Brit. J. Phil. Sci. **55** (2004), 411–434, axh303

In What Sense is the Kolmogorov-Sinai Entropy a Measure for Chaotic Behaviour?—Bridging the Gap Between Dynamical Systems Theory and Communication Theory

Roman Frigg

ABSTRACT

On an influential account, chaos is explained in terms of random behaviour; and random behaviour in turn is explained in terms of having positive Kolmogorov-Sinai entropy (KSE). Though intuitively plausible, the association of the KSE with random behaviour needs justification since the definition of the KSE does not make reference to any notion that is connected to randomness. I provide this justification for the case of Hamiltonian systems by proving that the KSE is equivalent to a generalized version of Shannon's communication-theoretic entropy under certain plausible assumptions. I then discuss consequences of this equivalence for randomness in chaotic dynamical systems.

1 *Introduction*
2 *Elements of dynamical systems theory*
3 *Entropy in communication theory*
4 *Entropy in dynamical systems theory*
5 *Comparison with other accounts*
6 *Product versus process randomness*

1 Introduction

For many years, chaos theory has been a hotly debated topic and its methods have been used to model a great variety of different situations. However, there is still controversy over the proper characterization of chaotic behaviour. Numerous criteria and indicators for the onset of chaos have been suggested, but none of these has gained the status of a 'canonical definition'. Intuitively, chaos has two faces: random behaviour and sensitive dependence on initial conditions. Accordingly, the great majority of proposals for characterizing chaos fall into one of the following two groups. The first group focuses on the seemingly random, stochastic, unpredictable or haphazard time evolution of chaotic systems and consequently tries to flesh out chaos in terms of randomness. The second group focuses on sensitive dependence on initial conditions. This leads to a study of the properties of trajectories in phase space, in particular their exponential instability: a slight variation in initial conditions produces significant changes in the long-term behaviour of trajectories.

This paper is concerned with the first group of proposals. The problem with this type of suggestion is that to characterize a system's behaviour as random or stochastic is not very illuminating, since these notions are as much in need of analysis as chaos itself. What does it mean to say that a dynamical system exhibits random behaviour?

Common physical (as well as philosophical) wisdom has it that ergodic theory fits the bill. More specifically, the claim is that the ergodic hierarchy provides a set of concepts which allows for an adequate characterization of random behaviour (see, e.g., Lichtenberg & Liebermann [1992], pp. 302–12; Ott [1993], pp. 261–2; Reichl [1992], pp. 47–53; Schuster [1988], pp. 203–7; Tabor [1989], pp. 167–74). A discussion of the entire hierarchy is beyond the scope of this paper, and for the present purpose only one of its notions is of importance—the Kolomogorov-Sinai Entropy (KSE, for short). This notion is crucial because it is generally assumed that the move from zero to positive KSE marks the transition from regular to chaotic behaviour. More precisely, the claim is that having positive KSE is a sufficient condition for chaos (Belot & Earman [1997], p. 155).

Although, at first glance, this might seem plausible, a closer look at the definition of the KSE casts doubt on the legitimacy of its use as a criterion for the presence of chaos. The definition is phrased in terms of the measure of subsets of the phase space and their time evolution and does not make reference to any notion that is connected to randomness. How can such a notion be an indicator for random behaviour?

Three suggestions have been made as to how to bridge this gap. First, connect the KSE to sensitive dependence on initial conditions (and thereby *de facto* reduce it to the second group of proposals); second, take algorithmic complexity to be a measure for randomness and relate the KSE to this notion, and; third, establish a link between the KSE and the information-theoretic notion of entropy. Among these options the third is the most widely used; the notion of information is almost habitually invoked when the interpretation of the KSE as a measure for random behaviour is discussed. Ironically, despite its frequent use and its undoubted attractiveness, it is the only one of these three proposals that has no theoretical grounding. In the first case, Pessin's theorem establishes a neat connection between the Liapunov exponents measuring the divergence of nearby trajectories and the KSE. (Roughly speaking, the theorem says that the KSE equals the sum over the positive Liapunov exponents.) In the second case, Brudno's theorem can be invoked, which basically says that if the phase space satisfies certain (unproblematic) conditions then the KSE is equal to the algorithmic complexity of almost all trajectories.

Surprisingly, there is no theorem which connects the KSE to the information-theoretic notion of entropy in roughly the same way in which Pessin's and Brudno's theorems link the KSE with Liapunov exponents and algorithmic complexity, respectively. Mathematically-minded authors either do not discuss the relationship between the KSE and communication theory at all (Arnold & Avez [1968], pp. 15–51; Cornfeld et al. [1982], pp. 246–57; Rudolph [1990], pp. 71–104; Cornfeld & Sinai [1980]); or they pay mere lip service to communication theory in that they attach the term 'information' to purely topological notions without elucidating how phase space topologies relate to the conceptual framework of information theory (Eckmann & Ruelle [1985], pp. 637–46; Keller [1998], pp. 43–50; Nadkarni [1998], pp. 63–4). Others seem to be aware of the fact that there is a gap to bridge but then content themselves with some rather loose and hand-waving remarks, mainly based on lax analogies (Billingsley [1965], 51–94; Mañé [1983], pp. 207–304; Parry [1981], pp. 52–73; Petersen [1983], pp. 227–48).[1] Nowhere in the literature could I find a clear argument connecting entropy in communication theory to its namesake in dynamical systems theory, despite the frequent mention of Shannon.

This raises the question of whether the interpretation of the KSE in terms of information can be vindicated. Is there a way to bridge the gap between the topological and the information-theoretic understanding of entropy? The aim of this paper is to provide such an argument for the case of Hamiltonian systems. More specifically, I prove that the KSE is equivalent to the information-theoretic notion of entropy given certain plausible assumptions. But before I embark on this project it seems worthwhile to appreciate what the problems are. First, communication theory and dynamical systems theory work with different conceptual frameworks. The former deals with a finite set of discrete messages and their combinations, while the latter considers a continuous measurable phase space on which an automorphism (a function mapping the phase space onto itself) is defined. Prima facie these two setups bear little,

[1] Petersen ([1983], pp. 229–34) discusses the problem in some detail but his arguments are of no help since the expression he gives for the entropy of a source is not correct.

if any, resemblance to each other. Second, the mathematical theories are entirely different. The KSE of an automorphism Φ is defined as $H_{\Phi} = \sup_{\alpha} \lim_{k\to\infty}(1/k)H(\alpha \vee \Phi\alpha \vee \cdots \vee \Phi^{k-1}\alpha)$ and there is not a single formula in communication theory that bears any similarity to this expression. For this reason, it is not possible to resort to formal analogy and consider information and topological measures as two interpretations of one calculus in the same way that classical probabilities and actual frequencies are interpretations of probability calculus, for instance. For these reasons, the question of how the concepts of entropy in these two disciplines fit together is not a trivial issue.

Critics might now ask, even if we grant that there is a problem, why we should bother with it. There are two ways to connect the KSE to unpredictability and randomness, why do we need a third one? Why not just stick with either Pessin's or Brudno's theorem? The answer to this question is that the different associations bring out different aspects of randomness that are all captured in the KSE. The connection of the KSE with communication theory adds to our understanding of the KSE because it brings out some aspects of the notion of randomness associated with the KSE that are particularly close to our physical intuitions and which are not made explicit by either Pessin's or Brudno's theorem. More to the point, based on communication-theoretic notions, one can make statements about the time span of unpredictability or the role that knowledge of the past history plays in making forecasts.

This gives rise to the following plan. After a short introduction into dynamical systems (Section **2**), I present the concept of entropy in communication theory (Section **3**) and discuss in what sense it can be considered a measure of random or unpredictable behaviour. Although I thereby follow the spirit of Shannon & Weaver ([1948]), the formal presentation differs significantly from theirs. I present a version of communication theory which is in several respects a generalization of the original theory. In particular, I consider messages whose probability of appearance is a function of the *entire* past history of the system, while Shannon and Weaver only consider Markov processes. The purpose of this is to facilitate the establishment of the connection to entropy in dynamical systems. This is what I do in Section **4**, where I prove an equivalence theorem for the KSE and the notion of entropy used in communication theory, this under the assumption that the measure defined on the phase space can be interpreted probabilistically. This establishes the sought-after connection between the two notions. On the basis of this result, I give a precise characterization of the kind of randomness we find in dynamical systems with positive KSE. I then compare this account to the notions of randomness we get from Pessin's or Brudno's theorems (Section **5**). In the last section, I point out that the main result of this paper has a bearing on the relation between product and process randomness, and I argue that it casts doubt on the recent claim that chaotic systems exhibit product but not process randomness.

