Advances in Complex Systems

AN INFORMATION-THEORETIC PRIMER ON COMPLEXITY, SELF-ORGANISATION AND EMERGENCE

MIKHAIL PROKOPENKO

Information and Communication Technologies Centre,
Commonwealth Scientific and Industrial Research Organisation,
Locked bag 17, North Ryde, NSW 1670, Australia
mikhail.prokopenko@csiro.au

FABIO BOSCHETTI

Marine and Atmospheric Research,
Commonwealth Scientific and Industrial Research Organisation,
Underwood Avenue, Floreat, WA, Australia
fabio.boschetti@csiro.au

ALEX J. RYAN

Defence Science and Technology Organisation,
West Avenue, Edinburgh, SA, Australia
alex.ryan@dsto.defence.gov.au

Received (received date)
Revised (revised date)

Complex Systems Science aims to understand concepts like complexity, self-organization, emergence and adaptation, among others. The inherent fuzziness in complex systems definitions is complicated by the unclear relation among these central processes: does self-organisation emerge or does it set the preconditions for emergence? Does complexity arise by adaptation or is complexity necessary for adaptation to arise? The inevitable consequence of the current impasse is miscommunication among scientists within and across disciplines. We propose a set of concepts, together with their information-theoretic interpretations, which can be used as a dictionary of Complex Systems Science discourse. Our hope is that the suggested unifying information-theoretic framework may facilitate consistent communications among practitioners, and provide new insights into the field.

Keywords: Complexity; Information Theory; self-organisation; emergence; predictive information; excess entropy; entropy rate; assortativeness; predictive efficiency; adaptation; self-referentiality; downward causation.

1. Introduction

Complex Systems Science studies general phenomena of systems comprised of many simple elements interacting in a non-trivial fashion. Currently, fuzzy quantifiers like 'many' and 'non-trivial' are inevitable. 'Many' implies a number large enough so that no individual component/feature predominates the dynamics of the system,

but not so large that features are completely irrelevant. Interactions need to be 'non-trivial' so that the degrees of freedom are suitably reduced, but not constraining to the point that the arising structure possesses no further degree of freedom. Crudely put, systems with a huge number of components interacting trivially are explained by statistical mechanics, and systems with precisely defined and constrained interactions are the concern of fields like chemistry and engineering. In so far as the domain of Complex Systems Science overlaps these fields, it contributes insights when the classical assumptions are violated.

It is unsurprising that a similar vagueness afflicts the discipline itself, which notably lacks a common formal framework for analysis. There are a number of reasons for this. Because Complex Systems Science is broader than physics, biology, sociology, ecology, or economics, its foundations cannot be reduced to a single discipline. Furthermore, systems which lie in the gap between the 'very large' and the 'fairly small' cannot be easily modelled with traditional mathematical techniques.

Initially setting aside the requirement for formal definitions, we can summarise our general understating of complex systems dynamics as follows:

(1) complex systems are 'open', and receive a regular supply of energy, information, and/or matter from the environment;
(2) a large, but not too large, ensemble of individual components interact in a non-trivial fashion;
(3) the non-trivial interactions result in internal constraints, leading to symmetry breaking in the behaviour of the individual components, from which *coordinated* global behaviour arises;
(4) the system is now more organised than it was before; since no central director nor any explicit instruction template was followed, we say that the system has *'self-organised'*;
(5) this coordination can express itself as *patterns* detectable by an external observer or as structures that convey new properties to the systems itself. New behaviours *'emerge'* from the system;
(6) coordination and emergent properties may arise from specific response to environmental pressure, in which case we can say the system displays *adaptation*;
(7) when adaptation occurs across generations at a population level we say that the system *evolved*;
(8) coordinated emergent properties give rise to larger *scale* effects. These interdependent sets of components with emergent properties can be observed as coherent entities at lower *resolution* than is needed to observe the components. The system can be identified as a novel unit of its own and can interact with other systems/processes expressing themselves at the same scale. This becomes a building block for new iterations and the cycle can repeat from 1. above, now at a larger scale.

The process outlined above is not too contentious, but does not address 'how' and 'why' each step occurs. Consequently, we can observe the process but we can not

understand it, modify it or engineer for it. This also prevents us from understanding what complexity is and how it should be monitored and measured; this equally applies to self-organisation, emergence, evolution and adaptation.

Even worse than the fuzziness and absence of deep understanding already described, is when the above terms are used interchangeably in the literature. The danger of not making clear distinctions in Complex Systems Science is incoherence. To have any hope of coherent communication, it is necessary to unravel the knot of assumptions and circular definitions that are often left unexamined.

Here we suggest a set of working definitions for the above concepts, essentially a dictionary for Complex Systems Science discourse. Our purpose is not to be prescriptive, but to propose a baseline for shared agreement, to facilitate communication between scientists and practitioners in the field. We would like to prevent the situation in which a scientist talks of emergence and this is understood as self-organisation.

For this purpose we chose an information-theoretic framework. There are a number of reasons for this choice:

- a considerable body of work in Complex Systems Science has been cast into Information Theory, as pioneered by the Santa Fe Institute, and we borrow heavily from this tradition;
- it provides a well developed theoretical bases for our discussion;
- it provides definitions which can be formulated mathematically;
- it provides computational tools readily available; a number of measures can be actually computed, albeit in a limited number of cases.

Nevertheless, we believe that the concepts should also be accessible to disciplines which often operate beyond the application of such a strong mathematical and computational framework, like biology, sociology and ecology. Consequently, for each concept we provide a 'plain English' interpretation, which hopefully will enable communication across fields.

2. An information-theoretical approach

Information Theory was originally developed by Shannon [65] for reliable transmission of information from a source X to a receiver Y over noisy communication channels, and includes fundamental concepts of source coding and channel coding.

Source coding quantifies the average number of bits needed to represent the result of an uncertain event – intuitively, it measures one's freedom of choice in selecting a message which captures information in the source X faithfully, i.e. without information loss. This quantity is known as *(information) entropy*. In the simplest cases, the amount of information can be measured by the logarithm (base 2) of the number of available choices. The entropy is a precise measure of the amount of freedom of choice (or the degree of randomness) contained in the process – a process with many possible states has high entropy. Formally, given a probability

distribution P over a (discrete) random variable X, the entropy is defined by

$$H(X) = -\sum_X p(x) \log p(x) \ , \tag{1}$$

and is the only measure satisfying three required properties:

- changing the value of one of the probabilities by a small amount changes the entropy by a small amount;
- if all the choices are equally likely, then entropy is maximal;
- entropy is independent of how the process is divided into parts.

The joint entropy of two (discrete) random variables X and Y is defined as the entropy of the joint distribution of X and Y:

$$H(X,Y) = -\sum_{X,Y} p(x,y) \log p(x,y) \ , \tag{2}$$

Mutual information $I(X;Y)$ is defined as the amount of information (in bits) that the received signal Y contains about the transmitted signal X, on average:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y|X) \ , \tag{3}$$

where $H(Y|X)$ is the conditional entropy of Y given X, also called the equivocation of Y about X. If X and Y are independent, the joint entropy is simply the sum of their individual entropies, and the mutual information is zero. Informally, we can state that

mutual information = receiver's diversity − equivocation of receiver about source (4)

Equivocation of Y about X may also be interpreted as non-assortativeness between Y and X: the degree of having no reciprocity in either a positive or negative way. This interpretation is particularly useful in dealing with assortative, disassortative, and non-assortative networks.

Channel coding establishes that reliable communication is possible over noisy channels if the rate of communication is below a certain threshold called the channel capacity. Channel capacity is defined as the maximum mutual information for the channel over all possible distributions of the transmitted signal X.

In the remainder of this work, we intend to point out how different concepts in Complex Systems Science can be interpreted via simple information-theoretic relationships. In particular, when suitable information channels are identified, the rest is often a matter of computation – the computation of "diversity" and "equivocation". The choice of channels is typically a task for modelers, while in biological systems the "embodied" channels are shaped by interactions with the environment.

There are other mathematical approaches, such as non-linear time series analysis, Chaos Theory, etc., that also provide insights to the concepts used by Complex Systems Science. We note that these approaches are outside the scope of this paper, as our intention is to provide a candidate unifying framework rather than a competing methodology. It is possible that Information Theory has not been widely

used in applied studies of complex systems because of the lack of clarity. We are proposing here to clarify the applicability and exemplify how different information channels can be identified and used.

3. Complexity

It is an intuitive notion that certain processes and systems are harder to describe than others. Complexity tries to capture this difficulty in terms of the amount of information needed for the description, the time it takes to carry out the description, the size of the system, the number of components in the system, the number of conflicting constraints, the number of dimensions needed to embed the system dynamics and related ideas. A large number of definitions have been proposed in the literature and since a review is beyond the scope of this work we refer the interested reader to http://bruce.edmonds.name/.

Here we propose to adopt as definition of complexity the amount of information needed to describe a process, a system or an object. This definition is computable (at least in one of its forms), is observer-independent (once resolution is defined), it applies to both data and models [12] and provides a framework within which self-organisation and emergence can also be consistently defined.

