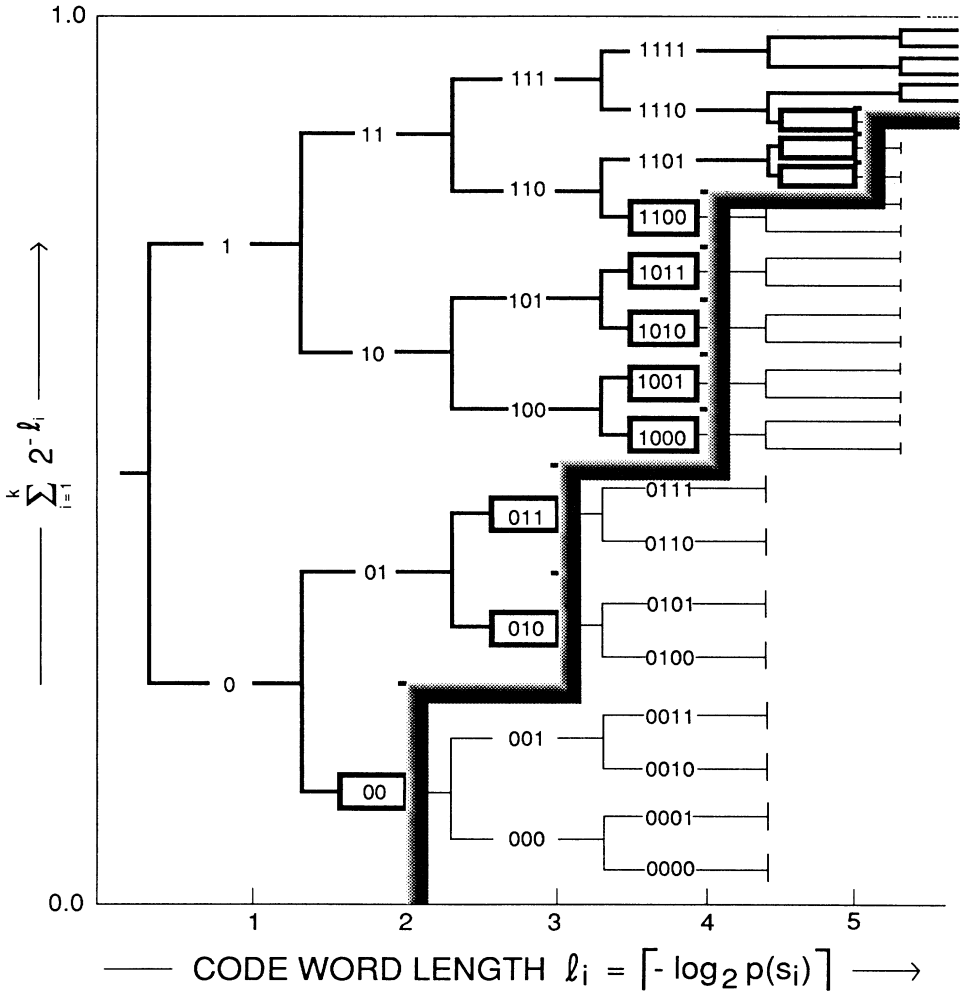

FIG. 1. Efficient instantaneous encoding of the states $\{s_i\}$ with probabilities $p(s_i)$ is illustrated above with the help of the lexicographic tree. The tree is plotted in a coordinate system, with the vertical axis corresponding to the cumulative probability and the horizontal axis labeled with the index of the "level" l of the tree (there are 2^l distinct branches at the level l). This establishes a correspondence between each l-digit-long binary sequence and a 2^{-l} section of the [0,1) interval. Determination of the code word corresponding to the state s_i begins with arranging all the states in the order of decreasing probabilities $[p(s_1) \geq p(s_2) \geq p(s_3) \geq \cdots]$. Code words are assigned in this order. The code word corresponding to the state s_k with probability $p(s_k)$ is picked out from the l_kth level of the tree, where $l_k = [\log_2[1/p(s_k)]]$. The actual sequence of digits assigned to s_k corresponds to the first still available branch on the level l_k of the tree. If a certain binary string is chosen as the code word (this is indicated by "framing" it in the figure), the corresponding branch and all its "descendants" at levels greater than (to the right of) l_k are "cut off." The tree is successively "pruned," and its whole sections contained in the intervals $[\sum_{i=1}^{k-1} 2^{-l_i}, \sum_{i=1}^{k} 2^{-l_i})$ (indicated by "ticks" on the shaded "pruning border") become unusable. This guarantees prefix-free coding. There will always be unassigned branches to encode states with the nonincreasing probabilities $p(s_i)$ with the code words of the length $[\log[1/p(s_i)]]$. This is apparent both from the figure and from the fact that $2^{-l_i} \leq p(s_i)$, which in turn guarantees that the Kraft inequality, Eq. (4.8), is satisfied. In fact, the proof of the Kraft inequality utilizes the lexicographic tree in this manner. The method shown in this figure implements Shannon-Fano coding, which is not optimal, but which suffices to establish the upper bound in the inequality (4.17).

V. PHYSICAL ENTROPY = MISSING INFORMATION + KNOWN RANDOMNESS

Section IV has demonstrated that the consequences of the radical shift of the paradigm used for the definition of entropy of thermodynamic ensembles would be, from the point of view of its quantitative estimates, barely noticeable. The average values of the algorithmic entropy and of the Gibbs-Shannon entropy of any concisely described ensemble are virtually identical. Indeed, in equilibrium all the sensible definitions of entropy must approximately coincide. Otherwise, given the excellent record of the Boltzmann-Gibbs approach, they would be experimentally ruled out.

Conceptual implications of the new definition of entropy are, on the other hand, very different from either Boltzmann or Gibbs ensemble definitions. Ensemble entropies bear an unmistakable formal similarity to Shannon's information-theoretic entropy. Indeed, already Maxwell and Boltzmann expressed the suspicion that entropy is a measure of ignorance.[1,7] A more direct connection between missing information and entropy was established by Szilard in his discussion of Maxwell's demon.[30]

Particularly influential works arguing in favor of the identification of entropy with missing information are due to Jaynes[31] and Brillouin.[32] Yet all but a few of the most fervent supporters of the information-theoretic interpretation will agree with the notion that individual states of physical systems can be either ordered or random regardless of our information about them. A broken glass—a frequent example of the increase of disorder in movies illustrating the thermodynamic arrow of time—is more random than the unbroken one not because our ignorance about it has dramatically increased when it was shattered by the hammer, but because it has become more disordered.

A. Definition of physical entropy

The intuitive notion of randomness is, we believe, formally expressed by the algorithmic definition of entropy. We have demonstrated that equilibrium systems are expected to have the same equilibrium thermodynamic properties regardless of which of the two paradigms is employed to define their entropy. Therefore we have the "objective" quantity—algorithmic randomness—which measures the disorder in the system. This approach is consistent and well defined for the systems considered by equilibrium statistical mechanics, but it does not suffice to give a full account of the operation of idealized thermodynamic engines. For, consider gas entering an appropriate chamber in the engine. Its initial state, from the standpoint of the operation of the engine, is defined by the few relevant macroscopic properties (pressure, volume of the chamber, temperature) which can be used to determine its entropy, and which can, in turn, be employed in the calculations of the amount of the work that can be extracted. The fact that the gas may be in some specific microstate does not enter into consideration at all. Indeed, the efficiency of the engine is largely determined by the fact that its design *ignores* all but a few *macroscopic* characteristics of the microstate. The work extracted is then limited not so much by the details of the microscopic initial state, but by the combination of these few usually macroscopic properties of that state that are "encoded" in the design of the engine—in the "algorithm" it follows. If this algorithm is unable to take advantage of all the features of the initial state of the gas, then it will consistently miss the additional opportunities to extract all of the extractable energy.