2 Elements of dynamical systems theory

An *abstract dynamical system* is a triple $\mathcal{M} = (M, \mu, \Phi_t)$ where (M, μ) is a measure space equipped with a one-parameter group Φ_t of automorphisms of (M, μ), Φ_t depending measurably on t (Arnold & Avez [1968], p. 7).

It is common to assume that M has manifold structure but this is not necessary; in what follows it can be any measurable space. M will be referred to as a 'phase space' of the system. Sometimes this term is reserved for symplectic manifolds but I will use it in a more general sense and refer to every measurable space on which an automorphism is defined as a 'phase space'.

The parameter t plays the role of time. As can easily be seen, the set of all Φ_t has group structure. In the sequel, I will use the following notational conventions: $\Phi_t(x)$ is the point in phase space onto which Φ_t maps the 'initial condition' x after time t has elapsed, and $\Phi_t(A)$ is the image of the subset $A \subseteq M$ under Φ_t. I write $\Phi_{t_i \to t_j}(A)$ to denote the image of A under the time development starting at t_i and ending at t_j.

Furthermore I assume, without loss of generality, that M is normalized: $\mu(M) = 1$. The last clause in the definition (Φ_t depending measurably on t) simply means that $\Phi_t A \cap B$ is measurable for all t, and for all $A, B \in M$, i.e. that $\mu(\Phi_t A \cap B)$ is a measurable function of t. In what follows, 'almost everywhere' means 'everywhere except, perhaps, on a set of measure zero'.

As a simple example of a dynamical system in this sense, think of the unit interval endowed with Φ_t: the shift $x \to x + t \bmod 1$ where μ is the usual Euclidean length of an interval.

The above definition is extremely general and in what follows I make the following restrictions:

(1) The transformation Φ_t is measure-preserving: $\mu(\Phi_t A) = \mu(A)$ for all subsets A of M and all times t. That is, in the sequel I restrict my attention to Hamiltonian systems. Some may feel a certain unease about this limitation because they tend to think about chaos in terms of attractors, which cannot occur in Hamiltonian systems. Those may become reconciled by the following two observations. First, important paradigm cases of chaotic systems are Hamiltonian systems, for instance the three-body problem, the Hénon-Heiles system, the autonomous double pendulum, and more generally KAM-type systems. Second, apart from attractors, which are ruled out by the conservation of phase volume, Hamiltonian systems can exhibit all features that are commonly taken to be distinctive of chaotic systems: Positive Liapunov exponents, sensitive dependence on initial condition, unpredictable time evolution, continuous power spectra, decaying autocorrelations, aperiodic orbits, the presence of a stretching and folding mechanism in phase space, and last but not least positive KSE.

(2) Nothing has been said so far about whether the parameter t is continuous or discrete. To keep things simple I restrict my attention to the discrete case; that is, I only consider systems in which time evolves in finite steps: t_1, t_2,…. Moreover, it is often the case that the Φ_{t_i}, $i = 1,2,\ldots$ are generated by an iterative application of one single automorphism Φ (standard examples like the cat map or the baker's transformation belong to this class). In this case we have $\Phi_{t_i} = \Phi^i$ and $\Phi_{t_i \to t_j}(A) = \Phi^{j-i}(A)$. Furthermore, I drop the subscript and just write (M, μ, Φ) for a dynamical system of this kind.

In what follows, partitions will play an important role. Roughly, a partition of M is a division of M into *finitely* many measurable sets. More precisely, a partition $\alpha = \{\alpha_i | i = 1, \ldots, n\}$ is a collection of non-empty, non-intersecting measurable sets that together cover M: the α_i are pairwise disjoint, $\alpha_i \cap \alpha_j = \emptyset$ for all $i \neq j$; and together the α_i cover M up to measure zero, $\mu(M - \bigcup_{i=1}^{n} \alpha_i) = 0$. The α_i are called 'atoms' or 'cells' of the partition. Furthermore, notice that if α is a partition of M then its picture under the automorphism Φ_t is also a partition. That is, if $\alpha = \{\alpha_i | i = 1, \ldots, n\}$ is a partition of M then $\Phi_t\alpha := \{\Phi_t\alpha_i | i = 1, \ldots, n\}$ is as well.

Given two partitions $\alpha = \{\alpha_i \,|\, i = 1, \ldots, n\}$ and $\beta = \{\beta_j | j = 1, \ldots, m\}$, their *least common refinement* $\alpha \vee \beta$ is defined as follows: $\alpha \vee \beta = \{\alpha_i \cap \beta_j | i = 1, \ldots, n;\ j = 1, \ldots, m\}$. Sometimes $\alpha \vee \beta$ is also called 'sum of α and β'. Figure 1 provides an example illustrating this.

We are now in a position to state the definition of the Kolmogorov-Sinai entropy H_Φ of an automorphism Φ (Arnold & Avez [1968], pp. 38–40):

$$H_\Phi := \sup_{\alpha} \lim_{k \to \infty} (1/k) H(\alpha \vee \Phi\alpha \vee \cdots \vee \Phi^{k-1}\alpha), \tag{1}$$

where the function on the right-hand side is the entropy of a partition (recall that $\alpha \vee \Phi\alpha \vee \cdots \vee \Phi^{k-1}\alpha$ is a partition as well) which is defined as follows: $H(\beta) := -\sum_{i=1}^{m} z[\mu(\beta_i)]$, $z(x) = x\log(x)$ if $x > 0$ and $z(x) = 0$ if $x = 0$; and $\sup_\alpha$ is the supremum over all possible finite partitions α of the phase space. (I shall discuss this definition in detail later on.)

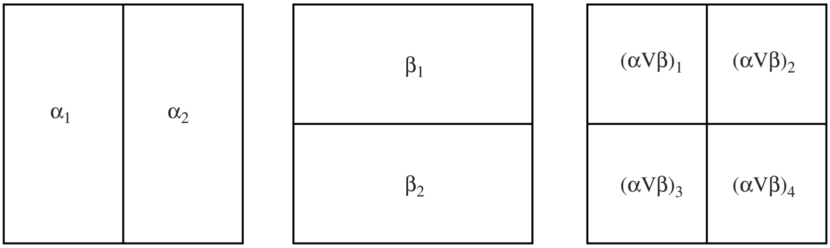

Figure 1. The sum of the partitions α and β

3 Entropy in communication theory

Consider the following situation: we have a source S producing messages which are communicated to a receiver R registering them (on a paper tape, for instance).[2] The messages may be of various types: sequences of letters, numbers, words, or any other symbols we might think of. The only restriction we impose is that the source uses discrete symbols and generates the message symbol by symbol. The product of this process is a string of symbols, the message, which can be of finite or infinite length.

More precisely, let $S_1, \ldots, S_n$ be the available symbols and let the process of the composition of a message start at time t_0. At that time, no symbol has been produced by S and the tape of R is blank. The first symbol is put out by S and sent to R at t_1, where it is registered; the second symbol is sent and registered at t_2 (we assume that $t_1 < t_2 < t_3 < \ldots$), and so on. The production of one symbol by the source (when it moves from t_i to t_{i+1}) will be referred to as a 'step'. As a result of this, at t_k the tape of R contains a string of k symbols: $S_{l_1}^{t_1} S_{l_2}^{t_2} \ldots S_{l_k}^{t_k}$, where all the l_i range over $1, \ldots, n$ (i.e. the number of symbols available). The time-superscripts have been added to indicate the order of reception (S_{l_1} has been sent and received at t_1, and so on). For instance, assuming that our symbols are letters, $h^{t_1} e^{t_2}\ l^{t_3} l^{t_4} o^{t_5}$ means that the letter 'h' has been received at time t_1, 'e' at t_2, and so on.