3.1. *Concept*

Algorithmic Complexity The original formulation can be traced back to Solomonoff, Kolmogorov and Chaitin, who developed independently what is today known as Kolmogorov-Chaitin or algorithmic complexity [16]. Given an entity (this could be a data set or an image, but the idea can be extended to material objects and also to life forms) the algorithmic complexity is defined as the length (in bits of information) of the shortest program (computer model) which can describe the entity. According to this definition a simple periodic object (a sine function for example) is not complex, since we can store a sample of the period and write a program which repeatedly outputs it, thereby reconstructing the original data set with a very small program. At the opposite end of the spectrum, an object with no internal structure cannot be described in any meaningful way but by storing every feature, since we cannot rely on any shared structure for a shorter description. It follows that a random object has maximum complexity, since the shortest program able to reconstruct it needs to store the object itself[a].

A nice property of this definition is that it does not depend on what language we use to write the program[b].

Statistical Complexity A clear disadvantage of the algorithmic complexity is that it can not be computed exactly but only approximated from above[c]. A criti-

[a]This follows from the most widely used definition of randomness, as structure which can not be compressed in any meaningful way.
[b]It can be shown that descriptions using different languages differ by constants.
[c]See the Chaitin theorem [15].

cism often made to this definition is that there are problems for which associating randomness to maximum complexity seems counter-intuitive. Imagine you throw a cup of rice to the floor and you want to describe the spatial distribution of the grains. In most cases you do not need to concern with storing the position of each individual grain; the realisation that the distribution is structure-less and that predicting the exact position of a specific grain is impossible is probably all you need to know. And this piece of information is very simple (and short) to store. There are applications for which our intuition suggests that both strictly periodic and totally random sequences should share low complexity.

One definition which addresses this concern is the statistical complexity [26], which attempts to measure the size of the minimum program able to reproduce the statistically significant features of the entity under analysis. In the rice pattern mentioned above, there is no statistical difference in the probability of finding a grain at different positions and the resulting statistical complexity is zero. Along with our previous non mathematical analogy we could say that for many purposes the actions of a mentally disabled patient, taking apparently random actions at any time can be described very briefly by just noticing the lack of apparent intentionality or purpose in the patient's behaviour.

Apart from implementation details, the conceptual difference between algorithmic and statistical complexity lies in how randomness is treated. Essentially, the algorithmic complexity implies a deterministic description of an object (it defines the information content of an individual sequence), while the statistical complexity implies a statistical description (it refers to an ensemble of sequences generated by a certain source) [37, 11]. As suggested by Boffetta et al. [11], which of these approaches is more suitable is problem-specific. In the previous analogy, the details of the behaviour of a mentally-disabled patient maybe of crucial importance to a psychiatrist, who may thus prefer an algorithmic approach, but less relevant to a more superficial observer, for whom a statistical description would suffice.

Excess entropy and predictive information As pointed out by Bialek et al. [8], our intuitive notion of complexity corresponds to statements about the underlying process, and not directly to Kolmogorov complexity. A dynamic process with an unpredictable and random output (large algorithmic complexity) may be as trivial as the dynamics producing predictable constant outputs (small algorithmic complexity) – while "really complex processes lie somewhere in between". Interestingly, however, the two extreme cases share one feature: the entropy of the output strings "either is a fixed constant or grows exactly linearly with the length of the strings", and corrections to the asymptotic behaviour do not grow with the size of the data set. Grassberger [37] identified the slow approach of the entropy to its extensive limit as a sign of complexity. Thus, subextensive components – which grow with time less rapidly than a linear function – are of special interest. Bialek et al. [8] observe that the subextensive components of entropy identified by Grassberger determine precisely the information available for making predictions – e.g. the complexity in a time series can be related to the components which are "useful"

or "meaningful" for prediction. We shall refer to this as *predictive information*. Revisiting the two extreme cases, they note that "it only takes a fixed number of bits to code either a call to a random number generator or to a constant function" – in other words, a model description *relevant to prediction* is compact in both cases.

The predictive information is also referred to as excess entropy [25, 23], stored information [66], effective measure complexity [37, 51, 33], complexity [50, 2], and has a number of interpretations.

3.2. *Information-theoretic interpretation*

3.2.1. *Predictive information*

In order to estimate the relevance to prediction, two distributions over a stream of data x are considered: a prior probability distribution for the futures, $\mathrm{P}(x_{future})$, and a more tightly concentrated distribution of futures conditional on the past data, $\mathrm{P}(x_{future}|x_{past})$, and define their average ratio

$$I_{pred}(T,T') = \left\langle log_2 \frac{\mathrm{P}(x_{future}|x_{past})}{\mathrm{P}(x_{future})} \right\rangle , \tag{5}$$

where $\langle \cdots \rangle$ denotes an average over the joint distribution of the past and the future, $\mathrm{P}(x_{future}|x_{past})$, T is the length of the observed data stream in the past, and T' is the length of the data stream that will be observed in the future. This average predictive information captures the reduction of entropy (in Shannon's sense) by quantifying the information that the past provides about the future:

$I_{pred}(T,T') = H(T') - H(T'|T)$ or informally,

$$\begin{aligned} \text{predictive information} = \text{total uncertainty about the future} - \\ \text{uncertainty about the future, given the past} \end{aligned} \tag{6}$$

The predictive information is always positive and grows with time less rapidly than a linear function, being subextensive. It provides a universal answer to the question of how much is there to learn about the underlying pattern in a data stream: $I_{pred}(T,T')$ may either stay finite, or grow infinitely with time. If it stays finite, this means that no matter how long we observe we gain only a finite amount of information about the future: e.g. it is possible to completely predict dynamics of periodic regular processes after their period is identified. For some irregular processes the best predictions may depend only on the immediate past (e.g. a Markov process, or in general, a system far away from phase transitions and/or symmetry breaking) – and in these cases $I_{pred}(T,T')$ is also small and is bound by the logarithm of the number of accessible states: the systems with more states and longer memories have larger values of predictive information [8]. If $I_{pred}(T,T')$ diverges and optimal predictions are influenced by events in the arbitrarily distant past, then the rate of growth may be slow (logarithmic) or fast (sublinear power). If the data allows us to learn a model with a finite number of parameters or a set of underlying rules describable by a finite number of parameters, then $I_{pred}(T,T')$ grows

logarithmically with a coefficient that counts the dimensionality of the model space (i.e. the number of parameters). Sublinear power-law growth may be associated with infinite parameter (or nonparametric) models such as continuous functions with some regularization (e.g. smoothness constraints) [9].

3.2.2. *Statistical complexity*

The statistical complexity is calculated by reconstructing a minimal model, which contains the collection of all situations which share a similar specific probabilistic future – the causal states – and measuring the entropy of the probability distribution of the states. The description of an algorithm which achieves such reconstruction and calculates the statistical complexity for 1D time series can be found in [63] and for 2D time series in [62].

In general, the predictive information is related to the statistical complexity $I_{pred}(T, T') \leq C_\mu$, which is the minimum average amount of memory needed to statistically reproduce the configuration ensemble to which the sequence belongs [70] – both predictive information and statistical complexity are measured in bits.

3.2.3. *Excess entropy*

Excess entropy is defined as a measure of the total apparent memory or structure in a source [24]:

$$E = \sum_{L=1}^{\infty} (h_\mu(L) - h_\mu), \tag{7}$$

where $h_\mu(L) = H(L) - H(L-1), L \geq 1$, for the entropy $H(L)$ of length-L sequences or blocks, and $h_\mu = \lim_{h\to\infty} \frac{H(L)}{L}$ is the source entropy rate – also known as per-symbol entropy, the thermodynamic entropy density, Kolmogorov-Sinai entropy, metric entropy, etc. The length-L entropy rate $h_\mu(L)$ is the average uncertainty about the L^{th} symbol, provided the $(L-1)$ previous ones are given [11]. As noted by Crutchfield and Feldman [24], the length-L approximation $h_\mu(L)$ typically over-estimates the entropy rate h_μ at finite L, and each difference $\lfloor h_\mu(L) - h_\mu \rfloor$ is the difference between the entropy rate conditioned on L measurements and the entropy rate conditioned on an infinite number of measurements – it estimates the information-carrying capacity in the L-blocks that is not actually random, but is due instead to correlations, and can be interpreted as the local (i.e. L-dependent) predictability [30]. The total sum of these local over-estimates is the excess entropy or intrinsic redundancy in the source. Thus, the excess entropy measures the amount of apparent randomness at small L values that is "explained away" by considering correlations over larger and larger blocks. Importantly, Crutchfield and Feldman [24] demonstrated that the excess entropy E can also be seen as either:

(1) the mutual information between the source's past and the future – exactly the predictive information $I_{pred}(T, T')$, if T and T' are semi-infinite, or

(2) the subextensive part of entropy $H(L) = E + h_\mu L$, as $L \to \infty$.

It was also shown that only the first interpretation holds in 2-dimensional systems [35].

Revisiting representation (6), we may point out that the total uncertainty $H(T')$ can be thought of as structural diversity of the underlying process. Similarly the conditional uncertainty $H(T'|T)$ can be related to structural non-conformity or equivocation within the process – a degree of non-assortativeness between the past and the future, or between components of the process in general. This analogy creates an alternative intuitive representation, which will be analysed in section 4:

$$\text{predictive information} = \text{excess entropy} = \text{diversity} - \text{non-assortativeness} \quad (8)$$

3.3. *Example – Thue-Morse process*

Predictive information grows logarithmically for infinite-memory Thue-Morse sequences $\sigma^k(s)$ which contain two units 0 and 1, and can be obtained by the substitution rules $\sigma^k(0) = 01$ and $\sigma^k(1) = 10$ (e.g. $\sigma^2(1) = 1001$, etc.). Such a process needs an infinite amount of memory to maintain its aperiodicity [24], and hence, its past provides an ever-increasing predictive information about its future.