A more sophisticated engine designed to miss fewer such opportunities could be operated by an IGUS (information gathering and using system), which, following each measurement, could perform a computation in an attempt to optimize the strategy for each individual case on the basis of the acquired data d. Following the measurement, with a definite outcome at hand, the entity operating the engine could also assess its ability to extract useful work. The question therefore arises: What quantity should it employ in the calculation of the net useful energy?

Motivated by such considerations, I conjecture that the quantity relevant for this purpose is the *physical entropy* $\mathcal{S}$. It is a sum of two contributions: (i) The algorithmic randomness $K(d)$ given by the size of the most concise description of the already available relevant data d, and (ii) the information about the actual microstate which is still missing, in spite of the availability of d, as measured by the Shannon conditional entropy

$$H_d = -\sum_k p_{k|d} \log_2 p_{k|d} \ ,$$

where $p_{k|d}$ is the conditional probability of the state s_k given d.

Definition. Physical entropy is the sum of the missing information and of the length of the most concise record expressing the information already at hand:

$$\mathcal{S}_d = H_d + K(d) \ . \tag{5.1}$$

It is apparent that $\mathcal{S}_d$ is a good thermodynamic potential. For, by the results of Sec. IV, $\mathcal{S}_d \cong \sum_k p_{k|d} K(s_k|d)$. Therefore, for systems in thermodynamic equilibrium, the ensemble average of $\mathcal{S}_d$ approximates the premeasurement ensemble entropy (that is, the entropy computed in absence of d). Below, we shall prove that the quantity defined by Eq. (5.1) enters into the formulation of thermodynamics for engines operated by entities that can acquire information through measurements and can process it in a manner analogous to Turing machines. We shall also drop the subscript d on $\mathcal{S}_d$ to simplify the notation.

The physical significance of the above formula is best understood in the example in which a sequence of measurements is being carried out on a system (see Fig. 2). Measurements change conditional probabilities $p_{k|d}$ of the microstates; as a result, H_d decreases. On the other hand, the "record tape" containing measurement outcomes is getting more and more data about the system. If these measurements are performed on an algorithmically random member of an equilibrium ensemble, then, by virtue of the results discussed in Sec. IV, the measurement will increase the size of the record by the amount

almost exactly equal to the increase of the statistical information—decrease of Shannon's entropy—about its state. This must be so, as, according to Eq. (4.17), the ensemble average of the sum $H_d+K(d)$ (that is, $\sum_d p_d[H_d+K(d)]$) is bounded from below by (and in fact, approximately equal to) the average $\sum_k p_k K(k)$, which, for an equilibrium thermodynamic ensemble $\mathcal{E}$, is approximately given by the algorithmic entropy of its typical member. Hence

$$\langle \mathcal{S}_d \rangle_d \simeq \langle K \rangle_{\mathcal{E}} \simeq H(\mathcal{E}) .$$

In the limit when the measurements are successful, and the microstate known precisely, $p_k=\delta_{kk'}$, the physical entropy of the system is given by the algorithmic randomness of the state in which the system is found.

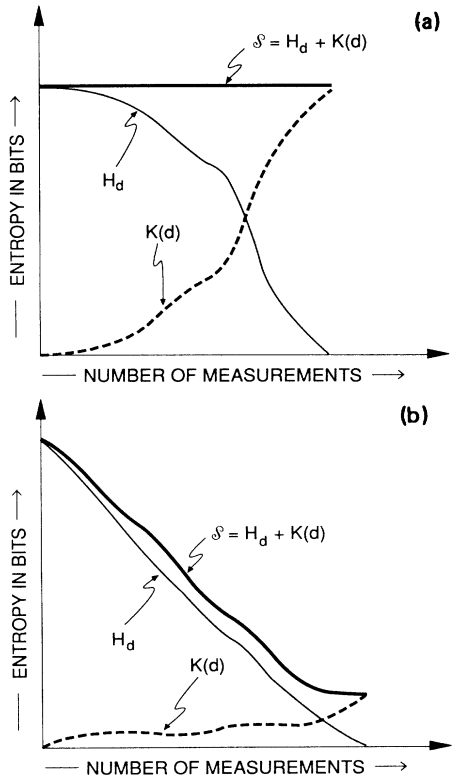

FIG. 2. Schematic illustration of the effect of data acquisition by measurements on (i) H_d, the information still missing in presence of the data d; (ii) $K(d)$, the algorithmic information content of the data; and (iii) the physical entropy $\mathcal{S}=H_d+K(d)$, observers internal measure of net work that can be extracted from the system. (a) For a typical (algorithmically random) microstate of the system the size $K(d)$ of the minimal record will increase by the amount equal to the decrease in H_d. Hence the physical entropy $\mathcal{S}$ will be constant, unchanged by the measurements. (b) For a regular microstate, the increase of the minimal size of the record describing the acquired data will be less than the decrease of the still missing information H_d. Consequently, the value of physical entropy can really decrease as a result of the measurement, and the observer can gain the ability to extract useful energy from the system.

Consider now a sequence of measurements of a "regular" (that is, simply describable) microstate. Before the measurements are carried out, the physical entropy is dominated by the missing information. As the measurements are carried out, the information about the state of the system is stored in the form of a concise program record. The size of such a record will be smaller in bits than the decrease of the missing information. In the end, the minimal program record, equal in the number of bits to the algorithmic entropy of the state, can be used to store the information. However, now this implies a relatively short record, as the state, by assumption, was algorithmically simple.

Physical entropy defined by Eq. (5.1) can be really decreased by a gain of information. When the state of the system is regular, the information about it can be recorded concisely, its binary image can be generated from a short program. Moreover, the length of the record need not be a monotonic function of the decrease in ignorance. Often it may become possible to compress the record only after it contains sufficient evidence of the underlying regular pattern.

Indeed, one can claim that the measurements are usually devised "in anticipation" of regularities. Approximate measurements of macroscopic properties performed by our senses are particularly adept at finding familiar patterns (see, e.g., Ref. 33 for an interesting discussion of visual perception in the computational context). Most of the information reaches our consciousness already preprocessed, with the records relatively concise compared with the number of hypothetical possibilities. Analyzing this point would take us, however, into a discussion of the theory of perception. While this is a fascinating subject, one can demonstrate how the record can be compressed by other, more rigorous and less anthropocentric means.

Information processing by means of computers can accomplish the goal of compressing the record. To demonstrate this, we consider a record r which can be generated from a more concise program r', by a universal computer (or by a special purpose computer attached to the measuring device). We also assume that this computer can operate reversibly.

The proof that computers can process information reversibly is due to Bennett.[34] Landauer[35] had earlier argued that only logically irreversible operations lead to loss of information and increase of entropy. Bennett took the next step to show that reversibility can be accomplished by storing the information normally erased in such irreversible operations, so that the computation can "backtrack" from the output to the input along the same logical route. A good summary of the subject can be found in recent reviews.[27,36,37]

To shorten the record without any thermodynamic cost one can use a reversible computer to replace r with r'.[34] Here is one strategy of accomplishing compression: The computer first uses r', which, we assume, it already contains, to generate one more copy of r. The newly generated r can be used to cancel the "old" but identical r. Reversible cancellation leaves one r behind.[27] The remaining r can be finally converted reversibly into r' by

running the program backwards—it has now all the "directions" to backtrack reversibly to r'.

As a result of all these reversible operations, the computer will be left with r' and the rest of the register previously occupied by r will be filled with $|r|-|r'|$ 0's. Our discussion demonstrates an important point: If the record r' is known to encode the same information as r more economically—that is, using fewer bits—the substitution of r with r' can be, in principle, used to "compress" information stored in computer memory at no cost in work or entropy increase.

It is important to emphasize that the above argument demonstrates the possibility of compression only when a more concise r' is *known* "by fiat" to generate r, the substitution can be accomplished reversibly. However, in general, as it follows from the undecidable aspects of the algorithmic information, it may be difficult if not impossible to determine whether a string r has a more concise description. Nevertheless, the ability to recognize concisely describable configurations will be of advantage in the task of decreasing physical entropy.