I now introduce in seven stages the notion of the entropy of a source S, which will be designed such that it serves as a measure of the receiver's average uncertainty about what message the source produces next (that is, it is a measure for unpredictability—I shall use 'uncertainty' and 'unpredictability' interchangeably in what follows). Thereby it also is a measure of the amount of information received. I shall come to that below.

Stage 1: To start with, consider a source which has just two symbols (0 and 1, say) it can send. What is the amount of uncertainty of the receiver about what message will crop up next? We answer this question by adopting the convention that the amount of uncertainty of the receiver in this case equals one. This is a reasonable choice: if we had just one symbol available, we would in fact know for sure what the receiver would indicate whenever we switched it on; there is no uncertainty. The simplest non-trivial situation is the one considered here, where two symbols are available. In this case, we are not sure what the next message will be (we could get either of the two messages), and it seems reasonable to say that the amount of uncertainty in this case is one because there is one choice to be made.

Before continuing with the development of the theory, I would like to make some remarks about information. It is one of the main insights of Shannon's *Mathematical Theory of Communication* that uncertainty is closely related to

[2] In what follows I assume that the channel is noiseless and deterministic, which basically means that there is a one-to-one correspondence between the input and the output messages.

information. If we are sure about what message we will receive next, we do not learn anything by actually receiving it. Therefore the amount of information transmitted is zero. If, on the other hand, there are several possibilities (e.g. if we do not know whether we will obtain 0 or 1), we do acquire information when we receive either of the two. For those who find this a bit contrived, consider Lewis Carroll's remark in *Through the Looking-Glass*: 'It is a very inconvenient habit of kittens (Alice had once made the remark) that, whatever you say to them, they always purr. "If they would only purr for 'yes' and mew for 'no,' or any rule of that sort," she had said, "so that one could keep up a conversation! But how *can* you talk with a person if they *always* say the same thing?" ' ([1998], p. 238). In short, uncertainty about what comes next and the transmission of information are two sides of the same coin.

For this reason, devising a measure of the amount of uncertainty about future events and the quantity of information transmitted amounts to the same. Consider again the previous example. What is the amount of information transmitted when R has registered '1' or '0' on its tape? For the same reasons outlined above, it is natural to say that the amount of information transmitted is one (in technical jargon, we get one 'binary digit', or 'bit' for short, of information). As a consequence, when, in what follows, we devise entropy as a measure of the amount of uncertainty, it also is a measure of the amount of information transmitted. However, the focus in the rest of the paper will be on uncertainty, not on information, though I use information jargon at times if this turns out to be convenient.

Two remarks about this concept of information should be made. First, contrary to the concept of meaning, which applies to a single message (receiving 'S_1', for instance, could mean, 'I love you', 'I hate you', 'happy birthday' or what have you), information is not concerned with individual messages, but only with the ensemble of *all messages a source could possibly send*. What makes a single message informative is not its meaning but the fact that it is selected from a set of possible messages. The more (different) messages the source could in principle send, the higher the information content of the one we actually get. Second, from what has been said so far, it is obvious that we are dealing here with a rather idiosyncratic concept of information which has little to do with the various senses the term 'information' has in ordinary discourse, such as knowledge or propositional content. Information in these senses has semantic features such as truth-values, something the communication-theoretic concept lacks. This has led many to criticize this concept as misconceived. Be this as it may, my focus in this paper is on uncertainty and not information, and for this reason I will not dwell on this issue here.

Stage 2: How do the ideas introduced in Stage 1 generalize to a source which can emit n different symbols? This question is best answered by looking at the restrictions we want to impose on a measure of the amount of uncertainty. (a) It must be a monotonically increasing function of n; the more possibilities there are, the greater our uncertainty about what comes next. (b) Additivity: When we add two sources of the same type, we want the amount of uncertainty to double. Informally, let I stand for uncertainty (or information) per step. Then we require I(source 1 + source 2) = I(source 1) + I(source 2) (see Shannon & Weaver ([1949], p. 32), for a justification of this assumption). (c) Finally, we should have $I = 1$ in the case of only two possibilities (see Stage 1). The only function which satisfies these criteria is the logarithm to the base 2. Hence, we have $I = \log(n)$, where 'Log' stands for the logarithm to the base 2.

Stage 3: So far we have implicitly assumed that all symbols occur with equal probability at any step k, i.e. that all S_i occur with probability $p_k = 1/n$ at step k (since the probabilities p_k do not actually depend on step k, I drop the subscript k in what follows). This is a perfectly good assumption in certain cases but it does not hold true generally. If, for instance, the symbols are letters of the alphabet, the probability that the next letter the sender emits is an 'a' is much higher than that for an 'x', since 'a' occurs much more often in English than 'x'. So we need a generalization of I to this case. Let $p(S_1), \ldots, p(S_n)$ (where $p(S_1) + \cdots + p(S_n) = 1$) be the respective probabilities that $S_1, \ldots, S_n$ occur. Shannon showed that a natural generalization of the above notion is the following (Shannon & Weaver [1949], pp. 48–53):

$$H_{step} := -\sum_{i=1}^{n} z[p(S_i)], \tag{2}$$

where $z(x) = x\log(x)$ if $x > 0$ and $z(x) = 0$ if $x = 0$. H_{step} is the measure of the uncertainty about what symbol will crop up at the next step; the greater H_{step} the less certain we are about the outcome. The use of the letter '*H*' instead of '*I*' is motivated by the fact that Eqn (2) has the same structure as the expression for the entropy in statistical mechanics, and for this same reason we also refer to it as 'entropy'. H_{step} is a natural generalization of I for the following reasons. First, if all events are equally probable ($p(S_i) = 1/n$ for all $i = 1, \ldots, n$), H_{step} coincides with the above notion, that is $H_{step} = \log(n)$, as some simple algebra immediately reveals. Second, it is continuous in the $p(S_i)$. Third, it has the 'right' behaviour: (a) Any change toward equalization of the probabilities $p(S_1), \ldots, p(S_n)$ increases H. In particular, H_{step} is maximal if all events are equally probable. (b) $H_{step} = 0$ if all $p(S_i)$ but one equal zero, i.e. if there is in fact no choice (*Ibid.*, p. 51).

Stage 4: So far nothing has been said about what the $p(S_i)$ are and on what they depend. In many cases the choice of a symbol at some particular time t_{k+1} does not depend on previous choices. However, for a general source, the probability that a particular symbol is chosen may depend on what has been chosen beforehand. For example, if the source is producing English prose, there are a number of limitations due to the orthography and syntax of the language. The probability of receiving a ‘u’, for instance, will rise dramatically each time a ‘q’ comes through, and it will be almost zero after an ‘x’. In short, successive symbols may not be chosen independently and their probabilities may depend on preceding letters. In the simplest case, a so-called Markov process, a choice depends only on the preceding letter and not on the ones before that. However, the choice may in general depend on the *entire previous history* of the process; that is, the choice of a symbol at t_{k+1} may depend on $S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k}$. It is natural to account for this by using conditional probabilities: the probability of receiving S_i at time t_{k+1} is $p(S_i^{t_{k+1}}/S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k})$. Since these probabilities may vary with k, the entropy may have a different value at every step. To make this explicit, I replace the subscript ‘step’ in Eqn (2) by ‘k’ to emphasise that we are considering the entropy at the k^{th} step of the process. The expression for the entropy then reads:

$$H_k(S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k}) := -\sum_{i=1}^{n} z[p(S_i^{t_{k+1}}/S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k})]. \tag{3}$$

This is a measure for the uncertainty about what symbol will show up at time t_{k+1} given that the previous history of the process (recorded on R’s tape) is $S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k}$.