An even faster rate of growth is also possible – typically, this happens in problems where predictability over long scales is "governed by a progressively more detailed description" as more data are observed [8]. This essentially produces an increase in the number of causal states, used by Crutchfield in defining statistical complexity.

3.4. *Example – periodic vs random processes*

The source entropy rate h_μ captures the irreducible randomness produced by a source after all correlations are taken into account [24]:

- $h_\mu = 0$ for periodic processes and even for deterministic processes with infinite-memory (e.g. Thue-Morse process) which do not have an internal source of randomness, and
- $h_\mu > 0$ for irreducibly unpredictable processes, e.g. independent identically distributed (IID) processes which have no temporal memory and no complexity, as well as Markov processes (both deterministic and nondeterministic), and infinitary processes (e.g. positive-entropy-rate variations on the Thue-Morse process).

The excess entropy, or predictive information, is a better measure of complexity as it increases with the amount of structure or memory within a process:

- E is finite for both periodic processes and random (e.g. it is zero for an IID process) – its value can be used as a relative measure: a larger period results in higher E, as a longer past needs to be observed before we can estimate the finite predictive information;

- E diverges logarithmically for complex processes due to an infinite memory (e.g. Thue-Morse process) – again, its value can be used as a relative measure estimating a number of parameters or rules in the underlying model;
- E exhibits a sublinear power law divergence for complex processes due to a nonparametric model of the underlying process (e.g. a continuous function with smoothness constraints) – here, the relative measure is the number of different parameter-estimation scales growing in proportion to the number of taken samples (e.g. the number of bins used in a histogram approximating the distribution of a random variable), i.e. a progressively more detailed description is required.

3.5. *Summary*

While entropy rate is a good identifier of intrinsic randomness, it suffers from the same drawbacks as the Kolmogorov-Chaitin (KC) complexity, to which it is strongly related. To reiterate, the KC complexity of an object is the length of the minimal Universal Turing Machine (UTM) program needed to reproduce it. The entropy rate h_μ is equal to the average length (per variable) of the minimal program that, when run, will cause an UTM to produce a typical configuration and then halt [34, 21, 49].

The relationships $I_{pred}(T, T') = E$ and $E \leq C_\mu$, suggest a very intuitive interpretation:

$$\text{predictive information} = \text{richness of structure} \leq \text{statistical complexity} = \text{memory for optimal predictions} \quad (9)$$

where the latter is defined as the entropy of causal states – all histories that have the same conditional distribution of futures [26, 61]. The causal states provide an optimal description of a system's dynamics in the sense that these states make as good a prediction as the histories themselves. The inequality in (9) means that the memory needed to perform an optimal prediction of the future configurations cannot be lower than the mutual information between the past and future themselves [34]: this relationship reflects the fact that the causal states are a reconstruction of the hidden, effective states of the process.

Specifying how the memory within a process is organized cannot be done within the framework of Information Theory [24], and a more structural approach based on the Theory of Computation must be used – this leads (via causal states) to ε-machines and statistical complexity C_μ.

4. Edge-of-chaos

According to the analysis presented in the preceding section, statistical complexity is small at both extremes (complete order and complete randomness), and is maximal in the region somewhere between the extremes. Moreover, in some "intermediate" cases, the complexity is infinite, and may be divergent at different rates. A few natural questions then arise:

- when does the complexity attains the maximum;
- how can we precisely identify this region;
- what happens to information dynamics within or near this region; and
- how such dynamics are related to self-organisation, adaptation, evolution, and emergence in general.

4.1. *Concept*

Cellular automata (CA) are discrete spatially-extended dynamical systems that have been used as models of many computational, physical and biological processes [52], and we shall use this abstraction in illustrating the edge-of-chaos phenomenon[d]. Langton [48] examined behaviour of CA in terms of the parameter λ – the fraction of rules with a given property in the rule table (e.g. the fraction of nonzero transitions). Varying the parameter within its range between ordered and chaotic extremes, he identified, for intermediate values of λ, an increase in the mutual information $I(A;B)$ between a cell and itself at the next time-step. An intriguing feature is that the average mutual information has a distinct peak at the transition point – an indication of a phase transition from "order" to "chaos" in CA. This study introduced the edge-of-chaos – the region where the CA behaviour shifts from the ordered regimes towards chaotic regimes, effectively approaching random dynamics.

4.2. *Information-theoretic interpretation*

A rule-space of 1-dimensional CA was characterised with the Shannon entropy of rules' frequency distribution [73]. More precisely, given a rule-table (the rules that define a CA), the input-entropy at time step t is defined as

$$S^t = -\sum_{i=1}^{m} \frac{Q_i^t}{n} \log \frac{Q_i^t}{n} \tag{10}$$

where m is the number of rules, n is the number of cells (system size), and Q_i^t is the number of times the rule i was used at time t across the CA. The input-entropy settles to fairly low values for ordered dynamics, but fluctuates irregularly within a narrow high band for chaotic dynamics. For the complex CA, order and chaos may predominate at different times causing the entropy to vary. A measure of the variability of the input-entropy curve is its variance or standard deviation, calculated over time. Wuensche [73] has convincingly demonstrated that only complex dynamics exhibits high variance of input-entropy, leading to automatic classification of the rule-space. Importantly, the peak of input-entropy variance points to a phase transition as well, indicating the edge-of-chaos.

[d]We would like to point out that the edge-of-chaos hypothesis is far from being accepted unanimously. The edge of chaos concept is also distinct from low dimensional chaos in dynamical systems, characterised by non-periodicity with few degrees of freedom. Nevertheless some connections have been made – see subsection 4.3 for a brief description.

Similarly, Wolfram [72] classified cellular automata rules qualitatively – according to their asymptotic behavior: class I (homogeneity); class II (periodicity); class III (chaos); class IV (complexity). The first class consists of CA that, after a finite number of time steps, produce a unique, homogeneous state (analogous to "fixed points" in phase space). From almost all initial states – such behaviour completely destroys any information on the initial state, i.e. complete prediction is trivial and complexity is low. The second class contains automata which generate a set of either stable or periodic structures (typically having small periods – analogous to "limit cycles" in phase space) – each region of the final state depends only on a finite region of the initial state, i.e. information contained within a small region in the initial state thus suffices to predict the form of a region in the final state. The third class includes infinite CA producing aperiodic ("chaotic") spatiotemporal patterns from almost all possible initial states – the effects of changes in the initial state almost always propagate forever at a finite speed, and a particular region depends on a region of the initial state of an ever-increasing size (analogous to "strange attractors" in phase space). While any prediction of the "final" state requires complete knowledge of the initial state, the regions are indistinguishable statistically as they possess no structure, and therefore the statistical complexity is low. The fourth class deals with automata that generate patterns continuously changing over an unbounded transient.

4.3. *Example – universal computation*

The fourth class CA existing at the edge-of-chaos were also shown to be capable of universal computation [72]. They support three basic operations (information storage, transmission, and modification) through static, propagating and interacting structures (blinkers, gliders, collisions). Importantly, the patterns produced along the transient are different in terms of generated structure, and in fact, their structural *variability* is highest among all four classes – i.e. the predictive information and the complexity of the class IV automata are highest. Casti [13] made an analogy between the complex (class IV) automata and quasi-periodic orbits in phase space, while pursuing deeper interconnections between CA, dynamical systems, Turing Machines, and formal logic systems – in particular, the complex automata producing edge-of-chaos dynamics were related to formal systems with undecidable statements (Gödel's Theorem).

4.4. *Example – graph connectivity*

Graph connectivity can be analysed in terms of the size of the largest connected subgraph (LCS) and its standard deviation obtained across an ensemble of graphs, as suggested by Random Graph Theory [32]. In particular, critical changes occur in connectivity of a directed graph as the number of edges increases: the size of the LCS rapidly increases as well and fills most of the graph, while the variance in the size of the LCS reaches a maximum at some critical point before decreasing. In

other words, variability within the ensemble of graphs grows as graphs become more and more are different in terms of their structure – this is analogous to different patterns in complex CA.

An information-theoretic representation can subsume this graph-theoretic model. For example, a feasible average measure of a complex network's heterogeneity is given by the entropy of a network defined through the link distribution. The latter can be defined via the simple degree distribution, the probability P_k of having a node with k links, or via the remaining degree distribution q_k: "the number of edges leaving the vertex other than the one we arrived along" [68]. The remaining degree distribution is useful in analysing how assortative, disassortative or non-assortative is the network. Assortative mixing (AM) in Graph Theory is the extent to which high-degree nodes connect to other high degree nodes [53]. In disassortative mixing (DM), high-degree nodes are connected to low-degree ones. Both AM and DM networks are contrasted with non-assortative mixing (NM), where one cannot establish any preferential connection between nodes.