B. Physical entropy and the demon's version of thermodynamics

To demonstrate that the use of the term "physical entropy" for the expression defining $\mathcal{S}$, Eq. (5.1), is legitimate, consider a transition taking the system from the initial state s_i to some final state s_f. We shall assume that this transition is accomplished sufficiently slowly to be thermodynamically reversible. To simplify the analysis, we shall also assume that the internal energy of the system is constant, so that the work extracted in the course of the transition is given by

$$\Delta W = T\Delta S \ . \tag{5.2}$$

Here T is the temperature, Boltzmann's constant $k_B=1$, and ΔS is the difference between the entropy of the final and initial state

$$\Delta S = S_f - S_i \ . \tag{5.3}$$

Equation (5.2) defines, for the purpose of this argument, the quantity that should be regarded as physical entropy. We shall demonstrate that when this transition $(s_i \to s_f)$ is controlled by an automated "demon"—a computer that can distinguish between different initial states and which maintains a current record of the state of the system—the physical entropy S must be given by Eq. (5.1). In short, we shall show that S which must be used in Eq. (5.2) is the physical entropy $\mathcal{S}$.

This arrangement is clearly inspired by the "Szilard's engine," in which the demon controls the operation, altering the course of the cycle depending on the outcome of the measurement, on the information it acquires about the new initial state of the system.[30,38,39] The initial state of the system is a single particle of gas located in either the left or the right chamber. The demon's memory contains the corresponding one-bit record. The final state—in the sense of our analysis—corresponds to the particle somewhere in the container. The demon must accordingly "reset" its memory to be able to initiate a new cycle. This "resetting"—as argued by Bennett[27,38] for the classical version of the Szilard engine— is thermodynamically costly and essential in preventing the demon from violating the second law. Indeed, the thermodynamic cost of resetting was, to some extent, anticipated by Szilard,[30] who, nevertheless, in other parts of his paper, viewed measurement rather than erasure as the fundamental thermodynamically costly operation. Bennett's argument was based on the observation by Landauer,[35] who, in the context of information processing operations, pointed out that the thermodynamic cost of the removal of information about the past states of the computer is $k_B T$ per bit. This conclusion was extended for the quantum version of the Szilard engine by the present author.[39]

In the general case considered here, we shall also insist that the final state of the computer memory should correctly reflect the demon's knowledge (or lack of it) about the state of the system. With this assumption in mind, we can now calculate the net work obtained by our computer-operated engine. The gain due to the change of Gibbs-Shannon entropy is given by the usual formula in a somewhat unusual notation:

$$\Delta W^{(+)} = T(H_f - H_i) \ . \tag{5.4}$$

Here H_f and H_i are Gibbs-Shannon entropies of the final and initial states, respectively. A standard textbook discussion of energy extraction stops here. However, the information-theoretic "bill" for this increase of energy has not been settled. The computer must update its memory to get rid of the no longer relevant record r_i about the initial state, and introduce the record r_f describing its knowledge about the final state. Below we shall prove that the cost of such an update for a computer operating in the environment of temperature T is no less than

$$\Delta W^{(-)} = T[K(r_f) - K(r_i)] = T(|r_f^*| - |r_i^*|) \ . \tag{5.5}$$

Here r_f^* and r_i^* are the minimal records—the most concise programs containing the information required to specify the ensembles describing states s_f and s_i. Our discussion will follow a similar but more complete account of the computational aspects of energy extraction and related issues (e.g., of the limitations imposed by undecidability on the thermodynamic efficiency) given elsewhere.[40]

To justify Eq. (5.5) we first note that the more verbose record r_f can be reversibly, i.e., without any work expenditure, substituted with r_f^* (providing, of course, that r_f^* is known[40]). The record of the initial state r_i plus the information about the operating procedure of the engine, which we assume is encoded in the computer library, suffice to compute r_f. This last computation could be also accomplished reversibly, but generally only at a cost of being left with r_f and the "historical" record of the computation path required to assure reversibility. This would lead to an accumulation of such a historical record with each engine cycle, which would slowly fill up the memory tape. Therefore the process would not be truly cyclic, and could not be used in the discussion of the first and of the second law of thermodynamics. Hence it would not suffice for our purpose.

Unless r_f and the operating procedure of the engine determine r_i uniquely, the computation of r_f from r_i

must be necessarily irreversible, as some of the information about the initial state is irreversibly erased from the memory. In short, a computation that disposes of some of the information about the initial state is logically irreversible, and hence, in accord with Landauer's remark, must be thermodynamically irreversible.

The key question relevant to our discussion is then: What is the least possible thermodynamic cost of computing r_f from r_i, in a manner which disposes of the record of the computation path by a universal computer operating in the environment of temperature T. We begin by noting that r_i contains obviously the information required to compute r_f:

$$K(r_i,r_f)=K(r_i)=|r_i^*| \ . \tag{5.6}$$

Moreover, r_f does constrain r_i partially. This fact can be expressed by noting that to compute r_i one can use r_f supplemented by the conditional information. According to the algorithmic information theory,

$$K(r_i,r_f)=K(r_f)+K(r_i|r_f^*) \ . \tag{5.7a}$$

In the above equation we have assumed that the computation of r_i is carried out from the minimal program r_f^* rather than directly from r_f; this is a technical assumption important in avoiding small (logarithmic) corrections[28] to the equation for the size of the conditional string. Another possibility of avoiding such corrections would involve assuming that the computation starts from r_f, but that $K(r_f)$ is also available. For, in accord with Eq. (2.6), one can write

$$K(r_f,r_i)=K(r_f)+K(r_i|r_f,K(r_f)) \ . \tag{5.7b}$$

Here $K(r_i|r_f,K(r_f))$ is the algorithmic content of the conditional information—the size of the minimal program to computer r_i given both r_f and $K(r_f)$. Moreover,

$$K(r_i,r_f)=K(r_f,r_i) \ . \tag{5.8}$$

We are now ready to show that no less than $|r_i^*|-|r_f^*|$ bits must be erased in the calculation of r_f from r_i. To this end, we note that reversible computation is capable of obtaining r_i from the record containing jointly r_f^* and $r^*_{i|f*}$, where $r^*_{i|f*}$ stands for the conditional string relevant in the context of Eq. (5.7a). Therefore one can also employ a reversible computer in the manner outlined before to substitute r_i with r_f^* and $r^*_{i|f*}$. Now to finish the update procedure, one can irreversibly erase the bits of $r^*_{i|f*}$. The number of bits that have to be erased is readily calculated from Eqs. (5.6)–(5.8). The length of the conditional information string is

$$|r^*_{i|f*}|=K(r_i|r_f,K(r_f))=|r_i^*|-|r_f^*| \ . \tag{5.9}$$

Erasure of $|r_i^*|-|r_f^*|$ bits of information in the environment of temperature T can be accomplished only at the expense of no less than $\Delta W^{(-)}$ of work, Eq. (5.5).

Note that these computations, Eqs. (5.6)–(5.9), recognize the fact that the relevant algorithmic randomness is defined with respect to the specific computer which happens to be operating the engine. Therefore computer-dependent error terms usually appearing in the discussion of the algorithmic randomness [$O(1)$] do not enter into consideration.

We have demonstrated the following.

Theorem 5.1.

The net work gained by an engine coupled with a computerized demon, which can perform measurements and control the operation of the engine, is no more than

$$\Delta W=\Delta W^{(+)}+\Delta W^{(-)}=T[(H_f+K_f)-(H_i+K_i)]$$
$$=T(\mathcal{S}_f-\mathcal{S}_i) \ . \tag{5.10}$$

This justifies our conjecture about the formula for physical entropy.