Stage 5: Now we generalize our question slightly. Instead of asking ‘What is the uncertainty about the $(k+1)^{\text{th}}$ symbol *given that* the message produced so far is $S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k}$?’ (to which Eqn (3) is an answer), we now ask ‘What is the uncertainty about the $(k+1)^{\text{th}}$ symbol *whatever* message has been produced so far?’. Or in other words: What is the uncertainty at t_{k+1} if we do not presuppose that the system has a particular previous history, namely $S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k}$? The answer seems clear: take the average of all $H_k(S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k})$ and, to do justice to the fact that not all histories are equally likely, weight each term with the probability of the respective history:

$$\bar{H}_k := \sum_{l_1,\dots,l_k=1}^{n} p(S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k})H_k(S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k}), \tag{4}$$

where

$$p(S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k}) := p(S_{l_1}^{t_1})p(S_{l_2}^{t_2}/S_{l_1}^{t_1})\dots p(S_{l_k}^{t_k}/S_{l_1}^{t_1}\cdots S_{l_{k-1}}^{t_{k-1}}). \tag{5}$$

Stage 6: On the basis of this result we can now define the entropy $\tilde{H}_k$ of the entire process of the composition of a message of length k. Since no step is privileged over the others, this can be effected by simply taking the average of the entropy at every step of the process:

$$\tilde{H}_k := \frac{1}{k}\sum_{j=0}^{k-1}\bar{H}_j \tag{6}$$

Stage 7: Now we can say that the entropy of the source itself, H_S, is the average of the uncertainty at every step if we let the process go on forever:

$$H_S := \lim_{k\to\infty}\tilde{H}_k \tag{7}$$

This is the *entropy of the source*. It is a measure for the average uncertainty over the entire process or, to put it differently, the average amount of information which the source conveys with every symbol the receiver prints. I will refer to this notion of entropy as 'communication-theoretic entropy', or CTE for short.[3]

From a technical point of view, the development of the theory is now complete. But what is its conceptual import? What does it mean for a source to have a positive CTE? And in what sense is a positive CTE a measure for random behaviour? In the remainder of this section, I shall discuss these questions.

Let us start by having a look at the probabilities involved. When probabilities are introduced into the theory, they are assumed to be given; there is no element in the theory that determines what their values are. For this reason one could also say that the set of possible messages $S_1, \dots, S_n$ together with the conditional probabilities of occurrence $p(S_i^{t_{k+1}}/S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k})$, $i = 1, \dots, n$ actually *defines* the source. Characteristically, these probabilities are past relative frequencies, and it is assumed that these relative frequencies will persist.

However, this 'natural order' of proceeding can, in a certain sense, be reversed: the entropy can be used to characterize the probabilities involved even if they are not explicitly known (leaving aside the question of how we get to know the entropy without knowing the probabilities). The point is the following. The notion of entropy has been set up in such a way that it is a measure for the average uncertainty per symbol over the entire process. For this reason, $H_S > 0$ expresses the fact that, on average, at every step there is some uncertainty about what the next symbol printed by the receiver will be.

[3] Note that if we assume that all the probabilities are independent (this is the case for Bernoulli processes, for instance), we have $H_S = H_{step}$. This is easy to see: for independent events, Eqn (4) becomes

$$\bar{H}_k := -\sum_{l_1,\dots,l_k=1}^{n} p(S_{l_1})p(S_{l_2})\dots p(S_{l_k})\sum_{i=1}^{n} z[p(S_i)]. \tag{8}$$

Now realize that the two sums separate and that the first one is just a sum over the probabilities of all strings of length k. For this reason we have: $\sum_{l_1,\dots,l_k=1}^{n} p(S_{l_1})p(S_{l_2})\dots p(S_{l_k}) = 1$. Hence $\bar{H}_k(S) = -\sum_{i=1}^{n} z[p(S_i)]$ and therefore $H_S = -\sum_{i=1}^{n} z[p(S_i)]$.

More precisely, whatever the past history of the system and whatever our knowledge about it (we may know it all), we are not sure as to what the next symbol that emerges will be. And this characteristic persists forever, there exists no 'cut-off time' t_c in the process from which on the past history of the system allows us to predict with certainty what its future will be.[4] This follows immediately from the definition of the CTE: $H_S = \lim_{k\to\infty}(1/k)\sum_{j=0}^{k-1}\bar{H}_j = \lim_{k\to\infty}(\bar{H}_0/k + \cdots + \bar{H}_{k-1}/k)$ is greater than zero only if there do not cease to be $\bar{H}_k$ greater than zero. Now recall that $\bar{H}_k$ is a measure for the uncertainty about what the message printed at time $k+1$ will be. Hence, if there do not cease to be $\bar{H}_k > 0$ as time goes on, there will always be times at which we are not sure about what is going to happen. As a consequence, we cannot predict with certainty what the future will be.[5] In terms of probabilities, this means that as the process goes on we never reach a stage where $p(S_i^{t_{k+1}}/S_{l_1}^{t_1}S_{l_2}^{t_2}\ldots S_{l_k}^{t_k})$ equals one for some particular symbol and zero for all the others.

Summing up, we can characterize a system with positive entropy as one in which the past history never conveys certainty onto what will happen at the next step and more generally in the future. Or to phrase it differently, even given the entire past history we are not able to predict with certainty what will happen in the future. It is in this sense that a process with positive entropy is random, and the magnitude of the entropy is a measure of how random it is.

If, on the other hand, H_S equals zero, then, on average, there is no uncertainty and we can predict what the future will look like. There is a subtlety, however. Zero entropy does not imply that the process is deterministic (by which I here simply mean that, given the state of a process at time t_k, there is exactly one state in which it can be at t_{k+1}). It is true that for a deterministic process $H_S = 0$ holds. But the converse is false: $H_S = 0$ does not imply that the process is deterministic: it just means that *on average* there is no freedom of choice. This does not preclude that the process is indeterministic at some particular instants of time.

4. Entropy in dynamical systems theory

To repeat, the KSE of an automorphism is defined as $H_\Phi = \sup_\alpha \lim_{k\to\infty}(1/k)H(\alpha \vee \Phi\alpha \vee \cdots \vee \Phi^{k-1}\alpha)$ and it is commonly used as a measure for the unpredictability of the dynamics. But as I explained in the

[4] I should note that there is a subtle difference between 'with probability equal to one' and 'certainty'. The latter implies the former but not vice versa. However, since this subtlety does not play any role in what follows, I shall use these two expressions interchangeably.

[5] This does not imply that there cannot be an isolated instant of time for which this is possible. As the above formula shows, $H_S > 0$ is compatible with there being, from time to time, some particular $\bar{H}_k$ that equals zero. The point is just that as k, i.e. time, goes on we never reach a point after which *all* $\bar{H}_k$ equal zero; and this is all we need to render the future unpredictable.

introduction, it is prima facie not clear whether this is legitimate or not. In this section I show that it *is*, by proving, under plausible assumptions, that the KSE is equivalent to the CTE.

Wanting to prove this equivalence we face the following problem: the messages we have been dealing with so far are discrete entities, whereas the phase space of a dynamical system is continuous. These do not seem to go together. This mismatch can be removed if we coarse-grain the phase space, i.e. if we work with a partition instead of the 'whole' phase space. Then it is no longer difficult to associate a dynamical system $\mathcal{M}$ with an information source of the type discussed above. This association is achieved as follows. Let $\alpha = \{\alpha_1, \ldots ,\alpha_n\}$ be a partition of the phase space and assume that the state of the system at t_0 is x (more needs to be said about the choice of a partition; I come back to this issue below at stage 8). Then trace the trajectory $\Phi_{t_i}(x)$ of x and take down on a paper tape at each time $t_i, i = 1,2,\ldots$, in what cell $\alpha_j, j = 1, \ldots, n$ of the partition $\Phi_{t_i}(x)$ is. That is, write down $\alpha_j^{t_1}$ if $\Phi_{t_1}(x) \in \alpha_j$ at time t_1 and so on. If we follow this procedure up to time t_k this generates the string $\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\cdots\alpha_{l_k}^{t_k}$, which is structurally identical to $S_{l_1}^{t_1}S_{l_2}^{t_2}\cdots S_{l_k}^{t_k}$. This is illustrated in Figure 2.

Now we need to find something in $\mathcal{M}$ corresponding to the probability $p(S_i)$ of choosing a particular symbol S_i. This is not too hard to get. By assumption, there is a normalized measure μ on M (that is $\mu(\mathrm{M}) = 1$) and it is a straightforward move to interpret this measure as a probability measure. More precisely, let μ reflect our ignorance about the real state of the system, and interpret $\mu(\alpha_i)$ as the probability of finding the system's state in α_i. Note, however, that although this move is quite natural, it is not necessary to interpret μ as the probability of finding the system's state in a particular cell of the partition. Not all measures reflect our ignorance about the system's real state; it could also simply be the spatial volume. However, this interpretation is perfectly *possible*, and it allows us to connect what happens in dynamical systems to communication theory as outlined above, and that is all we need for the time being.