Importantly, the conditional entropy $H(q|q')$ may estimate spurious correlations in the network created by connecting the vertices with dissimilar degrees – this noise affects the overall diversity or the average uncertainty of the network, but does not contribute to the amount of information (correlation) within it. Using the joint probability of connected pairs $q_{k,k'}$, one may calculate the amount of correlation between vertices in the graph via the mutual information measure, the information transfer, as

$$I(q) = H(q) - H(q|q') = \sum_{k=1}^{m} \sum_{k'=1}^{m} q_{k,k'} \log \frac{q_{k,k'}}{q_k q_{k'}}. \tag{11}$$

Informally,

$$\begin{aligned} &\text{transfer within the network} = \\ &\quad \text{diversity in the network} - \text{assortative noise in the network structure} \end{aligned} \tag{12}$$

This interpretation is analogous to the one suggested by the Equations (4) and (8), assuming that assortative noise is the non-assortative extent to which the preferential (either AM or DM) connections are obscured. In general, the mutual information $I(q)$ is a better, more generic measure of dependence: correlation functions, like the variance in the size of the LCS, "measure linear relations whereas mutual information measures the general dependence and is thus a less biased statistic" [68]. At the edge of chaos, the information transfer within the network attains its maximum. This case, however, is not easily achievable as observed by Sole and Valverde. In fact, the cases where entropy $H(q)$ and noise $H(q|q')$ are approximately equal, are most typical, suggesting that some intrinsic constraints dominate the search-space of possible network configurations.

4.5. *Summary*

The main Edge-of-Chaos hypothesis asserts that when biological systems must perform complex computation for survival, the process of evolution under natural selection tends to select systems near a phase transition between ordered and chaotic behavior [52] – i.e. the complex dynamic regime acts as an evolutionary attractor in the state space. The analysis summarised in this section emphasizes the role of structural variability (equivalent to predictive information) in complex systems – at the edge of chaos the behaviour is potentially structurally richer and therefore, has a higher predictive capacity.

5. Self-Organisation

Three ideas are implicit in the word self-organisation: a) the organisation in terms of global implicit coordination; b) the dynamics implicit in progressing in time from a not (or less) organised to an organised state; and c) the spontaneous arising of such dynamics. To avoid semantic traps, it is important to notice that the word 'spontaneous' should not be taken literally; we deal with open systems, exchanging energy, matter and/or information with the environment and made up of components whose properties and behaviours are defined prior to the organisation itself. The 'self' prefix merely states that no centralised ordering or external agent/template explicitly guides the dynamics. It is thus necessary to define what is meant by 'organisation' and how its arising or increase can be detected.

5.1. *Concept*

A commonly held view is that organisation entails an increase in complexity. Unfortunately the lack of agreement of what we mean by complexity leaves such definition somehow vague. For example, De Wolf and Holvoet [28] refer to complexity as a measure of redundancy or structure in the system. The concept can be made more formal by adopting the statistical complexity described above as a measure of complexity, as demonstrated in Shalizi [60] and Shalizi et al. [64]. This definition offers several of the advantages of the Computational Mechanics approach; it is computable and observer independent. Also, it captures the intuitive notion that the more a system self-organises, the more behaviours it can display, the more effort is needed to describe its dynamics. Importantly, this needs to be seen in a statistical perspective; while a disorganised system may potentially display a larger number of actual configurations, the distinction among several of them may not be statistically significant. Adopting the statistical complexity allows us to focus on the system configurations which are statistically different (causal states) for the purpose at hand. We thus have a measure which is based only on the internal dynamics of the system (and consequently is observer-independent) but which can be tuned according to the purpose of the analysis.

For an alternative definition of self-organisation based on thermodynamics and

the distinction between self-organisation and the related concept of self-assembly we refer the reader to Halley and Winkler [39].

5.2. *Information-theoretic interpretation*

In the scientific literature the concept of self-organisation referrs to both living and non living systems, ranging from physics and chemistry to biology and sociology. Kauffman [43] suggests that the underlying principle of self-organisation is the generation of constraints in the release of energy. According to this view, the constrained release allows for such energy to be controlled and channelled to perform some useful work. This work in turn can be used to build better and more efficient constraints for the release of further energy and so on; this principle is closely related to Kauffman's own definition of life [43]. It helps us to understand why an organised system with effectively less available configurations may behave and look more complex than a disorganised one to which, in principle, more configurations are available. The ability to constrain and control the release of energy may allow a system to display behaviours (reach configurations) which, although possible, would be extremely unlikely in its non-organised state. It is surely possible that 100 parrots move independently to the same location at the same time, but this is far more likely if they fly in a flock. A limited number of coordinated behaviours become implementable because of self-organisation, which would extremely unlikely to arise in the midst of a nearly infinite number of disorganised configurations. The ability to constrain the release of energy thus provides the self-organised system with behaviours that can be selectively chosen for successful adaptation.

However, Halley and Winkler [39] correctly point out that attention should paid to how self-organisation is treated if we want the concept to apply equally to both living and non-living systems. For example, while it is temping to consider adaptation as a guiding process for self-organisation, it then makes it hard to use the same definition of self-organisation for non-living systems.

Recently, Correia [19] analysed self-organisation motivated by embodied systems, i.e. physical systems situated in the real world, and established four fundamental properties of self-organisation: no external control, an increase in order, *robustness*[e], and interaction. All of these properties are easily interpretable in terms of information transfer. Firstly, the absence of external control may correspond to 'spontaneous' arising of information dynamics without any flow of information into the self-organising system. Secondly, an increase in order or complexity reflects simply that the predictive information is increased within the system or its specific part: $I_{pred}([t_1 - T, t_1], [t_1, t_1 + T']) < I_{pred}([t_2 - T, t_2], [t_2, t_2 + T'])$, and $C_\mu^{System}(t_2) > C_\mu^{System}(t_1)$, for $t_2 > t_1$ and positive T and T', where

[e]Although Correia refers to this as adaptability, according to the concepts in this paper he in fact defines robustness. This is an example of exactly the kind of issue we hope to avoid by developing this dictionary.

$C_\mu^{System}(t)$ is the statistical complexity at time t. In general, however, we believe that one may relax the distinction between these two requirements and demand only that in a self-organising system, the amount of information flowing from the outside $I_{Outside}$ is strictly less than the change in the predictive information's gain: $I_{Outside} < I_{pred}([t_2 - T, t_2], [t_2, t_2 + T']) - I_{pred}([t_1 - T, t_1], [t_1, t_1 + T'])$, or $C_\mu^{Outside} < C_\mu^{System}(t_2) - C_\mu^{System}(t_1)$, where $C_\mu^{Outside}$ is the complexity of the contribution from the outside.

In general, a system is robust if it continues to function in the face of perturbations [71]. Information-theoretically, robustness of a self-organising system to perturbations means that it may interleave stages of an increased information transfer within some channels (dominant patterns are being exploited) with periods of decreased information transfer (alternative patterns are being explored).

The interaction property is described by Correia [19] as follows: "minimisation of local conflicts produces global optimal self-organisation, which is evolutionarily stable." Minimisation of local conflicts, however, is only one aspect, captured in Equations (4), (8), and (12) as equivocation or non-assortativeness, and should be generally complemented by maximising diversity within the system.

5.3. *Example – self-organising traffic*

In the context of pedestrian traffic, Correia [19] argues that it can be shown that the "global efficiency of opposite pedestrian traffic is maximised when interaction rate is locally minimised for each component. When this happens two separate lanes form, one in each direction. The minimisation of interactions follows directly from maximising the average velocity in the desired direction." In other words, the division into lanes results from maximizing velocity (an overall objective or fitness), which in turn supports minimization of conflicts.

Another example is provided by ants: "Food transport is done via a trail, which is an organised behaviour with a certain complexity. Nevertheless, a small percentage of ants keeps exploring the surroundings and if a new food source is discovered a new trail is established, thereby dividing the workers by the trails [42] and increasing complexity" [19]. Here, the division into trails is again related to an increase in fitness and complexity.

These two examples demonstrate that when local conflicts are minimised, the degree of coupling among the components (i.e. interaction) increases and the information flows easier, thus increasing the predictive information. This means that not only the overall diversity of a system is important (more lanes or trails), but the interplay among different channels (the assortative noise within the system, the conflicts) is crucial as well.

5.4. *Example – self-regulating morphogenetic processes*

Modeling information processing in bio-systems is concerned with the nature of biological information and the ways in which it is processed in biological and ar-

tificial cells and tissues. There are a number of models involving self-organising or self-regulating processes within complex biological systems. For example, some morphogenetic models explore the requirements for spontaneous formation of regular and stable structures out of more or less homogeneous cell aggregations or cell sheets – i.e. the creation of a new form apparently without a detailed external blueprint. In particular, the biomechanical model of Beloussov and Grabovsky [7] reproduces two main categories of patterns, stationary cell domains and running waves. It interplays short- and long-range factors in a "causal chain" moving from the state of an initial homogeneous cell layer towards the complicated shapes of embryonic rudiments: "the local perturbations act in ensemble, rather than in a mosaic, one-to-one manner"[f]. The model involves a measure of the cells' mechanosensitivity (i.e. their ability to transform external mechanical stresses into active mechanochemical reactions), and changes in the tangential pressure due to cell extensions and contractions. An increase in mechanosensitivity is related to the formation of regular stationary structures and propagating waves, travelling across embryonic tissues, while an increase of the limits of tangential extension is essentially destructive, deteriorating regular structures and generating irregular waves. This interplay between two forces may be interpreted in information-theoretic terms: the information transfer is optimal when the diversity in mechanosensitivity is large, and the assortative noise (conflicts) inducing tensions is minimal. Spontaneous formation occurs, thus, within an optimal range of the information transfer and not at either of these extremes.

