

Self-Organisation in Dynamical Systems:

A limiting result

Richard Johns
Department of Philosophy
University of British Columbia
johns@interchange.ubc.ca
updated: June 2009

0. INTRODUCTION

Self organization, or “order for free”, is an important (and expanding) area of inquiry. Self-organized structures occur in many contexts, including biology. While these structures may be intricate and impressive, there are some limitations on the kinds of structure than can self-organize, given the dynamical laws. (William Paley pointed out, for example, that a watch cannot be produced by “the laws of metallic nature”.) In this paper I will demonstrate that certain fundamental *symmetries* in the laws of physics constrain self organization in an interesting way. Roughly speaking, structures that are both large and non-self-similar cannot self organize in any dynamical system.

1. WHAT IS SELF-ORGANISATION?

The term “self-organisation” (SO for short) is used to describe the emergence of an object or structure “by itself”, or “spontaneously”, within a dynamical system. Of course the structure isn’t entirely uncaused – it arises *from the dynamics*. The easiest way to make the notion of SO precise is to exclude other possible causes of the structure, as follows:

1. The appearance of the object does not require a special, “fine-tuned” initial state.
2. There is no need for interaction with an external system.
3. The object is likely to appear in a reasonably short time.

The first two conditions are clear enough, ruling out cases where the structure is latent in the initial state of the system (like an oak tree from an acorn), and where the structure comes from outside (like an artist carving a sculpture). The third condition rules out cases of dumb luck and dogged persistence. A purely random dynamics, for example, might produce a watch with some fantastically small probability, or with a large probability given some fantastically long time, but these are not cases of self organization.

There are many kinds of object that appear by self organisation. Crystals are one obvious example. The vortex created by draining the bath tub is another. Living organisms are a case where self-organisation is largely, although not entirely, responsible, according to the standard evolutionary picture. This case is of particular interest, and will be discussed separately in sections 10 and 11.

2. LIMITS TO SELF-ORGANISATION

It is obvious enough that there are limits to self-organisation, as even simple arithmetic will show. Any given set of dynamical laws might produce *some* kinds of object spontaneously, but cannot produce *all* kinds of object that way. Consider, for example, the first 1000 objects that a set of laws produces, from a random initial state. It is clear that there cannot be more than 1000 distinct objects that are guaranteed to be in this set. And, similarly, there cannot be more than 100,000 objects with a better than 1% chance of being in this set.

For any given set of dynamical laws, therefore, we can ask such questions as: “Which types of object can these laws produce?”, and “Which types of object cannot these laws produce?”

Of course these questions are not too precise, as in a stochastic system it might turn out that every conceivable object is a possible member of the first 1000 products, but it's still true for most dynamical systems that some kinds of structure that tend to be produced much more quickly and probably than others. The need to describe this situation precisely will lead us to the concept of *saliency* below. In short, even if any

object can be produced at any time, some objects are still far more salient than others, with respect to the dynamics.

3. DYNAMICAL SYMMETRIES

In examining the question of which objects tend to be produced by a given set of dynamical laws, one important feature of those laws will be the *symmetries* they contain. The idea that symmetry in a cause constrains its possible effects (and more generally the probability function over its possible effects) is familiar enough. In a deterministic world, for example, Buridan's ass will starve, since eating either bale will break the initial symmetry. And in a stochastic world, the two bales have equal probability of being eaten.

More generally I assume, in cases where two possible events A and B are symmetric with respect to both the dynamical laws and the initial state, that A and B have the same chance of occurrence.

In the following argument, I will focus on just two types of symmetry that dynamical laws typically possess.

(i) Invariance under spatial and temporal translation

(ii) Locality

The first symmetry is just the familiar idea that the laws of physics are the same everywhere and at all times. The second says that what happens in one place x at time t depends directly on what happened just prior to t in the neighbourhood of x . There is no direct action at a distance, or across times. (You may not think of this second property as a symmetry, but it is in some sense at least.)

The argument below is made in the context of cellular automata, rather than dynamical systems with continuous space and time, for simplicity. I hope that the results will generalise fairly easily, however.