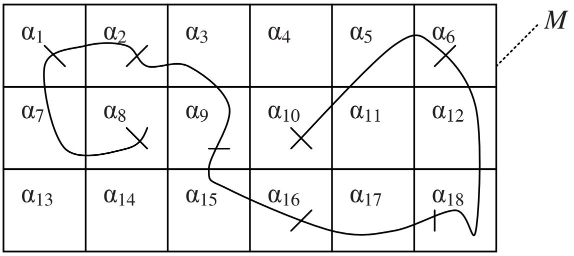

Figure 2. The generation of the string $\alpha_8^{t_1}\alpha_1^{t_2}\alpha_2^{t_3}\alpha_9^{t_4}\alpha_{16}^{t_5}\alpha_{18}^{t_6}\alpha_6^{t_7}\alpha_{10}^{t_8}$

Then, the following associations are made to connect dynamical systems to communication theory:

(a) The atoms α_i of the partition correspond to the symbols (messages) S_i of the source.
(b) The measure of an atom $\mu(\alpha_i)$, interpreted as the probability of finding the system's state in atom α_i, corresponds to the probability $p(S_i)$ of obtaining symbol S_i.
(c) The automorphism Φ_t corresponds to the source S, since they both do the job of generating the strings $\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\dots\alpha_{l_k}^{t_k}$ and $S_{l_1}^{t_1}S_{l_2}^{t_2}\dots S_{l_k}^{t_k}$ respectively.

With these associations at hand, it is now possible to carry over the notions introduced in the last section to the present context.

Stage 1: To begin with, consider a partition consisting of two atoms, $\alpha = \{\alpha_1, \alpha_2\}$, and assume that the state of the system at t_0 is x. Then trace the trajectory $\Phi_{t_i}(x)$ and take down at each time step in what cell $\alpha_j, j = 1,2$, $\Phi_{t_i}(x)$ is. This generates a string of exactly the same sort as the one we obtain from a source which can send only two symbols.

As in the case of the source, we are generally not sure what cell the system's state will be in next. Due to restrictions on the precision of the specification of initial conditions, we normally cannot know precisely what the system's initial state is, and this uncertainty is then propagated, or even amplified, by the dynamics of the system as time evolves. Therefore we gain information, i.e. remove uncertainty, when we learn that the system's state is in α_1 rather than α_2, say. Adopting the same convention as in the case of sources, we can say that the amount of uncertainty of the observer regarding what cell the system's state will be in at the next step is one bit.

Stage 2: Replacing this partition by one consisting of more than two atoms is the analogue to the transition from a source with two symbols to one with any number of symbols. The considerations concerning the general properties of information then carry over one-to-one. The more possibilities we have (i.e. the more cells the partition consists of), the greater the uncertainty about what happens next becomes. Combining two systems of the same sort should result in the doubling of the amount of uncertainty; and in the case of a partition with only two cells, the uncertainty must be unity in order to be consistent with stage 1. So, as in the case of the source, we set $I = \log(n)$.

Stage 3: By assumption, the measure $\mu(\alpha_j)$ is interpreted as the probability that at the next step the system will be in cell α_j; that is, we have $p(\alpha_j) = \mu(\alpha_j)$. This allows us to carry over Eqn (2) to the present context. We immediately obtain $H_{step} = -\sum_{i=1}^{n} z[\mu(\alpha_i)]$, which is commonly called the 'entropy of the partition α'. To be in accord with the notation commonly used in the literature, I write 'H(α)' instead of 'H_{step}':

$$H(\alpha) := -\sum_{i=1}^{n} z[\mu(\alpha_i)]. \tag{9}$$

Stage 4: In general, in dynamical systems also, the previous history affects future probabilities. Therefore Eqn (3) carries over to the present context:

$$H_k(\alpha;\ \alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\dots\alpha_{l_k}^{t_k}) := -\sum_{i=1}^{n} z[p(\alpha_i^{t_{k+1}}/\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\dots\alpha_{l_k}^{t_k})]. \tag{10}$$

Now we have to express the probabilities occurring in this expression in terms of μ. To this end, first spell out the conditional probability in terms of unconditional ones:

$$p(\alpha_i^{t_{k+1}}/\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\dots\alpha_{l_k}^{t_k}) = p(\alpha_i^{t_{k+1}}\ \&\ \alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\dots\alpha_{l_k}^{t_k})/p(\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\dots\alpha_{l_k}^{t_k}) \tag{11}$$

$$= p(\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\dots\alpha_{l_k}^{t_k}\alpha_i^{t_{k+1}})/p(\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\dots\alpha_{l_k}^{t_k}). \tag{12}$$

The latter equality follows immediately from the definition of a string.

Note that for any two instants of time t_i and t_j (where $t_i < t_j$) and any two subsets A and B of M the following holds:

$$p(A^{t_i}B^{t_j}) = \mu[\Phi_{t_i\to t_j}(A)\cap B] \tag{13}$$

The validity of this equation becomes transparent in Figure 3.

The generalization of this equality to any number of sets and instants of time is straightforward. Applying this generalisation to the above expressions yields:

$$p(\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\dots\alpha_{l_k}^{t_k}) = \mu(\alpha_{l_k}\cap\Phi_{t_{k-1}\to t_k}\alpha_{l_{k-1}}\cap\dots\cap\Phi_{t_1\to t_k}\alpha_{l_1}) \tag{14}$$

and

$$p(\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\dots\alpha_{l_k}^{t_k}\alpha_i^{t_{k+1}}) = \mu(\alpha_i\cap\Phi_{t_k\to t_{k+1}}\alpha_{l_k}\cap\dots\cap\Phi_{t_1\to t_{k+1}}\alpha_{l_1}). \tag{15}$$

Hence (10) becomes:

$$H_k(\alpha;\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\dots\alpha_{l_k}^{t_k}) := -\sum_{i=1}^{n} z\left[\frac{\mu(\alpha_i\cap\Phi_{t_k\to t_{k+1}}\alpha_{l_k}\cap\dots\cap\Phi_{t_1\to t_{k+1}}\alpha_{l_1})}{\mu(\alpha_{l_k}\cap\Phi_{t_{k-1}\to t_k}\alpha_{l_{k-1}}\cap\dots\cap\Phi_{t_1\to t_k}\alpha_{l_1})}\right]. \tag{16}$$

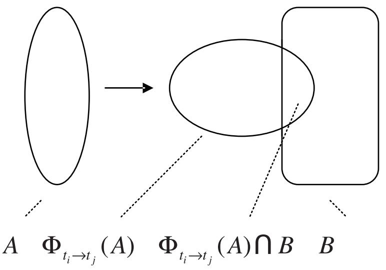

Figure 3. The probability of $A^{t_i}B^{t_j}$ equals $\mu[\Phi_{t_i\to t_j}(A)\cap B]$

Stage 5: Similarly, Eqn (4) carries over to dynamical systems easily:

$$\bar{H}_k(\alpha) := \sum_{l_1,\ldots,l_k=1}^{n} p(\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\ldots\alpha_{l_k}^{t_k})H_k(\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\ldots\alpha_{l_k}^{t_k}); \tag{17}$$

which is the entropy of the k^{th} step relative to the partition α. Inserting probabilities in terms of the measures, we obtain

$$\begin{aligned}\bar{H}_k(\alpha) := \sum_{l_1,\ldots,l_k=1}^{n} & \mu(\alpha_{l_k} \cap \Phi_{t_{k-1}\to t_k}\alpha_{l_{k-1}} \cap \cdots \cap \Phi_{t_1\to t_k}\alpha_{l_1}) \\ & \sum_{i=1}^{n} z\left[\frac{\mu(\alpha_i \cap \Phi_{t_k\to t_{k+1}}\alpha_{l_k} \cap \cdots \cap \Phi_{t_1\to t_{k+1}}\alpha_{l_1})}{\mu(\alpha_{l_k} \cap \Phi_{t_{k-1}\to t_k}\alpha_{l_{k-1}} \cap \cdots \cap \Phi_{t_1\to t_k}\alpha_{l_1})}\right].\end{aligned} \tag{18}$$

Stage 6: The entropy of the process of the composition of a string of length k is:

$$\tilde{H}_k(\alpha) := \frac{1}{k}\sum_{j=0}^{k-1}\bar{H}_j(\alpha). \tag{19}$$

Stage 7: On the basis of this, we define the entropy of an automorphism as follows:

$$H_{\Phi_t}(\alpha) := \lim_{k\to\infty}\tilde{H}_k(\alpha) \tag{20}$$

This is the entropy of the automorphism Φ_t *with respect to the partition* α.