5.5. *Example – self-controlling neural automata*

One possible pathway towards an optimal solution in the information-theoretic search-space is explored in the context of neural networks with dynamic synapses. Cortes et al. [20] studied neural automata (neurobiologically inspired cellular automata) which exhibit "chaotic itinerancy among the different stored patterns or memories". In other words, activity-dependent synaptic fluctuations ("noise") explore the search-space by continuously destabilising a current attractor and inducing random hopping to other possible attractors. The complexity and chaoticity of the resulting dynamics (hopping) depends on the intensity of the synaptic "noise" and on the number of the network nodes that are synchronously updated (we note again the two forces involved in the information transfer – diversity and non-assortativeness). Cortes et al. [20] utilised a quantitative measure – the entropy of neural activity over time (computed in frequency-domain), and related varying values of synaptic noise parameter F to different regimes of chaoticity. Decreasing entropy for mid-range values of F indicates a tendency towards regularization or

[f]We would like to point out that in this example, as well as in many others, there are external initial conditions, and this requires a precise estimation of the external influence $C_\mu^{Outside}$, contrasted with the increase in complexity $C_\mu^{System}(t_2) - C_\mu^{System}(t_1)$ during morphogenesis.

smaller chaoticity. Importantly, hopping controlled by synaptic noise may occur autonomously, without the need for external stimuli.

5.6. *Example – self-organising locomotion*

The internal channels through which information flows within the system are observer-independent, but different observers may select different channels for a specific analysis. For example, let us consider a modular robotic system modelling a multi-segment snake-like (salamander) organism, with actuators ("muscles") attached to individual segments ("vertebrae"). A particular side-winding locomotion emerges as a result of individual control actions when the actuators are coupled within the system and follow specific evolved rules [56, 55]. There is no global coordinator component in the evolved system, and it can be shown that the amount of predictive information between groups of actuators grows as the modular robot starts to move across the terrain – that is, the distributed actuators become more coupled when a coordinated side-winding locomotion is dominant. Faced with obstacles, the robot temporarily loses the side-winding pattern: the modules become less organised, the strength of their coupling is decreased, and rather than exploiting the dominant pattern, the robot explores various alternatives. Such exploration temporarily decreases self-organisation, i.e. the predictive information within the system. When the obstacles are avoided, the modules "rediscover" the dominant side-winding pattern by themselves, recovering the previous level of predictive information and manifesting again the ability to self-organise without any global controller. Of course, the "magic" of this self-organisation is explained by properties defined *a priori*: the rules employed by the biologically-inspired actuators have been obtained by a genetic programming algorithm, while the biological counterpart (the rattlesnake *Crotalus cerastes*) naturally evolved over long time. Our point is simply that we can measure the dynamics of predictive information and statistical complexity as it presents itself within the channels of interest.

In summary, the fundamental properties of self-organisation are immediately related to information dynamics, and can be studied in precise information-theoretic terms when the appropriate channels are identified.

6. Emergence

Nature can be observed at different levels of resolution, be these intended as spatial or temporal scales or as measurement precision. For certain phenomena this affects merely the level of details we can observe. As an example, depending on the scale of observation, satellite images may highlight the shape of a continent or the make of a car; similarly, the time resolution of a temperature time series can reflect local stochastic (largely unpredictable) fluctuations or daily periodic (fairly predictable) oscillations. There are classes of phenomena though, which when observed at different levels, display behaviours which appear *fundamentally* different. The quantum phenomena of the 'very small' and the relativistic effects of the 'very large' do not

seem to find obvious realisations in our everyday experience at the middle scale; similarly the macroscopic behaviour of a complex organism appears to transcend the biochemistry it derives from. The apparent discontinuity between these radically different phenomena arising at different scales is usually, broadly and informally, defined as emergence.

Attempts to formally address the study of emergence have sprung at regular intervals in the last century or so (for a nice review see Corning [18]), under different names, approaches and motivations and is currently receiving a new burst of interest. Here we borrow from Crutchfield [22], who, in a particularly insightful work, proposes a distinction between two phenomena which are commonly viewed as expression of emergence: pattern formation and 'intrinsic' emergence.

6.1. *Concept*

Pattern Formation. In pattern formation we imagine an observer trying to 'understand' a process. If the observer detects some patterns (structures) in the system, they can then employ such patterns as tools to simplify their understanding of the system. As an example, a gazelle which learns to correlate hearing a roaring to the presence of a lion, will be able to use it as warning and flee danger. Not being able to detect the pattern 'roaring = lion close by' would require the gazelle to detect more subtle signs, possibly needing to employ more attention and thus more effort. In this setting the observer (gazelle) is 'external' to the system (lion) it needs to analyse.

Intrinsic emergence. In intrinsic emergence, the observer is 'internal' to the system. Imagine a set of traders in an economy. The traders are locally connected via their trades, but no global information exchange exists. Once the traders identify an 'emergent' feature, like the stock market, they can employ it to understand and *affect* the functioning of the system itself. The stock market becomes a mean for global information processing, which is performed by the agents (that is, the system itself) to affect their *own* functioning.

6.2. *Information-theoretic interpretation*

Given that a system can be viewed and studied at different levels, a natural question is "what level should we choose for our analysis"? A reasonable answer could be "the level at which it is easier or more efficient to construct a workable model". This idea has been captured formally by Shalizi [60] in the definition of Efficiency of Prediction. Within a Computational Mechanics [61] framework, Shalizi suggests:

$$e = \frac{E}{C_\mu} \tag{13}$$

where e is the Efficiency of Prediction, E is the excess entropy and C_μ the statistical complexity discussed above. The excess entropy can be seen as the mutual information between the past and future of a process, that is, the amount of information

observed in the past which can be used to predict the future (i.e. which can be usefully coded in the agent instructions on how to behave in the future). Recalling that the statistical complexity is defined as the amount of information needed to reconstruct a process (that is equivalent to performing an optimal prediction), we can write informally:

$$e = \frac{\text{how much can be predicted}}{\text{how difficult it is to predict}} \tag{14}$$

Given two levels of description of the same process, the approach Shalizi suggests is to choose for analysis the level which has larger efficiency of prediction e. At this level, either:

- we can obtain better predictability (understanding) of the system (E is larger), or
- it is much easier to predict because the system is simpler (C_μ is smaller), or
- we may lose a bit of predictability (E is smaller) but at the benefit of much larger gain in simplicity (C_μ is much smaller).

We can notice that this definition applies equally to pattern formation as well as to intrinsic emergence. In the case of pattern formation, we can envisage the scientist trying to determine what level of enquiry will provide a better model. At the level of intrinsic emergence, developing an efficient representation of the environment and of *its own functioning within the environment* gives a selective advantage to the agent, either because it provides for a better model, or because it provides for a similar model at a lower cost, enabling the agent to direct resources towards other activities.

6.3. *Example – the emergence of thermodynamics*

A canonical example of emergence without self-organisation is described by Shalizi [60]: thermodynamics can emerge from statistical mechanics. The example considers a cubic centimeter of argon, which is conveniently spinless and monoatomic, at standard temperature and pressure, and sample the gas every nanosecond. At the micro-mechanical level, and at time intervals of 10^{-9} seconds, the dynamics of the gas are first-order Markovian, so each microstate is a causal state. The thermodynamic entropy (calculated as $6.6 \cdot 10^{20}$ bits) gives the statistical complexity C_μ. The entropy rate h_μ of one cubic centimeter of argon at standard temperature and pressure is quoted to be around $3.3 \cdot 10^{29}$ bits per second, or $3.3 \cdot 10^{20}$ bits per nanosecond. Given the range of interactions $R = 1$ for a first-order Markov process, and the relationship $E = C_\mu - Rh_\mu$ [34], it follows that the efficiency of prediction $e = E/C_\mu$ is about 0.5 at this level. Looking at the macroscopic variables uncovers a dramatically different situation. The statistical complexity C_μ is given by the entropy of the macro-variable energy which is approximately 33.28 bits, while the entropy rate per millisecond is 4.4 bits (i.e. $h_\mu = 4.4 \cdot 10^3$ bits/second). Again, the assumption that the dynamics of the macro-variables are Markovian, and the

relationship $E = C_\mu - Rh_\mu$ yield $e = E/C_\mu = 1 - Rh_\mu/C_\mu = 0.87$. If the time-step is a nanosecond, like at the micro–mechanical level, then $e \approx 1$, i.e. the efficiency of prediction approaches maximum. This allows Shalizi to conclude that "almost all of the information needed at the statistical-mechanical level is simply irrelevant thermodynamically", and given the apparent differences in the efficiencies of prediction at two levels, "thermodynamic regularities are emergent phenomena, emerging out of microscopic statistical mechanics" [60].

7. Adaptation and Evolution

Adaptation is a process where the behaviour of the system changes such that there is an increase in the mutual information between the system and a potentially complex and non-stationary environment. The environment is treated as a black box, meaning an adaptive system does not need to understand the underlying system dynamics to adapt. Stimulus response interactions provide feedback that modifies an internal model or representation of the environment, which affects the probability of the system taking future actions.