Proper accounting for the cost of the "information disposal" is essential in arriving at this formula. I have demonstrated that such accounting must be carried out in terms of the minimally sized programs capable of describing ensembles corresponding to the initial and final states of the active medium in the engine. Again, I have assumed that the minimal programs are already available and ready to use.

Possession of minimal programs can only increase the efficiency of the engine. ΔW in Eq. (5.2) refers to the *maximal* extractable energy. Therefore, if one can prove that Eq. (5.5) for the minimal cost of erasure holds when the efficiency of compression is optimized, this assumption can be made without danger: The computer cannot do better than we have assumed above, and we were trying to find out the maximal attainable efficiency.

However, in this sense, the above discussion assumes that the most difficult part of the procedure—finding the minimal algorithm for generation of the record describing the initial state given the minimal program describing the final state—has been somehow accomplished. In reality, the inability to guess or derive the most concise description will be the rule. It will increase the cost of erasure—more bits will have to be erased if the compression does not attain the limit defined by the algorithmic information content—and the efficiency of the engine will decrease. I shall discuss this fundamental inefficiency of information processing mandated by the undecidability and its relation to the thermodynamic efficiency of engines elsewhere in more detail (see Ref. 40). However, it is already tempting to conjecture that it provides the motivation for efficient acquisition and processing of information in systems which exist in and depend on an evolving nonequilibrium environment.

With the above caveats in mind we conclude that the absolute best Maxwell's demon can do is given by physical entropy which does include the algorithmic randomness of the relevant states in its definitions. This observation, expressed formally by Theorem 5.1, allows one to extend the range of applicability of suitably modified thermodynamics to engines operated by computers and by other systems capable of acquiring and processing information. The key result relevant to the discussion of the thermodynamic costs of information processing can be stated as follows. Suppose string r_2 can be computed from the program r_1 by a Turing machine T. If we demand that the computation should start with only r_1

on the tape (which assumes that r_1 is self-delimiting) and end with nothing but r_2, then the Turing machine must erase no less than $K(r_1)-K(r_2)$ bits in the process.[40]

The crucial result for statistical mechanics is the recognition that the entropy of an object can be properly defined as a sum of its statistical entropy with the ensemble defined by the data d and of the algorithmic information content associated with the description of that ensemble. This new definition of physical entropy will have to be reexamined both in the context of the standard issues of thermodynamics (e.g., second law) and from the point of view of its implications for the theory of measurements in classical and especially quantum physics.

VI. DISCUSSION

In this paper we have considered three measures of entropy. Boltzmann-Gibbs-Shannon entropy has an information-theoretic flavor and is a subjective property of the microstates in the following sense. The same microstate can be endowed with a very different value of the BGS entropy depending on the ensemble it is assigned to. This entropy is an (objective) property of an ensemble rather than of a microstate. However, ensembles are usually defined subjectively in a manner that may depend on the state of the knowledge of the observer. Moreover, both in classical and quantum physics the system is presumably in just one of the microstates constituting the ensemble.[41]

Algorithmic entropy is defined for a single microstate. It is difficult to calculate exactly, but typically relatively easy to estimate. In spite of the very different conceptual setting, its value is related to the number of microstates in much the same way as the Boltzmann-Gibbs-Shannon entropy. Therefore one could shift the foundations of thermodynamics and statistical mechanics from the ensemble to the algorithmic basis without endangering any of its key conclusions. Algorithmic entropy is an objective property of a microstate, except for the $O(1)$ corrections which are related to the size of the description—the algorithmic information content—of the universal computer used to define it.

Each of these two quantities is a satisfactory measure of disorder when applied from the outside of a complete (that is, including the Maxwell demon-type observer) system. However, *physical entropy,* the sum of the algorithmic randomness of the available data and of the still missing information, is necessary to discuss the process of energy extraction from the "inside," from the viewpoint of the observer.

In the experience of this author the only systems that are interested in the amounts of extractable work and capable of estimating entropy always seem to belong to the category of "observers." Therefore it is tempting to argue that the only entropy that should ever be used is the physical entropy. Indeed, even the traditional ensemble entropies are defined subject to the knowledge of the type of ensemble and its macroscopic parameters, which constitute the relevant data. From this point of view, the usual statistical entropy is the physical limit applicable in the case of almost complete ignorance, $K(d)\ll H_d$. Algorithmic entropy of a microstate is the opposite limit.

Having said all that, one must recognize that the statistical entropy is almost always an excellent approximation of the physical entropy—we are invariably in the vicinity of such a thermodynamic limit. Moreover, statistical entropy is far more convenient to calculate and to use, and there is certainly no reason to try to replace it with physical entropy in engineering applications of thermodynamics. One should, on the other hand, use physical entropy $\mathcal{S}$ whenever issues of principle are considered from the internal point of view of the information gathering and using systems.

To further investigate the connection between the physical and statistical entropy it is useful to recast the discussion of Sec. V in the ensemble language. Consider a system and a fine partition of its phase space $\xi=\{c_0,c_1,\ldots,c_N\}$. The standard entropy is

$$H(\xi)=-\sum_i p(c_i)\log_2 p(c_i)\ . \tag{6.1}$$

Suppose now that the Maxwell demon can make a dissipationless measurement corresponding to a coarse partition β. Suppose also that ξ is a refinement of β. The demon can now operate by (a) measuring the observable associated with the partition β to determine which $b\in\beta$ contains the microstate, (b) constraining the system to lie within b in a dissipation-free manner, and (c) extracting the work by relaxing the constraints in a slow isothermal reversible fashion.

Before the process begins with step (a) and after it is completed with step (c), the entropy is $H(\xi)$. However, at the intermediate step (b), it is equal to the conditional Shannon entropy $H(\xi|\beta)$. Thus the work extracted in the course of a cycle, when averaged over all the possible measurement outcomes equals

$$\Delta W=T[H(\xi)-H(\xi|\beta)]=T\mathcal{I}(\xi{:}\beta)\ . \tag{6.2}$$

Here $\mathcal{I}(\xi{:}\beta)$ is the mutual Shannon information defined by the statistical, ensemble analog of Eq. (2.7). Mutual information is a measure of correlation between two ensembles.[12,13]

So far, the demon has gained ΔW of work at the price of having its memory cluttered by the no longer relevant historical information about a "used-up" measurement outcome. Clearly, it would be possible for the demon to operate the "engine" outlined above *ad infinitum* only if its memory had infinite capacity. However, in order to achieve a truly cyclic process, the memory of the demon should be returned to the initial uncluttered state. This, according to Landauer[35] and Bennett,[27] can be accomplished by resetting the demon's memory at a cost given by k_BT times the number of erased bits.

In order to operate at maximum efficiency, the demon has to minimize that cost. Hence it must "compress" the block of memory occupied by the results of past measurements to the absolute minimum. Such compression can be achieved at no additional cost as long as it does not alter the information content. The absolute minimum in the size of the description is achieved by the minimal program b^*, the length of which defines the algorithmic en-

tropy of the measurement outcome with respect to this universal computer U which models operations of Maxwell's demon. As the optimal sizes of programs are related to the Shannon entropy via Eq. (4.17), the average algorithmic randomness—typical number of memory bits required to represent the outcome of the measurement—is almost exactly equal to $\mathcal{I}(\xi:\beta)$. Thus, already on the ensemble level, one can conclude that the demon will, at best, get nowhere.

We have just described the process of energy extraction by the demon in the ensemble language. Why should one then bother introducing "physical entropy," which is clearly more complicated than the probabilistic ensemble quantities?