Stage 8: At the beginning of this section I mentioned in passing that more needs to be said about the choice of a partition. The reason for this is that there is an important disanalogy between a source and a dynamical system. In the case of a Source S, the set of possible messages $(S_1,\ldots,S_n)$ is a part of the definition of the source, and hence it is no longer an issue later on. This is not so with the partition α, which is no constitutive part of the dynamical system $\mathcal{M}$. Rather, it has been 'imposed' on the system.

This is a problem because the values we obtain for $H_{\Phi_t}(\alpha)$ essentially depend on the choice of the partition α. If we choose α conveniently enough (which we always can, no restrictions having been imposed on α), there will be no uncertainty left, whatever the properties of Φ_t (this can be achieved, for instance, by choosing the trivial partition, i.e. the partition whose only atom is M itself). Hence, prima facie $H_{\Phi_t}(\alpha)$ tells us more about our choice of α than about the properties of the automorphism Φ_t.

This may pose no problem if the partition α is what we are ultimately interested in. But for the most part we are interested in the automorphism Φ_t itself rather than the partition, which is merely an auxiliary device. For this reason we have to eliminate the dependence on α and get to a notion of entropy which is no longer dependent on any particular partition.

This can be achieved by defining the *entropy of the automorphism* as the supremum of $H_{\Phi_t}(\alpha)$ over all finite measurable partitions (Arnold & Avez [1968], p. 40):

$$H_{\Phi_t} = \sup_{\alpha} H_{\Phi_t}(\alpha). \tag{21}$$

The choice of the supremum is motivated by the following considerations. From the point of view of the dynamical system, there is no privileged partition—one is just as good as any other. Therefore it is interesting to discuss how $H_{\Phi_t}(\alpha)$ behaves as a function of α, when α ranges over all finite measurable partitions. As I just observed, one can always find a partition such that $H_{\Phi_t}(\alpha) = 0$; and from the definition of $H_{\Phi_t}(\alpha)$ it follows that it cannot be negative. Hence zero is an infimum of $H_{\Phi_t}(\alpha)$. However, this is not a very informative result if we want to know something about the automorphism Φ_t, since this holds true regardless of what Φ_t is. So what about the supremum? This is an interesting question because the supremum really depends on Φ_t. Some automorphisms are such that we simply cannot find a partition with respect to which there is much uncertainty, while with others things get as unpredictable as we may want. For this reason the supremum of $H_{\Phi_t}(\alpha)$, unlike the infimum, tells us a great deal about the automorphism. More specifically, it informs us about the maximal magnitude of uncertainty we can encounter in a system governed by Φ_t.

But there is a problem: the expression in Eqn (21) does not bear any resemblance whatsoever to the standard definition of the KSE. I now solve this problem by proving that H_{Φ_t}, as defined above, is equivalent to the standard definition.

To this end, I first have to introduce a technical device, the so-called conditional entropy. Let α and β be two partitions. The *conditional entropy of* α *with respect to* β is defined as follows (Arnold & Avez [1968], p. 37):

$$H(\alpha/\beta) := \sum_{j=1}^{m} \mu(\beta_j) \sum_{i=1}^{n} z\left[\frac{\mu(\alpha_i \cap \beta_j)}{\mu(\beta_j)}\right]. \tag{22}$$

Then note that the standard definition of the KSE assumes that the flow is generated by the iterative application of the same automorphism Φ. So we have $\Phi_{t_i} = \Phi^i$ and $\Phi_{t_i \to t_j}(A) = \Phi^{j-i}(A)$ (see Section **2**). Given this, one can prove **Theorem 1** (the proof can be found in the Appendix):

$$\bar{H}_k(\alpha) = H(\alpha/\Phi\alpha \vee \Phi^2\alpha \vee \cdots \vee \Phi^k\alpha). \tag{23}$$

Then the entropy of the process as given in Eqn (19) reads:

$$\tilde{H}_k(\alpha) = \frac{1}{k}\left[H(\alpha) + H(\alpha/\Phi\alpha) + \cdots + H(\alpha/\Phi\alpha \vee \cdots \vee \Phi^{k-1}\alpha)\right]. \tag{24}$$

This can be facilitated considerably by using **Theorem 2** (the proof can be found in the Appendix):

$$H(\alpha \vee \Phi\alpha \vee \cdots \vee \Phi^k\alpha) = H(\alpha) + H(\alpha/\Phi\alpha) + \cdots + H(\alpha/\Phi\alpha \vee \cdots \vee \Phi^k\alpha). \tag{25}$$

Hence,

$$\tilde{H}_k(\alpha) = \frac{1}{k} H(\alpha \vee \Phi\alpha \vee \cdots \vee \Phi^{k-1}\alpha). \tag{26}$$

Inserting Eqn (26) first into (20) and then (21) we obtain

$$H_\Phi = \sup_\alpha \lim_{k\to\infty} (1/k) H(\alpha \vee \Phi\alpha \vee \cdots \vee \Phi^{k-1}\alpha) \tag{27}$$

and this is the definition of the entropy of an automorphism which we were aiming for. Gathering the pieces together, we have proven the

Equivalence Theorem:

$$\begin{aligned} H_\Phi &= \sup_\alpha \lim_{k\to\infty} (1/k) H(\alpha \vee \Phi\alpha \vee \cdots \vee \Phi^{k-1}\alpha) \\ &= \sup_\alpha \lim_{k\to\infty} \frac{-1}{k} \sum_{j=0}^{k-1} \sum_{l_1,\ldots,l_k=1}^{n} p(\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\ldots\alpha_{l_k}^{t_k}) \\ &\quad \sum_{i=1}^{n} z[p(\alpha_i^{t_{k+1}}/\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\ldots\alpha_{l_k}^{t_k})]. \end{aligned} \tag{28}$$

Since, by construction, the last term in this equation is equivalent to the CTE, the sought-after connection between the notion of entropy in dynamical system theory and in information theory is established.

As a consequence, everything that was said at the end of Section **3** about the unpredictable behaviour of a source can be carried over to dynamical systems one-to-one. However, a proviso with regard to the choice of a partition must be made. The exact analogue of the CTE is $H_\Phi(\alpha)$ and not H_Φ, which is defined as the supremum of $H_\Phi(\alpha)$ over all partitions α. For this reason, the characterization of randomness devised in the context of communication theory strictly speaking applies to $H_\Phi(\alpha)$ rather than H_Φ. However, there is a close connection between the two: whenever $H_\Phi > 0$, there is trivially at least one partition for which $H_\Phi(\alpha) > 0$. In this case, Φ is random in precisely the way described above with respect to this partition, and more generally with respect to all partitions for which $H_\Phi(\alpha) > 0$. For this reason, statements about H_Φ and $H_\Phi(\alpha)$ naturally translate into one another.

This said, we obtain the following characterization: if an automorphism has positive KSE, then whatever the past history of the system, we are on average not able to predict with certainty in what cell of the partition the system's state will lie next. And this will be the case forever: there is no ‘cut-off time’ after which we have gathered enough information to predict what will happen in the entire future. We can collect as much knowledge about the system’s past as we like and we are still left uncertain about its future. On average, we are just never sure about what happens next, since the past history does not convey certainty onto what will happen in the future (however, we may be certain of what happens at the next step at some isolated instants of time). In short, the past history does not determine the future.

Moreover, the magnitude of the KS entropy is a measure of how great our failure to predict the future will be; the greater the entropy, the more uncertain we are about the future.

From this it follows immediately that the dynamics obey the so-called 0–1 law of probability theory. This law states that even if we have complete knowledge of the process’ behaviour in the past, the only events which we can predict with certainty at the next step are those which have a probability of either 0 or 1 *independently* of the past history (see Batterman [1993], p. 60).