The conclusion of this argument is that, in a dynamical system with the two symmetries stated above, the only large structures that can have high salience are *regular*,

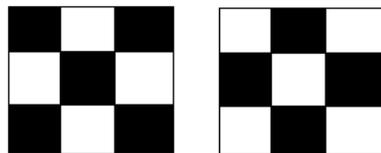
or *self-similar* ones. More precisely, I will show that a large object with high salience must be highly determined by its *local structure*. Let us therefore define this term.

4. LOCAL STRUCTURE

Suppose you are provided with a square grid of cells, 1000 cells wide and 1000 cells high, for a million cells in all. Each cell may be filled with either a white or a black counter. You're also provided with a black-and-white digital photograph, which has one million pixels in a 1000 by 1000 grid. You're given the task of placing counters into your grid to produce an exact copy of the "target" photograph. Simple enough? To make it more of a challenge, let's suppose that you can view the target only through a thin straw, which permits you to see only a 3 by 3 block of pixels at one time. Also, when you're looking at such a "local block", you can have no idea where in the target image it is.

These two constraints, of being able to see only local blocks, and being ignorant of their positions, may or may not greatly hamper one's ability to complete the task.

Figure 1



Suppose, for example, that the local blocks turn out to be of only two different kinds, as in Figure 1 above. In this case, the target is clearly one of two different things, so you are bound to complete the task in your first two attempts.

If, on the other hand, when you look through the straw you see all 512 possible kinds of local block, and with about the same frequency, then it's a lot more difficult, for the target image may be any one of a rather large set of possible states. One can then do no better than guess which one it is. You will complain that the task is practically impossible.

One way to describe the situation is in terms of “local structure”. Looking at the image through the straw tells you its local structure. We can define the local structure more precisely as a function from the 512 block types to their frequencies in the image. The difficulty of this task then depends on the extent to which the target image is determined by its local structure. In the first case, where there were only two block types, the local structure almost completely determined the image. In the second case, however, the image was largely undetermined by the local structure.

Using this notion of local structure we can define the *irregularity* of an image \mathbf{s} in terms of the number N of possible images that have the same local structure as \mathbf{s} . For reasons of convenience, I actually define the irregularity of \mathbf{s} as $\log N$.

Suppose the target is \mathbf{s} , and \mathbf{s}' is locally equivalent to \mathbf{s} . (I.e. \mathbf{s} and \mathbf{s}' have the same local structure.) Let $F^r \mathbf{s}$ be the event that you manage to produce \mathbf{s} among the first r attempts. We then see that, in the absence of additional information, so that you’re reduced to guessing the global structure of the target, $P(F^r \mathbf{s}) = P(F^r \mathbf{s}')$. It is possible to see this equality as a result of *symmetry* in your information, even though it is not a straightforward geometrical symmetry. Your information about the target image (i.e. the local structure of the target) is symmetric with respect to of the N -membered set of images with that local structure, in the sense that it does not allow you to single out any member of the set.

5. LOCAL DYNAMICS

Dynamical laws, as stated in Section 3, operate at the local level. Thus they are restricted in something like the way one is restricted by looking at the target through a straw. But there’s an important difference: Instead of looking at the *target* through a straw, the dynamical law looks at the *present state* of the system through a straw. To see how this works, let’s consider the image problem again.

Suppose that when you look at the target image through the straw you see all the 512 kinds of block, in equal frequency. You complain that the task cannot be done by any *clever* means, but only by sheer luck, (very) dogged persistence, or both. In response,

a new problem is set, where you are shown the image all at once, which turns out to be a portrait of Abraham Lincoln. But the catch is that you're now only allowed to look at *your own* grid through the straw, not knowing which block you're looking at. You decide the next colour of a given cell after examining just its present colour and those of the surrounding eight cells. (You're not allowed to take into account any knowledge of other parts of *your* grid, but you can use your knowledge of the target.)