The answer is straightforward: Physical entropy allows the demon to give its "private and personal" account of the attempts to extract energy. Each of these individual attempts is associated with the definite outcome of the measurement. Each of them involves a computation in an attempt to represent the measurement outcome in the simplest possible way. Maximum efficiency can be ascertained only when the computation leads to a minimal description. Moreover, the demon will miss the opportunity to extract all of the energy from these outcomes for which it will be unable to produce the most concise description because of the incompleteness. Hence the incompleteness emerges as one of the reasons for inefficiency. One could, of course, ascertain optimal performance by furnishing the demon ahead of time with the complete list of all the possible outcomes arranged in the order corresponding to probability—a "Huffman code" for a given ensemble. This would, however, require enormous, possibly infinite memory as well as a prior knowledge of the ensemble. Moreover, it would necessarily be a very inflexible strategy, locked to a single, definite ensemble. In practice, living systems that attempt to extract useful energy from their surroundings are faced with a variety of unanticipated circumstances. Therefore, rather than store a ready solution for every possibility, it is advantageous to let some elementary equivalent of computation choose the strategy. Fortunately, there are usually sufficiently many opportunities to allow for a decrease of physical entropy to sustain their activity. Moreover, on longer time scales it may be of enormous advantage to modify the information-processing system. Such modifications take advantage of the remaining subjectivity of the definition of algorithmic information content—of its dependence on the computer U with respect to which it is calculated.

Consequently, while introducing physical entropy does not improve thermodynamic limits on engine efficiency, it does allow for their discussion in a very different context, which ties statistical physics with the theory of computation, and which provides a physical motivation for efficient information processing. Further implications may include a combined physico-computational insight into biological systems and their evolution. Ensemble formulations provide neither a motivation nor a language to explore such questions. Physical entropy—tailored to represent information from the viewpoint of the system that acquires and makes use of it—appears to be an indispensable tool. It forces one to adopt a hybrid—partly probabilistic and partly algorithmic—definition of entropy.

It is interesting to contemplate an extension of the approach developed in this paper to quantum theory. We shall pursue this subject in more detail elsewhere. Two brief remarks anticipating some of the conclusions of the more complete discussion are nevertheless appropriate. Let us first note that the algorithmic entropy of a quantum system can be defined in much the same way as that of the ideal Boltzmann gas. Pure states can have different algorithmic entropies depending on how "random" and "disordered" they are. The choice of the basis in the Hilbert space can be regarded as equivalent to the choice of the "grid." All simply defined choices of the basis are expected to result in a similar algorithmic entropy for the same state.

Physical entropy of a quantum system can be defined for a mixture in a manner analogous to the "classical" physical entropy defined in Sec. V. However, the discussion of the problem of measurement is now somewhat different: Measurements performed on a mixture will decrease missing information and alter the record. Measurements performed on pure states can only alter these states (and update the records about them). This distinction as well as other issues usually raised in the context of quantum measurements are sufficiently complicated to warrant a separate paper.

One may inquire as to why this algorithmic aspect of physical entropy was not considered earlier. There are, of course, obvious historical reasons: Questions concerning entropy, Maxwell demons, etc., were generally considered either settled or hopeless by the second part of this century, and only rather recently was the computer-oriented point of view of information sufficiently developed to be of help in the issues of principle.

The success of the ensemble definition of entropy, combined with the fact that ensemble and algorithmic estimates of entropy almost always coincide, is the other important reason for concealing the alternative view of entropy. This is not to claim that the role of randomness was underestimated: The influential paper of Ehrenfest's[12] emphasized the role of randomness in the origin of irreversibility and interpreted Boltzmann's and Gibb's definitions from that point of view. Prigogine and his co-workers[9,43] have advocated the use of the formalism in which entropy is defined through probability distributions, but in a manner that has no information-theoretic connotations. These efforts, however, invariably associated randomness with probabilities. While this association is not incorrect, it bypasses the alternative, and conceptually very fruitful, algorithmic approach.

VII. CONCLUSIONS

Questions concerning physics and computation have been so far focused on the basic limitations placed by physical laws on information-processing systems. By contrast, this paper is concerned with the problem of a "complementary" nature: I have explored some of the

implications that information acquisition and processing have for physics.[44]

The definition of the physical entropy proposed here is made possible by the algorithmic definition of randomness. It is made necessary to the desire to discuss the function of the Maxwell's demon from its own perspective. Moreover, the demon's role can be successfully played by an automaton capable of (i) acquiring information through reversible measurements, (ii) processing it in a manner analogous to the universal computer, and (iii) adopting strategies that aim to optimize its behavior so that, for example, it can economically extract useful energy from the available sources.

Physical entropy, which must be used in the discussion of the laws of thermodynamics from the internal point of view of such an automated engine, is the sum of the missing information and of the thermodynamic cost of updating the memory with the record of the measurement outcomes. It provides one with a new perspective of the process of measurement. It extends the Boltzmann-Gibbs-Shannon missing information paradigm by taking into account the cost of storage of the information about the system. It is no longer necessary for the parameters defining ensembles of concern in the process of energy extraction to be few and "macroscopic:" They can be numerous and correspond to microscopic data, but the price for this information must be taken into consideration in evaluating the efficiency of the engines controlled by computers.

APPENDIX A: INDISTINGUISHABILITY AND THE SACKUR-TETRODE EQUATION

Consider a gas of N particles distributed "at random" in volume V in D-dimensional space. To within a predetermined accuracy $\Delta V=\Delta_x^D$ the location of each of these particles is described by D integers. When these integers are algorithmically random, the length of a program needed to encode the location of a single particle is $\sim\log_2 V/\Delta V$. In order to take care of all N particles, a program of typical length

$$K_V \simeq N\log_2(V/\Delta V)+O(\log_2(ND))+O(1) \qquad \text{(A1)}$$

is usually needed. Note that this formula has only a small self-delimiting correction $O(\log_2(ND))$ rather than the correction $O(\log_2 K_V)$ of the type appearing in Eq. (2.5). This correction does not depend on the actual algorithmic randomness of the microstate, but rather solely on the particle number and dimensionality. It can therefore be incorporated as a part of the constant $O(1)$ correction for the purpose of further discussion.

Such a small upper limit on the size of the correction comes from the recognition that the most economical way of writing a self-delimiting program is to specify once and for all the sizes of blocks of digits describing coordinates of particles in the phase space. The other possibility would be to write a program in which each of the coordinates is encoded by means of a separate self-delimiting subroutine. In some cases this would save space on the input tape. However, as the reader is encouraged to verify, for a random distribution of particles, the additional expenditure would be $\sim D^{-1}\log_2 V/\Delta V$.

This brief discussion outlining two different possible programs that can generate the same binary image illustrates an important point: Estimates of algorithmic entropy are quite insensitive to the details of the method of encoding. We can now consider the space on the program tape required to accomplish a similar "encoding" procedure for the momentum of an individual particle. In each degree of freedom a typical number of digits will be $\log_2(p/\Delta_p)$. As the expected value of one component of p is $(mkT)^{1/2}$, a typical size of the program required to encode all of the components for one particle must be, to leading order, $D\log_2\sqrt{mkT}/\Delta_p$. The estimate of algorithmic entropy is then

$$K=K_V+K_p\cong N\left[\log_2 V/\Delta V+\frac{D}{2}\log_2\frac{mkT}{(\Delta_p)^2}\right]+O(1)\,. \qquad \text{(A2)}$$

The expression for entropy, Eq. (A2), has a sensible form, but results in a counterintuitive prediction. Consider N gas particles in a container separated in the middle by a partition. Suppose that the partition is removed. Particles which were on the opposite sides of the partition in approximately equal numbers can now begin to mix. Using Eq. (A2) to calculate the difference of entropies before and after the removal of the partition results in a change of the value of the "candidate entropy," Eq. (A2):

$$\delta K_{\Delta V}=N\log_2 2=N\,. \qquad \text{(A3)}$$

However, change of entropy should be zero, as no real event occurs as a result of the removal of the partition.

The cause of the increase of entropy by $\delta K_{\Delta V}^{(d)}$ can be traced to the fact that we have not taken advantage of the indistinguishability of the particles in devising the program designed to encode the state of the system. Information about individual particles comes in separate blocks, each of which carries an "implicit" index (i.e., it arrives as a "number 17 input block" on a tape). In other words, the special-purpose machine L we have described at the beginning of Sec. III could print in the pixels occupied by the particles not just 1 to signify that the corresponding cell in the phase space is occupied: It could also print out a label of each particle implicitly encoded in its order of appearance on the input data tape.