5 Comparison with other accounts

In the last section, I presented a discussion of the unpredictability we find in systems with positive KSE. This characterization is, to a large extent at least, not new. However, it is only the link between the KSE and the information-theoretic notion of entropy which makes a justification of this characterization possible. In other words, it is the result obtained in the previous section that puts this characterization on firm ground. In this section, I briefly show why this is so.

As I mentioned at the beginning, there are two other methods to link the KSE with unpredictability or randomness: (1) Pessin’s theorem, which relates the KSE to positive Liapunov exponents or (2) Brudno’s theorem, which connects it to algorithmic complexity. I will now briefly discuss these options and explain where the differences between them and an approach based on the CTE lie.

(1) *Liapunov exponents:* Sensitive dependence on initial conditions is a distinguishing feature of chaotic behaviour. Initially arbitrarily close points in the phase space produce markedly different trajectories. For this reason, the slightest vagueness in the specification of the initial state renders long-term predictions impossible because two initially indistinguishable states will evolve into two distinguishable ones. Characteristically, trajectories in chaotic systems diverge exponentially and Liapunov exponents (LE) prove a good quantitative measure for the average rate of exponential divergence of two trajectories. Hence, positive LE are indicative of unpredictable behaviour. For this reason it is desirable to link the KSE to positive LE. And this is what Pessin’s theorem achieves by stating that $H\Phi$ is basically equal to the sum

of the positive LE of the system (see, for instance, Eckmann & Ruelle [1985], p. 394; Lichtenberg & Liebermann [1992], p. 304).

This is a valid argument. But it does not take us as far as we can go. Nothing has been said about the time span of the unpredictablity (will it last forever?); nothing has been said about how quickly predictions break down (after one time step? after two? after ten?), and no mention of the past history has been made (how does knowledge of the past history influence our predictive abilities?). But these questions are of great interest in the context of physics.

(2) *Algorithmic complexity:* An important account defines randomness in terms of algorithmic complexity. Roughly speaking, the algorithmic complexity (AC) of a sequence (here $\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\ldots\alpha_{l_k}^{t_k}$) is the length of the shortest computer program we have to provide in order to get a universal Turing machine to reproduce (compute) the sequence. We then define a sequence to be random if the shortest program of this sort has essentially the length of the sequence itself (that is, the program basically says 'print $\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\ldots\alpha_{l_k}^{t_k}$'). (For details, see Cover & Thomas [1991], pp. 144–82).

This notion of randomness can be connected to the KSE by invoking Brudno's theorem, which states that for almost all trajectories (i.e. sequences $\alpha_{l_1}^{t_1}\alpha_{l_2}^{t_2}\ldots\alpha_{l_k}^{t_k}$) the AC of the trajectory equals the KSE of the system (Brudno [1978]; for discussions, see Alekseev & Yakobson [1981]; Batterman & White [1996]). Hence we can say that the KSE is a measure of random behaviour in the sense of AC.

This is a very elegant way to interpret the KSE in terms of randomness. But is it really what we need? I think that this account is less attractive than it initially appears. The term 'randomness' may refer to many different things in different contexts, and it is beyond the scope of this paper to discuss the variety of options. However, in the context of dynamical systems, what we mean by 'random behaviour' is unpredictable behaviour. At the most basic level, we say that an event is random if there is no way to predict its occurrence with certainty. Likewise, a random process is one for which we are not able to predict what happens next. That is, what we have in mind when we call the behaviour of a dynamical system 'random' is our inability to predict its future behaviour, and any definition of randomness we employ in this context must somehow do justice to this intuition. But this is not the case with AC. It does not make reference to prediction and it is not clear how a connection between AC and predictability might be established since it is concerned only with the reproduction of a *previously given* sequence.

6 Product versus process randomness

I would like to conclude this paper by discussing a consequence of the above result for the notion of randomness characteristic of chaotic systems. Two basic types of randomness have been distinguished, namely process and product randomness (see Earman [1986], pp. 137–8). First, we are faced with *process randomness* (also referred to as *genesis randomness*) if we are faced with a process which operates without a hard and fast principle. A process involving genuine chance, for instance, belongs to this category. Second, the output of a process exhibits *product randomness* (also called *performance randomness*) if it is lacking a discernible pattern or if it is simply 'out of shape' in one sense or another. It is clear that product randomness is not an absolute concept. As in the case of simplicity, we have strong and pervasive intuitions, but it is difficult to cash out in an objective sense what these amount to; what we consider as patternless depends (to some extent at least) on our point of view.

These two notions of randomness do not generally coincide. On the one hand, a sequence that is random in the product sense need not necessarily be the output of a genuinely random process. So-called 'random number generators' in digital computers are a case in point. They are programs that are set up in such a way that the sequences they produce look random, but all the program performs are simple arithmetical manipulations of numbers which do not involve any stochastic element. On the other hand, process randomness is no sure guarantee for performance randomness, though it leads to strong expectation of a random product. It is in principle possible that a random process accidentally produces a highly ordered sequence. For instance, it is possible that if we flip a coin 1000 times we obtain 1000 heads.

For this reason it is interesting to note that in the case of a system with positive KSE the extensions of these two notions coincide as a consequence of the above theorem. The argument runs as follows. First, AC is commonly taken to be a notion of product randomness because it defines randomness in terms of the computational power needed to reproduce a *given* sequence. Second, my discussion of the CTE shows that it is a notion of process randomness: the focus is on the process, in that we ask at every step what the uncertainty about the next step is. Third, Brudno's theorem states that the KSE is equivalent to the AC for almost and trajectories of a system. The above theorem states that the CTE is equivalent to the KSE. Hence, AC is equivalent to the CTE for almost all trajectories. The punch line of this is that the last equivalence equates notions of process and product randomness. This means that whenever a dynamical system behaves randomly in a process sense (cashed out in terms of CTE), almost all of its trajectories exhibit product randomness (in the sense AC), and vice versa. In short, product and process randomness are extensionally equivalent.

This has a bearing on the type of randomness we find in chaotic systems. It has been claimed recently, for instance, that chaotic systems exhibit only product but not process randomness: 'If there is to be randomness in chaotic models, it must be randomness in the product sense—since, by hypothesis, we are there dealing with models with thoroughly deterministic dynamics (the "processes" are entirely non-random)' (Smith [1998], p. 149). However, if we grant that K-systems are chaotic,[6] then doubt is cast on this claim since K-systems exhibit both product and process randomness.

Proponents of the argument in question might now counter that the underlying dynamics are thoroughly deterministic and for this reason there cannot be any process randomness at the very 'basic level'. True, at the level of 'mathematical' trajectories and exact initial conditions there is no randomness. But this reply is beside the point: chaos and randomness only become an issue in dynamical systems once the dynamics are discussed at the coarse-grained level; as long as we assume that unlimited precision is available, there is no unpredictability or any other symptom of chaos. But once we go to the coarse-grained level, the system exhibits both product and process randomness.

Appendix: Proofs of Theorems 1 and 2

In order to prove the two main theorems, five lemmas are needed. The proof of Lemmas 1 and 3 can be found in Arnold & Avez ([1968], p. 38); the other proofs are trivial.