7.1. *Concept*

The three essential functions for an adaptive mechanism are generating variety, observing feedback from interactions with the environment, and selection to reinforce some interactions and inhibit others. Without variation, the system cannot change its behaviour, and therefore it cannot adapt. Without feedback, there is no way for changes in the system to be coupled to the structure of the environment. Without preferential selection for some interactions, changes in behaviour will not be statistically different to a random walk. First order adaptation keeps sense and response options constant and adapts by changing only the probability of future actions. However, adaptation can also be applied to the adaptive mechanism itself [38]. Second order adaptation introduces three new adaptive cycles: one to improve the way variety is generated, another to adapt the way feedback is observed and thirdly an adaptive cycle for the way selection is executed. If an adaptive system contains multiple autonomous agents using second order adaptation, a third order adaptive process can use variation, feedback and selection to change the structure of interactions between agents.

From an information-theoretic perspective, variation decreases the amount of information encoded in the system, while selection acts to increase information. Since adaptation is defined to increase mutual information between a system and its environment, the information loss from variation must be less than the increase in mutual information from selection.

For the case that the system is a single agent with a fixed set of available actions, the environmental feedback is a single real valued reward plus the observed change in state at each time step, and the internal model is an estimate of the future value

of each state, this model of first order adaptation reduces to reinforcement learning (see for example [69]).

For the case that the system contains a population whose generations are coupled by inheritance with variation under selective pressure, the adaptive process reduces to evolution. Evolution is not limited to DNA/RNA based terrestrial biology, since other entities, including prions and artificial life programs, also meet the criteria for evolution. Provided a population of replicating entities can make imperfect copies of themselves, and not all the entities have an equal capacity to survive, the system is evolutionary. This broader conception of evolution has been coined universal Darwinism by Dawkins [27].

7.2. *Information-theoretic interpretation*

Adami [1] advocated the view that "evolution increases the amount of information a population harbors about its niche". In particular, he proposed physical complexity – a measure of the amount of information that an organism stores in its genome about the environment in which it evolves. Importantly, physical complexity for a population X (an ensemble of sequences) is defined in relation to a specific environment Z, as mutual information:

$$I(X, Z) = H_{max} - H(X|Z) \tag{15}$$

where H_{max} is the entropy in the absence of selection, i.e. the unconditional entropy of a population of sequences, and $H(X|Z)$ is the conditional entropy of X given Z, i.e. the diversity tolerated by selection in the given environment. When selection does not act, no sequence has an advantage over any other, and all sequences are equally probable in ensemble X. Hence, H_{max} is equal to the sequence length. In the presence of selection, the probabilities of finding particular genotypes in the population are highly non-uniform, because most sequences do not fit the particular environment. The difference between the two terms in 15 reflects the observation that "If you do not know which system your sequence refers to, then whatever is on it cannot be considered information. Instead, it is potential information (a.k.a. entropy)". In other words, this measure captures the difference between potential and selected (filtered) information:

$$\begin{aligned}\text{physical complexity} = \text{how much data can be stored} - \\ \text{how much data irrelevant to environment is stored}\end{aligned} \tag{16}$$

Comparing this with the information transfer through networks, Equation (12), as well as analogous information dynamics Equations (4) and (8), we can observe a strong similarity: "how much data can be stored" is related to diversity of the network, while "how much data irrelevant to environment is stored" (or "how much conflicting data") corresponds to assortative noise in the network. In short, natural selection increases physical complexity by the amount of information a population contains about its environment. Adami argued that physical complexity must

increase in molecular evolution of asexual organisms in a single niche if the environment does not change, due to natural selection, and that "natural selection can be viewed as a filter, a kind of semipermeable membrane that lets information flow into the genome, but prevents it from flowing out". In general, information may flow out and it is precisely this dynamic that creates larger feedback loops in the environment.

7.3. *Example – perception-action loops*

The information transfer can also be interpreted as the acquisition of information from the environment by a single adapting individual: there is evidence that pushing the information flow to the information-theoretic limit (i.e. maximization of information transfer) can give rise to intricate behaviour, induce a necessary structure in the system, and ultimately adaptively reshape the system [44, 45]. The central hypothesis of Klyubin et al. is that there exists "a local and universal utility function which may help individuals survive and hence speed up evolution by making the fitness landscape smoother", while adapting to morphology and ecological niche. The proposed general utility function, *empowerment*, couples the agent's sensors and actuators via the environment. Empowerment is the perceived amount of influence or control the agent has over the world, and can be seen as the agent's potential to change the world. It can be measured via the amount of Shannon information that the agent can "inject into" its sensor through the environment, affecting future actions and future perceptions. Such a perception-action loop defines the agent's actuation channel, and technically empowerment is defined as the capacity of this actuation channel: the maximum mutual information for the channel over all possible distributions of the transmitted signal. "The more of the information can be made to appear in the sensor, the more control or influence the agent has over its sensor" – this is the main motivation for this local and universal utility function [45]. Other examples highlighting the role of information transfer in guiding selection of spatiotemporally stable multi-cellular patterns, well-connected network topologies, and coordinated actuators in a modular robotic system are discussed in [59, 58, 55, 56].

8. Self-referentiality

Patterns of intrinsic emergence may form as a result of observer-independent processes – nevertheless, these are patterns only with respect to a certain observation structure, imposed on or selected in the environment *a priori*, and not necessarily by the "observer". An interesting aspect, however, is that the choice of such a structure is not entirely decoupled from the environment. In fact, this choice often depends on the emergent phenomena. So the process can be best characterised in terms of tangled hierarchies exhibiting Strange Loops: "an interaction between levels in which the top level reaches back down towards the bottom level and influences it, while at the same time being itself determined by the bottom level" [41].

8.1. *Concept*

This view involves the concept of *downward causation* [36, 10, 40]: a feature is emergent if it has some sort of causal power on lower level entities. While common views of emergence assume that lower level entities must have an "upward" causation on the emergent features, this approach requires a 2-way causal relation. As an example, we can imagine individuals organising into a community (e.g. the Bar problem [3]). Their actions affect how the community develops (upward causality) and the development of the community itself affects the behaviour and interaction of the individuals (downward causality). An important feature of such systems is that, in the absence of explicit information sharing, each individual may need to "best-guess" what the other individuals may do over time. Systems that fall into this category are sometimes called self-referential problems – situations where the forecasts made by the individuals serve to create the world they are trying to forecast [6, 5]. In these systems, the "best" thing to do depends not on a rational decision model (which does not exist), but on what everyone else is doing, creating a heterogeneous mix of decision strategies that continually co-evolve over time [3, 14, 4]. Even if the individuals are allowed to communicate and share some information, the heterogeneity prevails [31].

8.2. *Information-theoretic interpretation*

A closely related problem is the minority game [17]: a repeated game where N (odd) agents have to choose one out of two alternatives at each time step, and those who happen to be in the minority win. As noted by Batten [6], "seemingly simple at first glance, the game is subtle in the sense that if all players analyze the situation identically, all will choose the same alternative and therefore all will lose". The agents may share information on the outcomes of the game in past M rounds (memory). In the minority game, there is a second order phase transition in the space of the parameter $\alpha = 2^M/N$, between a symmetric phase (where no past information is available to agents) and an asymmetric phase (where past information is available). An important feature of minority games is the possibility for some agents (risk-averters) to imitate others (risk-takers) in selecting their actions. As observed by Slanina [67], there is an optimal level of imitation, beyond which the whole collective starts to perform worse in terms of the average gain of the agents: "moderate imitation can be beneficiary, while exaggerated one can be harmful". In addition, this optimal level increases with decreasing memory length M: the less information is shared, the more imitation is allowed.

8.3. *Example – shortest path formation by ants*

Another well-known example of downward causation is shortest path formation with ants: ants indirectly interact through changes in their environment by depositing pheromones and forming a pheromone trail. This mechanism for increasing shared

information is known as *stigmergy*. The probability that an ant chooses a trail increases with the number of ants that chose the same path in the past – the process is thus characterised by a positive feedback loop (a form of autocatalytic behaviour called *allelomimesis*) [29]. Since the pheromone evaporates (loss of information), the shortest path lasts longer than the alternatives, and the ants have a higher probability of using and reinforcing it. Thus, not only the shortest path emerges as a result of allelomimesis and stigmergy (upward causation), it also influences the movement of the ants because they follow the pheromone trail (downward causation) [28]. Importantly, we again observe diversity in ant behaviours (path explorers and path exploiters).

8.4. *Summary*

The observed diversity in agent behaviours (risk-takers and risk-averters, or explorers and exploiters) creates information dynamics: loss of information, or "forgetting" effect [46], and gain of information, or "information selectivity" [47]. Our conjecture is that self-referentiality eventually forces a split in information dynamics, breaking the cycle between upward and downward causation. As a result of such a break, some form of memory (implicit or explicit) emerges in the system in order to support both diversity and imitation.

In summary, in self-referential systems there is also a relationship between heterogeneity, the level of non-assortative mixing, and the information shared within the system – similar to the information transfer relationships (4), (8), (12), and (16). A precise information-theoretic analysis of this relationship for self-referential systems is a subject of future research (see Parunak et al. [54] and Prokopenko et al. [57] for preliminary analysis).