To compress further the program and to take advantage of the particle indistinguishability, we can employ a modified version of L, our special purpose computer. Instead of the coordinates of individual particles in one of the dimensions, one can supply only differences in their locations.

Any coordinate of the phase space would do. Consider, for instance, spatial coordinate x_1. Integers needed to describe the distribution of particles in this dimension will be now smaller. Where, before, a number of the order of an average value of x, $\langle x\rangle\sim L_x/2$, was required, now a smaller number $\sim\langle(\Delta x)^2\rangle^{1/2}=\langle x\rangle/N$ will be sufficient. This one small number has to be followed by $D-1$ other numbers characterizing other coordinates of

the same particle, which have retained their previous values. The algorithmic entropy of the gas in a D-dimensional cubic box with the volume $V=L^D$ is therefore given by

$$K=N\log_2\frac{(L/N)}{\Delta V^{1/D}}+(D-1)N\log\frac{L^{D-1}}{(\Delta V)^{(D-1)/D}}+N\frac{D}{2}\log_2\frac{mkT}{(\Delta_p)^2}+O(1)\ ,$$

which readily simplifies. We are therefore led to the key result of this section.

Theorem A1. Algorithmic randomness of a typical microstate of an ideal gas of indistinguishable particles is given by

$$K=N\left[\log_2\frac{V}{N\Delta V}+\frac{D}{2}\log_2\frac{mkT}{(\Delta_p)^2}\right]+O(1)\ . \qquad \text{(A4)}$$

This is the Sackur-Tetrode equation for the entropy of the monoatomic gas. Its additive properties are entirely satisfactory. In particular, the experiment with a partition being removed now leads to $\delta K=0$, in accord with the physically motivated expectations.

APPENDIX B: HOW OBJECTIVE IS ALGORITHMIC RANDOMNESS?

Algorithmic entropy is sufficiently different from the traditional definition of entropy to be regarded with suspicion. There were several points that were left out from the discussion in Sec. III in an attempt to gain a first quick, physical insight into the relation between the algorithmic randomness defined through its information content and the equilibrium entropy of an idealized physical system. We return to them now.

The first of them concerns the container in which the gas is enclosed: It is conceivable that the container will be so complex in shape that the program required to specify it will be comparable in size with the algorithmic randomness of the particle configuration inside. Yet one would like the entropy to characterize the state of the gas alone. A "brute force" solution to this problem is to insist on simple containers. A more sophisticated solution is to employ conditional information. That is, one can define the entropy of the gas alone as the number of additional bits that must be supplied—given a description of the container—to specify the state of the gas with the assumed resolution.

The second issue that was glossed over is more important. It concerns the grid that was used to discretize the system. The first, simple aspect of this issue concerns the size adopted for individual cells. Obviously, in the limit of infinitesimally small cells, algorithmic entropy of a typical microstate of the system diverges logarithmically. This difficulty is not new. Before the advent of the quantum theory entropy was also thought to be logarithmically divergent for the same reason.[7] Entropy is finite only because the volume of the cells in the phase space is limited by the quantum of action. There is, however, a way of partially dealing with this problem which does not call on the quantum: One can compare different configurations of the gas on the grid with the same resolution. While the absolute value of algorithmic randomness is still undetermined, it is now possible, at least in principle, to compute differences of algorithmic randomness of different configurations.

While the problem of the logarithmic divergence of entropy with the resolution of the grid is rather similar for both the statistical and the algorithmic approach, the issue of the shape of the grid is not. This problem is responsible for the ambiguity in defining the coarse-grained entropy S_G of Gibbs[2]. By shifting cells defining coarse-grained entropy one can usually make S_G assume any value between its equilibrium value and the smallest value consistent with the fine-grained distribution. (This difficulty is particularly acute in the context of Gibb's "proof" of the second law. If the grid defining coarse graining is continually deformed with the same evolution Hamiltonian that generates evolution of the fine-grained distribution, then the coarse-grained entropy will be constant. Hence, unless one restricts possible grids in some suitable fashion, Gibb's arguments in favor of the validity of the second law are obviously incomplete.)

A natural resolution of this ambiguity is to insist that the grid should be "simple." The algorithmic information theory allows one to give a rigorous definition of what is simple: A simple grid is concisely describable by the (specific) universal computer U. For instance, a grid with hexagonal cells may be more appropriate for some configurations, and it is still concisely describable. A simple and natural requirement—to include the prescription for the grid as a part of the program describing the state of the gas—will eliminate the advantage of using fanciful grids. In other words, when the optimal strategy of describing the state of the system includes using an unusual and complex coordinate system, it still is a perfectly legal strategy, as long as the cost of specifying the grid is counted as part of the cost of this strategy.

The difficulty of Gibb's "coarse-graining" approach arose from the inability to measure the complexity of the different coarse-graining schemes and to account for them in units of entropy: All grids, no matter how "perverse," had to be regarded as equally valid. Algorithmic randomness settles this difficulty to some extent: It makes it possible to quantify the intuitive concept of "simplicity." In particular, this allows one to rule out co-evolving grids used in arguments against coarse graining simply because they are too complex.

In view of the importance of the concept of coarse graining and the long-standing concern with its subjectivity, it is useful to delineate more precisely what can and what cannot be accomplished by using algorithmic prescriptions. The discussion above assumes (i) a definite universal computer and (ii) a definite metric on the phase space, which must be unaffected by the choice of the grid. Given these two assumptions one can expect to arrive at the algorithmic definition of entropy which, for a given computer, does not significantly depend on the deformations of the grid.

Algorithmic randomness is by definition equal to the size of the smallest message program for a given universal computer needed to reproduce the state of the system. Once the resolution is fixed by demanding, for example,

that the sum of squares of the differences between the locations of the real microstate particles and their "images" on the plot, evaluated in some natural units (e.g., units related to the total volume of the phase space available to the system), is less than the specified tolerance, the grid can be varied in attempts to meet the criterion with the shortest message. There is just one smallest size for such a program. This minimal size is the one that gives the algorithmic entropy of the system. Such minimization includes possible variations of the grid.

At first, one might worry that this freedom to choose the grid makes K completely grid dependent. This is not the case. If this were true, one could prove that the size of the message, including both the description of the grid and the location of the system within the grid, can be made arbitrarily small *for a given Turing machine* U by making a judicious grid choice. The theory of information and coding proves that this must be impossible. For, if it were the case, one could violate Shannon's noiseless channel coding theorem[12,13] by first encoding messages in particle configurations on the source end, then using the "variable grid trick" to make message programs very short, and finally decoding it on the other end with the help of the computer.

Eliminating the subjectivity associated with the coarse graining does not, however, banish it altogether. Rather it reemerges under the guise of the definition of a universal computer employed to do the plotting of microstates. In this new form, it is, however, easier to deal with and quantify than in the form of coarse graining. In particular, all concisely describable Turing machines (i.e., all computers that have algorithmically short descriptions) will yield comparable estimates for algorithmic complexity. They define (approximately) "standard" algorithmic randomness. More generally, the algorithmic complexity of a machine with a long description can differ from such a standard by no more than the size of its description.

Issues of the freedom of choice of coarse graining, subjectivity of algorithmic randomness, and the resulting estimates of entropy are of course particularly important in the discussion of the second law. Therefore we shall return to them in a future paper in more detail. For the purpose of the present paper, which is mostly concerned with the definition of physical entropy by specific, individual observers, it is entirely sufficient to adopt the natural grid defined by the observer's measurements and use it to define algorithmic randomness.