Lemma 1: $H(\alpha \vee \beta) = H(\alpha) + H(\beta/\alpha)$
Lemma 2: $H(\alpha) = H(\Phi\alpha)$
Lemma 3: $\Phi(\alpha \vee \beta) = \Phi\alpha \vee \Phi\beta$
Lemma 4: $H(\alpha \vee \beta) = H(\beta \vee \alpha)$
Lemma 5: $\vee$ is associative: $\alpha \vee \beta \vee \gamma = (\alpha \vee \beta) \vee \gamma = \alpha \vee (\beta \vee \gamma)$

Proof of Theorem 1: For the case of an automorphism generated by a mapping we have $\Phi_{t_i \to t_j}(A) = \Phi^{j-i}(A)$ (see above). Then (17) becomes:

$$\bar{H}_k(\alpha) = -\sum_{l_1,\ldots,l_k=1}^{n} \mu(\alpha_{l_k} \cap \cdots \cap \Phi^{k-1}\alpha_{l_1}) \sum_{i=1}^{n} z\left[\frac{\mu(\alpha_i \cap \Phi\alpha_{l_k} \cap \cdots \cap \Phi^k \alpha_{l_1})}{\mu(\alpha_{l_k} \cap \cdots \cap \Phi^{k-1}\alpha_{l_1})}\right] \tag{29}$$

Using the fact that Φ is area preserving we get $\mu(\alpha_{l_k} \cap \cdots \cap \Phi^{k-1}\alpha_{l_1}) = \mu(\Phi\alpha_{l_k} \cap \cdots \cap \Phi^k \alpha_{l_1})$. Plugging this into Eqn (29) and taking the associativity of set intersection into account we obtain:

$$\bar{H}_k(\alpha) = -\sum_{l_1,\ldots,l_k=1}^{n} \mu(\Phi\alpha_{l_k} \cap \cdots \cap \Phi^k\alpha_{l_1}) \sum_{i=1}^{n} z\left[\frac{\mu(\alpha_i \cap \{\Phi\alpha_{l_k} \cap \cdots \cap \Phi^k \alpha_{l_1}\})}{\mu(\Phi\alpha_{l_k} \cap \cdots \cap \Phi^k\alpha_{l_1})}\right] \tag{30}$$

[6] I take it that being a K-system is sufficient for chaos but it is clearly not necessary. KAM-type systems, for instance, exhibit chaotic beaviour but they are not K-systems. Furthermore, dissipative systems, to which the notion of being a K-system does not apply, can be chaotic.

Now note that what the first sum effectively does is sum over all elements of a partition consisting of all intersections $\Phi\alpha_{l_k} \cap \cdots \cap \Phi^k \alpha_{l_1}$. This partition, however, is just $\Phi\alpha \vee \cdots \vee \Phi^k\alpha$. Furthermore, compare Eqn (30) with the definition of the conditional entropy in Eqn (22). We then obtain: $\bar{H}_k(\alpha) = H(\alpha/\Phi\alpha \vee \cdots \vee \Phi^k\alpha)$.

Proof of Theorem 2 by weak induction on *k*:

Base case:

$H(\alpha \vee \Phi\alpha) = H(\alpha) + H(\alpha/\Phi\alpha)$. *Proof*: $H(\alpha \vee \Phi\alpha) = H(\Phi\alpha \vee \alpha)$, by Lemma 4, and $H(\Phi\alpha \vee \alpha) = H(\Phi\alpha) + H(\alpha/\Phi\alpha)$ by Lemma 1. Now use Lemma 2 and get $H(\Phi\alpha) + H(\alpha/\Phi\alpha) = H(\alpha) + H(\alpha/\Phi\alpha)$.

Inductive step:

$H(\alpha \vee \Phi\alpha \vee \cdots \vee \Phi^{k+1}\alpha) = H(\alpha) + \cdots + H(\alpha/\Phi\alpha \vee \cdots \vee \Phi^{k+1}\alpha)$. *Proof*: Consider $H(\alpha \vee \Phi\alpha \vee \cdots \vee \Phi^{k+1}\alpha)$. With Lemmas 5 and 4 this is $H([\Phi\alpha \vee \cdots \vee \Phi^{k+1}\alpha] \vee \alpha)$, and now applying Lemma 1 yields $H(\Phi\alpha \vee \cdots \vee \Phi^{k+1}\alpha) + H(\alpha/[\Phi\alpha \vee \cdots \vee \Phi^{k+1}\alpha])$. Lemmas 2 and 3 together with the fact that Φ is measure-preserving give: $H(\alpha \vee \cdots \vee \Phi^k\alpha) + H(\alpha/[\Phi\alpha \vee \cdots \vee \Phi^{k+1}\alpha])$. With the induction hypothesis, this is $H(\alpha \vee \Phi\alpha \vee \cdots \vee \Phi^{k+1}\alpha) = H(\alpha) + \cdots + H(\alpha/[\Phi\alpha \vee \cdots \vee \Phi^{k+1}\alpha])$.

Acknowledgements

Jossi Berkovitz, Robert Bishop, Nancy Cartwright, Jean-Michel Delhotel, Carl Hoefer, Fred Kronz, Samuel Kutter, Chuang Liu, Michael Redhead, Orly Shenker and Max Steuer have read earlier drafts of the paper and their comments were of great help to me.

Department of Philosophy, Logic and Scientific Method
London School of Economics
Houghton Street
London WC2A 2AE
UK
Email: r.p.frigg@lse.ac.uk

References

Alekseev, V. M. & Yakobson, M. V. [1981]: ‘Symbolic Dynamics and Hyperbolic Dynamical Systems’, *Physics Reports*, **75**, pp. 287–325.

Arnold, V. I. & Avez, A. [1968]: *Ergodic Problems of Classical Mechanics*, New York, NY & Amsterdam: W. A. Benjamin.

Batterman, R. [1993]: ‘Defining Chaos’, *Philosophy of Science*, **60**, pp. 43–66.

Batterman, R. & White, H. [1996]: 'Chaos and Algorithmic Complexity', *Foundations of Physics*, **26**, pp. 307–336.

Belot, G. & Earman, J. [1997]: 'Chaos out of Order: Quantum Mechanics, the Correspondence Principle and Chaos', *Studies in the History and Philosophy of Modern Physics*, **28**, pp. 147–82.

Billingsley, P. [1965]: *Ergodic Theory and Information*, New York, NY: Wiley.

Brudno, A. A. [1978]: 'The Complexity of the Trajectory of a Dynamical System', *Russian Mathematical Surveys*, **33**, pp. 197–8.

Carroll, L. [1998]: *Alice's Adventures in Wonderland and Through the Looking-Glass*, London: Penguin.

Cornfeld, I. P., Fomin, S. V. & Sinai, Y. G. [1982]: *Ergodic Theory*, Berlin & New York, NY: Springer.

Cornfeld, I. P. & Sinai, Y. G. [1980]: 'Entropy Theory of Dynamical Systems', in Y. G. Sinai (*ed.*), 1980, *Dynamical Systems II. Ergodic Theory with Applications to Dynamical Systems and Statistical Mechanics*, Berlin & New York, NY: Springer, pp. 36–58.

Cover, T. M. & Thomas, J. M. [1991]: *Elements of Information Theory*, New York, NY & Chichester: Wiley.

Earman, J. [1986]: *A Primer on Determinism*, Dordrecht: Reidel.

Eckmann, J.-P. & Ruelle, D. [1985]: 'Ergodic Theory of Chaos and Strange Attractors', *Review of Modern Physics*, **57**, Part I, pp. 617–656.

Keller, G. [1998]: *Equilibrium States in Ergodic Theory*, Cambridge: Cambridge University Press.

Lichtenberg, A. J. & Liebermann, M. A. [1992]: *Regular and Chaotic Dynamics*, 2^{nd} edn, Berlin & New York, NY: Springer.

Mañé, R. [1983]: *Ergodic Theory and Differentiable Dynamics*, Berlin & New York: Springer.

Nadkarni, M. G. [1998]: *Basic Ergodic Theory*, Basel: Birkhäuser.

Ott, E. [1993]: *Chaos in Dynamical Systems*, Cambridge: Cambridge University Press.

Parry, W. [1981]: *Topics in Ergodic Theory*, Cambridge: Cambridge University Press.

Petersen, Karl [1983]: *Ergodic Theory*. London and New York, NY: Cambridge UP.

Reichl, L. E. [1992]: *The Transition to Chaos in Conservative Classical Systems: Quantum Manifestations*, Berlin & New York, NY: Springer.

Rudolph, D. J. [1990]: *Fundamentals of Measurable Dynamics: Ergodic Theory on Lebesgue Spaces*, Oxford: Clarendon Press.

Schuster, H. G. [1988]: *Deterministic Chaos: An Introduction*, 2^{nd} edn, Weinheim: Physik Verlag.

Shannon, C. E. & Weaver, W. [1949]: *The Mathematical Theory of Communication*, Urbana, Chicago, IL & London: University of Illinois Press.

Smith, P. [1998]: *Explaining Chaos*, Cambridge: Cambridge University Press.

Tabor, M. [1989]: *Chaos and Integrability in Nonlinear Dynamics*, New York, NY: John Wiley & Sons.