9. Discussion and Conclusions

By studying the processes which result from the local interaction of relatively simple components, Complex System Science has accepted the audacious aim of addressing problems which range from physics to biology, sociology and ecology. It is not surprising that a common framework and language which enable practitioners of different field to communicate effectively is still lacking. As a possible contribution to this goal we have proposed a framework within which concepts like complexity, emergence and self-organisation can be described, and most importantly, distinguished.

Figure 1 illustrates some relationships between the concepts introduced in this paper. In particular, it shows two levels of an emergence hierarchy that are used to describe a complex system. The figure depicts dynamics that tend to increase complexity as arrows from left to right, and increases in the level of organisation as arrows from bottom to top. The concepts can be related in numerical order as follows. (1) demonstrates self-organisation, as components increase in organisation

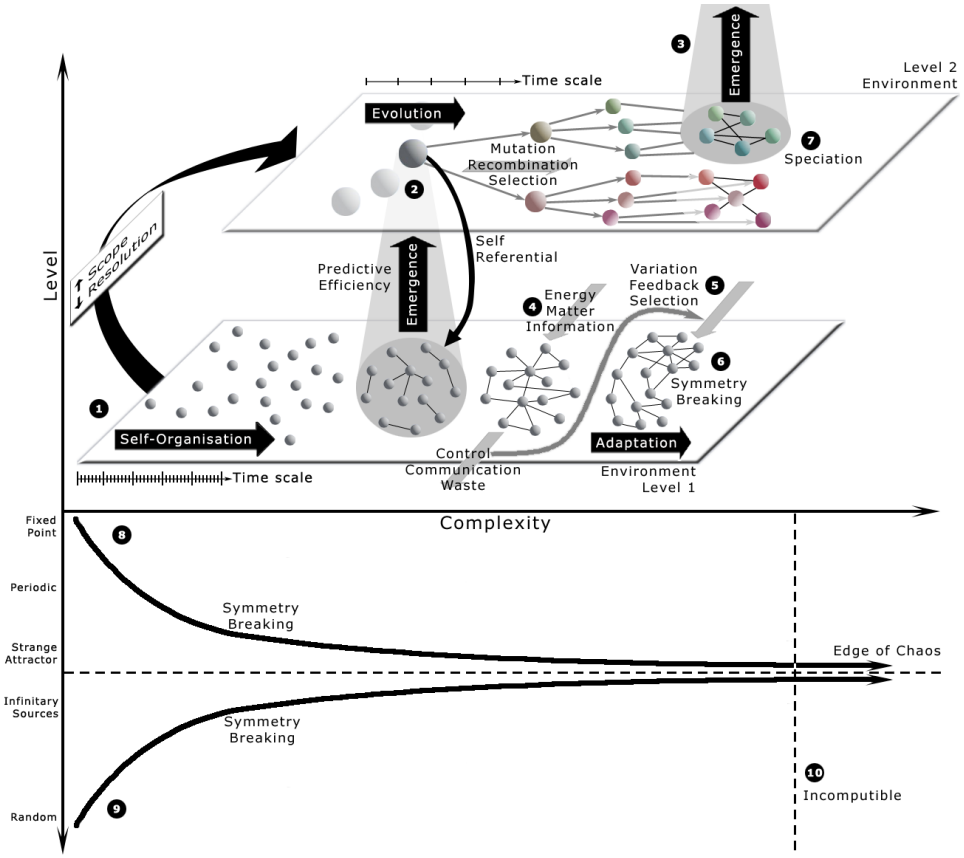

Fig. 1. A systems view of Complex Systems Science concepts.

over time. As the components become more organised, interdependencies arise constraining the autonomy of the components, and at some point it is more efficient to describe tightly coupled components as an emergent whole (or system). (2) depicts a lower resolution description of the whole, which may be self-referential if it causally affects the behaviour of its components. Note that Level 2 has a longer time scale. The scope at this level is also increased, such that the emergent whole is seen as one component in a wider population. As new generations descend with modification through mutation and/or recombination, natural selection operates on variants and the population evolves. (3) shows that interactions between members of a population can lead to the emergence of higher levels of organisation: in this case, a species is shown. (4) emphasises flows between the open system and the

environment in the Level 1 description. Energy, matter and information enter the system, and control, communication and waste can flow back out into the environment. When the control provides feedback between the outputs and the inputs of the system in (5), its behaviour can be regulated. When the feedback contains variation in the interaction between the system and its environment, and is subject to a selection pressure, the system adapts. Positive feedback that reinforces variations at (6) results in symmetry breaking and/or phase transitions. (7) shows analogous symmetry breaking in Level 2 in the form of speciation.

Below the complexity axis, a complementary view of system complexity in terms of behaviour, rather than organisation, is provided. Fixed point behaviour at (8) has low complexity, which increases for deterministic periodic and strange attractors. The bifurcation process is a form of symmetry breaking. Random behaviour at (9) also has low complexity, which increases as the system's components become more organised into processes with "infinitary sources" [24]: e.g. positive-entropy-rate variations on the Thue-Morse process and other stochastic analogues of various context-free languages. This class of behaviour ("infinitary sources") is depicted in (1). The asymptote between (8) and (9) is interpreted as the 'edge of chaos', where the complexity can grow without bound. Beyond some threshold of complexity at (10), the behaviour is incomputable: it cannot be simulated in finite time on a Universal Turing Machine.

For our discussion we chose an information-theoretical framework. There are four primary reasons for this choice:

(1) it enables clear and consistent definitions and relationships between complexity, emergence and self-organisation in the physical world;
(2) the same concepts can equally be applied to biology;
(3) from a biological perspective, the basic ideas naturally extend to adaptation and evolution, which begins to address the question of why complexity and self-organisation are ubiquitous and apparently increasing in the biosphere; and
(4) it provides a unified setting, within which the description of relevant information channels provides significant insights of practical utility.

As noted earlier, once the information channels are identified by designers of a physical system (or naturally selected by interactions between a bio-system and its environment), the rest is mostly a matter of computation. This computation can be decomposed into "diversity" and "equivocation", as demonstrated in the discussed examples.

Information Theory is not a philosophical approach to the reading of natural processes, rather it comes with a set of tools to carry out experiments, make predictions, and computationally solve real-world problems. Like all toolboxes, its application requires a set of assumptions regarding the processes and conditions regarding data collections to be satisfied. Also, it is by definition biased towards a view of Nature as an immense information processing device. Whether this view and these tools can be successfully applied to the large variety of problems Complex

Systems Science aims to address is far from obvious. Our intent, at this stage, is simply to propose it as a framework for a less ambiguous discussion among practitioners from different disciplines. The suggested interpretations of the concepts may be at best temporary place-holders in an evolving discipline – hopefully, the improved communication which can arise from sharing a common language will lead to deeper understanding, which in turn will enable our proposals to be sharpened, rethought and even changed altogether.

Acknowledgements

This research was conducted under the CSIRO emergence interaction task (http://www.per.marine.csiro.au/staff/Fabio.Boschetti/CSS_emergence.htm) with support from the CSIRO Complex Systems Science Theme (http://www.csiro.au/css). Thanks to Cosma Shalizi and Daniel Polani for their insightful contributions.

References

[1] C. Adami. What is complexity? *Bioessays*, 24(12):1085–1094, 2002.
[2] D. Arnold. Information-theoretic analysis of phase transitions. *Complex Systems*, 10:143–155, 1996.
[3] W. B. Arthur. Inductive behaviour and bounded rationality. *The American Economic Review*, 84:406–411, 1994.
[4] D. F. Batten. *Discovering Artificial Economics*. Westview Press, New York, 2000.
[5] D. F. Batten. Are some human ecosystems self-defeating? *Environmental Modelling and Software*, in press, 2006.
[6] F. Batten, D. Simulating human behaviour: the invisible choreography of self-referential systems. In *The International Environmental Modelling & Software Society (IEMSS) Conference 2004*, Osnabruck, Germany, 2004.
[7] L. V. Beloussov and V. I. Grabovsky. A common biomechanical model for the formation of stationary cell domains and propagating waves in the developing organisms. *Computer Methods in Biomechanics and Biomedical Engineering*, 8(6):381–91, 2005.
[8] W. Bialek, I. Nemenman, and N. Tishby. Complexity through nonextensivity. *Physica A*, 302:89–99, 2001.
[9] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neur. Comp.*, 13:2409–2463, 2001.
[10] M. H. Bickhard. Emergence. In E. C. F. N. O. C. P. V. Andersen, P. B., editor, *Downward Causation*, pages 322–348. University of Aarhus Press, Aarhus, Denmark, 2000.
[11] G. Boffetta, M. Cencini, M. Falcioni, and A. Vulpiani. Predictability: a way to characterize complexity. *Physics Reports*, 356:367–474, 2002.
[12] F. Boschetti and N. Grigg. Mapping the complexity of ecological models. *submitted to Complexity*, 2006.
[13] J. L. Casti. Chaos, godel and truth. In J. L. Casti and A. Karlqvist, editors, *Beyond Belief: Randomness, Prediction, and Explanation in Science*. CRC Press, 1991.
[14] J. L. Casti. Would-be worlds: the science and surprise of artificial worlds. *Computers, Environment and Urban Systems*, 23:193–203, 1999.
[15] G. J. Chaitin. Information-theoretic limitations of formal systems. *Journal of the ACM*, 21:403–424, 1974.