APPENDIX C: SECOND LAW AND EVOLUTION OF A DISCRETE DYNAMICAL SYSTEM

The aim of this appendix is to discuss the behavior of the algorithmic randomness in the course of the dynamically reversible evolution of a discrete system. We consider *reversible* transformations of a *finite* string s:

$$s_{t+1}=Us_t \ . \tag{C1}$$

Equation (C1) can be regarded as a deterministic law of motion generating the phase-space trajectory of a simple example of an elementary, idealized system. We will demonstrate how the reversible, deterministic, dynamical evolution can lead to an increase of the algorithmic randomness. We shall conclude that (i) algorithmic randomness $K(s_t)$ tends to increase in the course of evolution if the initial state is algorithmically simple, $K(s_0) \ll |s_0| \sim \log_2 W$; (ii) $K(s_t)$ approaches the equilibrium value consistent with the Boltzmann estimate $\log_2 W$ of entropy; and (iii) it fluctuates around this equilibrium.

We shall require the transformation U to be reversible to establish a closer analogy with the Hamiltonian dynamics. The old problem of statistical mechanics—the fundamental incompatibility of reversible dynamics with the second law—can be posed only when U is reversible. That is, we assume that there should exist U^{-1} such that

$$s_t=U^{-1}s_{t+1} \ . \tag{C2}$$

We shall also require U to be concisely describable, i.e., for a typical instant t

$$K(s_t,s_{t+1})-K(s_t,s_t)\equiv K(U)\ll K(s_t,s_t)=K(s_t)+O(1) \ .$$

Ciphers used in cryptography provide an example of such transformations.[45] Another, related example is afforded by "random-number" generators used in Monte Carlo programs.[46] For example, dynamical evolution of a discrete system can be described by

$$s_{t+1}=\lambda s_t+\mu (\mathrm{mod}\ \mathcal{P}) \ . \tag{C3}$$

Above, the string s is treated as an integer which it represents in the binary notation. Constants λ, μ, and $\mathcal{P}$ characterize U. It is not difficult to find examples where these numbers are easily describable.

The phase space in which the string s evolves is a set of all strings that correspond to natural numbers contained in the interval $[0,\mathcal{P}]$. The evolution shall be called *perfectly ergodic* if after $\mathcal{P}$ iterations the system would have passed through all states of the phase space. The evolution shall be called *quasiergodic* of order n if only $\mathcal{P}/n$ states were occupied in that time. Evolutions for which $n<\log_2\log_2\mathcal{P}$ shall be termed ergodic.

For example, the random-number generator, Eq. (C3), is perfectly ergodic when $\mathcal{P}$, λ, and μ satisfy the following conditions.[44]

(i) The greatest common divisor of μ and $\mathcal{P}$ is 1.

(ii) $\lambda\equiv 1(\mathrm{mod}\ q)$ for every prime factor q of $\mathcal{P}$.

(iii) $\lambda=1(\mathrm{mod}\ 4)$ if $\mathcal{P}$ is a multiple of 4.

In particular, when $\mathcal{P}=2^N$ (which means that $\mathcal{P}$ is concisely describable), $\lambda=4\alpha+1$ and $\mu=2\beta+1$, where α and β are arbitrary natural numbers, transformation given by Eq. (C3) will result in a perfectly ergodic random number generator. Hence it is quite easy to construct a concisely describable, ergodic U, with a very large volume of the phase space given by

$$W=\mathcal{P} \ . \tag{C4}$$

The analogy between classical dynamics and transformations described by Eq. (C3) can be pursued further by noting that the n separate orbits of the quasiperiodic discrete maps can be associated with the constants of motion.

Consider now the behavior of the algorithmic complexity of the time-dependent state of the system s_t in the course of the dynamical evolution defined by Eq. (C3). Let us first point out that this problem can be formulated in two different, but equivalent ways: One can either (i) consider the relative complexity of the initial string s_0 and the evolved string $s_t = U^t s_0, K(s_t/s_0)$, or (ii) one can inquire directly about $K(s_t)$. It is the first option that allows one to pose the problem of the apparent irreversibility of dynamical evolutions in an "observer-independent" manner. For one may now propose to measure the increase of algorithmically defined entropy by

$$\Delta K = K(s_t|s_0) \ . \tag{C5}$$

Therefore the entropy will be defined *relative* to the initial configuration. This configuration may or may not appear "simple" to an observer. Below, we shall see that the relative entropy will typically increase, although with fluctuations, over time scales smaller than the "Poincaré recurrence time."

On the other hand, if we were to assume that the initial state, which was presumably prepared by the observer, were algorithmically simple $[K(s_0) \ll \log_2 \mathcal{P}]$, then, instead of the conditional randomness of Eq. (C5), one can directly estimate

$$\Delta K \cong K(s_t) \ . \tag{C6}$$

This equation will be valid whenever $K(s_t) \gg K(s_0)$, that is, for an overwhelming majority of the phase space.

To prove that ΔK will tend to increase on time scales less than the Poincaré recurrence time $\mathcal{P}/n$, we must show that a program capable of reproducing the pair (s_0, s_0) is much less complex than the program reproducing (s_0, s_t). We can anticipate that the difference of complexity should be of the order of the typical randomness of the state s_t. Typically, it should be similar to the average complexity of an algorithmically random number of order $\mathcal{P}/n$:

$$\Delta K \simeq \log_2 \mathcal{P}/n \simeq \log_2 \mathcal{P} \ . \tag{C7}$$

Above, and further in this appendix we assume that the evolution of the system is ergodic.

This expectation, Eq. (C7), appears to lead to a paradox. After all, the evolution from s_0 to s_t can be readily expressed as

$$s_t = U^t s_0 \ . \tag{C8}$$

Why should s_t be significantly more difficult to describe than s_0, if U is, by assumption, concisely describable? It is straightforward to estimate the complexity of the program that starts from s_0 and proceeds to generate recursively all states of the string s in its phase space as given by $K(s_0) + K(U) + 0(1)$. This number can be clearly made much smaller than $\log_2 \mathcal{P}$. How can one then expect a program that generates s_t from s_0 to be, for a typical t, as long as $\log_2 \mathcal{P}$?

The requirement that the program must have a unique output, emphasized already in Sec. II, provides a resolution of this apparent paradox. Apart from the information about s_0 and U the program must contain the information on when to halt the calculation. This is given by t itself, and typical t is of order $\mathcal{P}/n$. Consequently, a binary string of the length $\sim \log_2 t$ must be added to the simple program generating the whole trajectory. The increase of the algorithmic randomness, as defined by Eqs. (C5)–(C8), will then be bounded from above by

$$\Delta K \simeq \log_2 t \ . \tag{C9}$$

It is not difficult to see that this estimate of entropy satisfies, for t significantly smaller than the recurrence period $\mathcal{P}$, not only the Boltzmann H theorem, but its generalized version

$$(-1)^\nu \left[\frac{d}{dt}\right]^\nu (-\Delta S) \geq 0 \tag{C10}$$

as well. These results strongly indicate that the algorithmic complexity can be regarded as an interesting contender for the "objective" (that is, relatively observer-independent) definition of physical entropy. Moreover, it demonstrates that to give a rigorous definition of entropy one need not try to argue that dynamical evolutions transform definite, pure states into mixtures. Probability is not indispensable as the foundation of statistical mechanics.

It is of course of great interest to inquire as to what is the relation between the algorithmic complexity and the more traditional recipes for calculating entropy. It is also of fundamental importance to consider how our idealized model of a dynamical system can be used to deduce the behavior of entropy in real physical systems.