[16] G. J. Chaitin. *The limits of mathematics: a course on information theory & limits of formal reasoning*. Springer, New York, 1997.
[17] D. Challet and Y.-C. Zhang. On the minority game: analytical and numerical studies. *Physica A*, 256(514):check this!, 1998.
[18] P. A. Corning. The re-emergence of "emergence": A venerable concept in search of a theory. *Complexity*, 7(6):18–30, 2002.
[19] L. Correia. Self-organisation: a case for embodiment. In *Proceedings of The Evolution of Complexity Workshop at Artificial Life X: The 10th International Conference on the Simulation and Synthesis of Living Systems*, pages 111–116, 2006.
[20] J. M. Cortes, J. Marro, and J. J. Torres. Control of neural chaos by synaptic noise. *Biosystems*, in press, 2006.
[21] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.
[22] J. Crutchfield. The Calculi of Emergence: Computation, Dynamics, and Induction. *Physica D*, 75:11–54, 1994.
[23] J. P. Crutchfield and D. P. Feldman. Statistical complexity of simple one-dimensional spin systems. *Phys. Rev. E*, 55(2):1239R1243R, 1997.
[24] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos*, 13(1):25–54, 2003.
[25] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica 7D*, pages 201–223, 1983.
[26] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108, 1989.
[27] R. Dawkins. Universal darwinism. In D. Bendall, editor, *Evolution from Molecules to Men*. Cambridge University Press, 1983.
[28] T. De Wolf and T. Holvoet. Emergence versus self-organisation: Different concepts but promising when combined. In S. Brueckner, G. D. M. Serugendo, A. Karageorgos, and R. Nagpal, editors, *Engineering Self-Organising Systems*, page 115. Springer, 2005.
[29] M. Dorigo, V. Maniezzo, and A. Colorni. The ant system: optimization by a colony of cooperating agents. *IEEE Trans. Syst. Man Cybern. B*, 26(1):113, 1996.
[30] W. Ebeling. Prediction and entropy of nonlinear dynamical systems and symbolic sequences with lro. *Physica D*, 109:4252, 1997.
[31] B. Edmonds. Gossip, sexual recombination and the el farol bar: modelling the emergence of heterogeneity. *Journal of Artificial Societies and Social Simulation*, 2(3):1–21, 1999.
[32] P. Erdos and A. Renyi. On the strength of connectedness of random graphs. *Acta Mathematica Scientia Hungary*, 12:261–267, 1961.
[33] K.-E. Eriksson and K. Lindgren. Structural information in self-organizing systems. *Physica Scripta*, 35:38897, 1987.
[34] D. P. Feldman and J. P. Crutchfield. Discovering noncritical organization: Statistical mechanical, information theoretic, and computational views of patterns in one-dimensional spin systems. Technical Report 98-04-026, SFI Working Paper, 1998.
[35] D. P. Feldman and J. P. Crutchfield. Structural information in two-dimensional patterns: Entropy convergence and excess entropy. *Physical Review E*, 67, 2003.
[36] J. Goldstein. The singular nature of emergent levels: Suggestions for a theory of emergence. *Nonlinear Dynamics, Psychology, and Life Sciences*, 6(4), 2002.
[37] P. Grassberger. *Int. J. Theor. Phys.*, 25:907938, 1986.
[38] A.-M. Grisogono. Co-adaptation. In *SPIE Symposium on Microelectronics, MEMS and Nanotechnology*, volume Paper 6039-1, Brisbane, Australia, 2005.
[39] J. Halley and D. Winkler. Towards consistent concepts of self-organization and self-

assembly, in prep. 2006.
[40] F. Heylighen. Modelling emergence, world futures. *Journal of General Evolution, special issue on creative evolution*, check this!, 1991.
[41] D. R. Hofstadter. *Godel, Escher, Bach: An eternal golden braid*. Vintage Books, New York, 1989.
[42] S. P. Hubbell, L. K. Johnson, E. Stanislav, B. Wilson, and H. Fowler. Foraging by bucket-brigade in leafcutter ants. *Biotropica*, 12(3):210213, 1980.
[43] S. A. Kauffman. *Investigations*. Oxford University Press, Oxford, 2000.
[44] A. S. Klyubin, D. Polani, and C. L. Nehaniv. Organization of the information flow in the perception-action loop of evolved agents. In *Proceedings of 2004 NASA/DoD Conference on Evolvable Hardware*, page 177180. IEEE Computer Society, 2004.
[45] A. S. Klyubin, D. Polani, and C. L. Nehaniv. All else being equal be empowered. In M. S. Capcarr'ere, A. A. Freitas, P. J. Bentley, C. G. Johnson, and J. Timmis, editors, *Advances in Artificial Life, 8th European Conference, ECAL 2005*, volume 3630 of LNCS, page 744753. Springer, 2005.
[46] L. Lancieri. *Memory and forgetting, two complementary mechanisms to characterize the various actors of the Internet and their interactions*. Phd thesis, University of Caen, 2000.
[47] L. Lancieri. Reusing implicit cooperation, a novel approach in knowledge management. *TripleC (Cognition, Cooperation, Communication) international journal*, 2(1):28–46, 2004.
[48] C. Langton. Computation at the edge of chaos: Phase transitions and emergent computation. In S. Forest, editor, *Emergent Computation*. MIT, 1991.
[49] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, New York, 2nd edition, 1997.
[50] W. Li. On the relationship between complexity and entropy for markov chains and regular languages. *Complex Systems*, 5(4):381399, 1991.
[51] K. Lindgren and M. G. Norhdal. Complexity measures and cellular automata. *Complex Systems*, 2(4):409440, 1988.
[52] M. Mitchell, P. T. Hraber, and J. P. Crutchfield. Revisiting the edge of chaos: evolving cellular automata to perform computations. *Complex Systems*, 7:89139, 1993.
[53] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(208701), 2002.
[54] H. V. D. Parunak, S. Brueckner, M. Fleischer, and J. Odell. Co-x: Defining what agents do together. In *Proceedings of Workshop on Teamwork and Coalition Formation*. AAMAS, 2002.
[55] M. Prokopenko, V. Gerasimov, and I. Tanev. Evolving spatiotemporal coordination in a modular robotic system. In *Proceedings of the 9th International Conference on the Simulation of Adaptive Behavior (SAB'06)*, pages 558–569, Rome, Italy, 2006.
[56] M. Prokopenko, V. Gerasimov, and I. Tanev. Measuring spatiotemporal coordination in a modular robotic system. In L. Rocha, L. Yaeger, M. Bedau, D. Floreano, R. Goldstone, and A. Vespignani, editors, *Artificial Life X: Proceedings of The 10th International Conference on the Simulation and Synthesis of Living Systems*, pages 185–191, Bloomington IN, USA, 2006.
[57] M. Prokopenko, D. Polani, and P. Wang. Optimizing potential information transfer with self-referential memory. In *Proceedings of 5th International Conference on Unconventional Computation (UC06)*, pages 228–242, York, UK, 2006.
[58] M. Prokopenko, P. Wang, M. Foreman, P. Valencia, D. C. Price, and G. T. Poulton. On connectivity of reconfigurable impact networks in ageless aerospace vehicles. *Journal of Robotics and Autonomous Systems*, 53(1):36–58, 2005.
[59] M. Prokopenko, P. Wang, D. C. Price, P. Valencia, M. Foreman, and F. A. J. Self-

organizing hierarchies in sensor and communication networks. *Artificial Life, Special Issue on Dynamic Hierarchies*, 11(4):407–426, 2005.

[60] C. Shalizi. *Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata*. PhD thesis, University of Michigan, 2001.

[61] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104:819–881, 2001.

[62] C. R. Shalizi and K. L. Shalizi. Optimal nonlinear prediction of random fields on networks. *Discrete Mathematics and Theoretical Computer Science*, AB(DMCS):11–30, 2003.

[63] C. R. Shalizi and K. L. Shalizi. Blind construction of optimal nonlinear recursive predictors for discrete sequences. In M. Chickering and J. Joseph Halpern, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference*, pages 504–511, Arlington, Virginia, 2004. AUAI Press.

[64] C. R. Shalizi, K. L. Shalizi, and R. Haslinger. Quantifying self-organization with optimal predictors. *Physical Review Letters*, 93(11):11870114, 2004.

[65] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October, 1948.

[66] R. Shaw. *The Dripping Faucet as a Model Chaotic System*. Aerial Press, Santa Cruz, California, 1984.

[67] F. Slanina. Social organization in the minority game model. *Physica A*, 286:367–376, 2000.

[68] R. V. Sole and S. Valverde. Information theory of complex networks: on evolution and architectural constraints. In E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai, editors, *Complex Networks*, volume 650 of *Lecture Notes in Physics*. Springer, 2004.

[69] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, The MIT Press, Cambridge, 1998.

[70] D. P. Varn. *Language Extraction from ZnS*. Phd thesis, University of Tennessee, 2001.

[71] A. Wagner. *Robustness and Evolvability in Living Systems*. Princeton University Press, Princeton, NJ, 2005.

[72] S. Wolfram. Universality and complexity in cellular automata. *Physica D*, 10, 1984.

[73] A. Wuensche. Classifying cellular automata automatically: Finding gliders, filtering, and relating space-time patterns, attractor basins, and the z parameter. *Complexity*, 4(3):47–66, 1999.