The "equilibrium value" of entropy, given by the typical size of the program which must be used to generate s_t is given by

$$\Delta K \cong \log_2 \mathcal{P} + \log_2 \log_2 \mathcal{P} \ . \tag{C11}$$

There is one prevailing reason for large fluctuations that approach $\Delta K \sim 0$: recurrences which occur on the time scale of the order of $\mathcal{P}$. As in the usual discussions of classical mechanics, they are forced upon the system by the finite volume of the accessible phase space, given here directly by $\mathcal{P}$. Apart from these rare occurrences ΔK may fluctuate significantly above the equilibrium value whenever the string s_k is either intrinsically concisely describable, or can be easily generated by transforming s_0. Of course, these two cases need not be discussed separately when s_0 is itself concisely describable. Small fluctuations around the equilibrium value can be discussed more naturally in the context of the relationship between complexity and traditional, probabilistically defined entropies, which we have considered in Sec. IV. However, it is easy to anticipate their origin. Clearly, for some special time $t_V \sim \mathcal{P}$ there may be a concise algorithm V capable of generating s_{t_V} from s_0 which does not employ the transformation U. When $K(V) \ll K(U^t) \sim K(t)$, the algorithmic complexity of s_{t_V} relative to s_0 will be small, and a large and very unlikely fluctuation—the spontaneous departure from the state of equilibrium—will have occurred.

It is important to emphasize that while the above discussion is clearly relevant for the dynamical understanding of the second law, it is only a part of the whole story. In particular, in the above example, entropy increases only very slowly (it gets "half way" to its equilibrium value only after a time $\sim \mathcal{P}^{1/2}$), which is inconsistent with the relaxation of systems such as the Boltzmann gas. Moreover, the system is stable in the sense that a collection of several natural states—binary strings—always leads to a similar collection with the same number of natural states at a later time. This is not the case in systems which, like Boltzmann gas, exhibit exponential relaxation and where a single "cell" in the phase space is rapidly smeared by the dynamical evolution over many "natural" cells. Indeed, the example we have considered here can be regarded as "integrable."

I shall discuss relaxation from the algorithmic point of view in more realistic systems elsewhere in more detail. The point of this appendix was only to demonstrate that integrability of the system does not preclude the possibility that in the course of its evolution it will reach algorithmically random—disordered—configurations even if the initial configuration were regular and the Hamiltonian responsible for the evolution of the system were simple.

ACKNOWLEDGMENTS

I would like to thank Charles Bennett, Murray Gell-Mann, James Hartle, Seth Lloyd, William Unruh, and John Wheeler for discussions and comments on the manuscript.

[1]L. Boltzmann, Wien Ber. **76**, 373 (1977). See S. G. Brush, *Kinetic Theory* (Pergamon, Oxford, 1965), Vol. 1, (1966), Vol. 2, and (1972), Vol. 3, for translation of this and other relevant papers.

[2]J. W. Gibbs, *Elementary Principles in Statistical Mechanics* (Yale University Press, London, 1928).

[3]J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton University Press, Princeton, 1955).

[4]H. Grad, Comm. Pure Appl. Math. **14**, 323 (1961).

[5]K. G. Denbigh and J. S. Denbigh, *Entropy in Relation to Incomplete Knowledge* (Cambridge University Press, Cambridge, England, 1985).

[6]P. C. W. Davies, *Physics of Time Asymmetry* (University of California Press, Berkeley, 1977).

[7]S. G. Brush, Ref. 1, Vols. 1–3.

[8]A. Wehrl, Rev. Mod. Phys. **50**, 221 (1978).

[9]I. Prigogine, *From Being to Becoming* (Freeman, San Fancisco, 1980).

[10]J. Ford, Phys. Today **36**, (4) **40**, (1983), and references therein.

[11]*Quantum Theory and Measurement,* edited by J. A. Wheeler and W. H. Zurek (Princeton University Press, Princeton, 1983); H. Everett, in *The Many-Worlds Interpretation of Quantum Mechanics,* edited by B. S. DeWitt and N. Graham (Princeton University Press, Princeton, 1973); W. H. Zurek, in *Information Transfer in Quantum Measurements: Irreversibility and Amplification,* Vol. 94 of *Nato Advanced Study Institute Series* in *Quantum Optics, Experimental Gravitation and Measurement Theory,* edited by P. Meystre and M. O. Scully (Plenum, New York, 1983), pp. 87–106.

[12]W. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).

[13]A. I. Khintchin, *Information Theory* (Dover, New York, 1957); R. W. Hamming, *Coding and Information Theory* (Prentice Hall, Englewood Cliffs, NJ, 1986).

[14]S. W. Hawking, Phys. Rev. D **13**, 191 (1976).

[15]J. B. Hartle and S. W. Hawking, Phys. Rev. D **28**, 2960 (1983).

[16]R. J. Solomonoff, Inf. Control **7**, 1 (1964).

[17]A. N. Kolmogorov, Inf. Transmission **1**, 3 (1965).

[18]G. J. Chaitin, J. Assoc. Comput. Mach. **13**, 547 (1966).

[19]A. N. Kolmogorov, IEEE Trans. Inf. Theory **14**, 662 (1968).

[20]A. N. Kolmogorov, Usp. Mat. Nauk. **23**, 201 (1968).

[21]G. J. Chaitin, J. Assoc. Comput. Mach. **22**, 329 (1975), and references therein.

[22]G. J. Chaitin, Sci. Am. **232**(5), 47 (1975).

[23]G. J. Chaitin, IBM, J. Res. Dev. **21**, 350 (1977).

[24]A. K. Zvonkin and L. A. Levin, Usp. Mat. Nauk. **25**, 602 (1970).

[25]P. Martin-Löf, Inf. Control **9**, 602 (1966).

[26]G. J. Chaitin, *Algorithmic Information Theory* (Cambridge University Press, Cambridge, England, 1987), and references therein.

[27]C. H. Bennett, Int. J. Theor. Phys. **21**, 905 (1982).

[28]P. Gacs, Dokl. Akad. Nauk SSSR **218**, 1265 (1974) [Sov. Phys.—Dokl. **15**, 1477 (1974)].

[29]L. A. Levin, Dokl. Akad. Nauk SSSR **227**, 1293 (1976) [Sov. Phys.—Dokl. **17**, 522 (1976)].

[30]L. Szilard, Z. Phys. **53**, 840 (1929), English translation in Behav. Sci. **9**, 301 (1964), reported in *Quantum Theory and Measurement,* Ref. 11.

[31]E. T. Jaynes, Phys. Rev. **106**, 620 (1957); **108**, 171 (1957).

[32]L. Brillouin, *Science and Information Theory,* 2nd ed. (Academic, London, 1962).

[33]D. Hofstader, *Gödel, Escher, Bach* (Vintage, New York, 1980).

[34]C. H. Bennett, IBM J. Res. Dev. **17**, 525 (1973).

[35]R. Landauer, IBM J. Res. Dev. **3**, 183 (1961).

[36]C. H. Bennett, IBM J. Res. Dev. **32**, 16 (1988).

[37]R. Landauer, Found. Phys. **16**, 551 (1986).

[38]C. H. Bennett, Sci. Am. **255**(11), 108 (1987).

[39]W. H. Zurek, in *Frontiers of Nonequilibrium Statistical Mechanics,* edited by G. T. Moore and M. O. Scully (Plenum, New York, 1986), pp. 151–161.

[40]W. H. Zurek, *Nature* (to be published); and unpublished.

[41]W. H. Zurek, Phys. Rev. D **24**, 1516 (1981); **26**, 1862 (1982); in Ref. 39, pp. 145–149.

[42]P. Ehrenfest and T. Ehrenfest, *The Conceptual Foundations of the Statistical Approach,* English translation (Cornell University Press, Ithaca, 1956).

[43]I. Prigogine, C. George, F. Henin, and L. Rosenfeld, Chem. Scr. **4**, 5 (1974).

[44]Such a "reversal" of the approach was advocated in the context of quantum measurements by J. A. Wheeler, see, e.g., IBM J. Res. Dev. **32**, 4 (1988), and references therein. See Ref. 38 for comments on the relevance of computation to Wheeler's ideas.

[45]J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods* (Wiley, New York, 1964).

[46]A. G. Konheim, *Cryptography, a Primer* (Wiley, New York, 1981).