For the sake of clarity it may help to present this new problem in a different way. For each time t your assistant looks at the state of your grid at t , and prepares a " t -sheet", which lists all the local blocks in the state at time t . Every *individual* block is listed, not just each *type* of block, so that there are exactly as many blocks as cells. (Each cell is, of course, at the centre of exactly one local block.) The blocks are however listed in random order, so that you have no information about the location of each block in the grid. You move through the t -sheet, making a decision about each block, whether to keep the central cell as it is or change it to the other colour. This decision is based entirely on its colour and those of the surrounding 8 blocks, not on any other blocks in the t -sheet. The decision may be either deterministic or probabilistic, however. The most general case therefore is that you have a set of 512 different "toggle probabilities", i.e. probabilities for toggling (changing) the central cell, based on the 512 different possible colour combinations of that cell and its surrounding cells. Your assistant takes all these toggle/keep decisions and uses them to update the state of the grid, and then provide you with a $(t+1)$ -sheet. Then you make a similar set of decisions about the blocks on the $(t+1)$ -sheet, and so on. In making these decisions you are exactly mimicking the work done by a (local and invariant) dynamical law of a cellular automaton.

In the new problem the difficulty has been shifted. Instead of having restricted information about the target image, you have even more tightly restricted information about the present state of the grid. (Your information is slightly less in the new problem than in the old, since in the new problem you never see the entire local structure of your grid. You just see one local block at a time.) How do the two problems compare in difficulty? In Appendix 1 the following answer is demonstrated.

Theorem 1 The new problem is at least as hard as the old one.

Theorem 1 is specifically used to convert results about the old problem into results about dynamical systems. In fact, some of the results in this paper apply primarily to the old problem. But then, using Theorem 1, we derive a corresponding result about the new problem, i.e. about dynamical systems.

6. SALIENCE

As mentioned in Section 2, the notion of *salience* is needed to express the fact that some objects tend to appear more quickly than others, in a given dynamical system, from a random initial state.

I define the salience of an object, with respect to a dynamical law, as follows. First we define the *r-salience* of an object \mathbf{s} :

Definition Let the proposition $F^r \mathbf{s}$ say that \mathbf{s} is among the first r objects that appear in the history. Then $Sal^r(\mathbf{s}) = P(F^r \mathbf{s})/r$.

Note that if the object \mathbf{s} tends to be produced fairly quickly, from most initial states, then its *r-salience* will be quite high for some (low) values of r . If \mathbf{s} is rarely produced, on the other hand, for almost all initial states, then its *r-salience* will be low for all r . This fact suggests that we define salience from *r-salience* in the following way.

Definition $Sal(\mathbf{s}) = \max_r \{Sal^r(\mathbf{s})\}$.

In other words, we define the *salience* of \mathbf{s} as its *maximum r-salience*, over all values of r . Thus, if an object tends to be produced quickly by the dynamics, so that its *r-salience* is quite high, for some small r , then its salience will also be quite high. An object that is unlikely to be produced in any short time will have low salience.

For convenience I also define the “dynamical complexity” of \mathbf{s} as the log of its salience.

Definition $Comp(\mathbf{s}) = -\log Sal(\mathbf{s})$.

Note that if \mathbf{s} and \mathbf{s}' are locally equivalent, then they are equally likely to be guessed, by a player who views the target through a straw, as in the old problem. In other words, for such a player: $P(F^r\mathbf{s}) = P(F^r\mathbf{s}')$. We then have the following theorem (see Appendix 2 for a proof).

Theorem 2 If \mathbf{s} is one of N objects in a locally-equivalent set $S = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N\}$, then $Sal(\mathbf{s}) \leq 1/N$.

Further, according to Theorem 1, which states that a local dynamical law is at least as severely restricted as such a player, we infer that $Sal(\mathbf{s}) \leq 1/N$ for a dynamical system as well.

It will be useful to consider the salience of an n -bit binary string in a couple of rather trivial dynamical systems. The first such system, which we'll call the *completely random system*, is one which each cell evolves completely at random, and independently of the other cells. It's as if the content of each cell at each time is determined by the outcome of a fair coin toss, with there being a separate toss for each cell at each time.

In considering the salience (and hence dynamical complexity) of an n -bit string here, the size of the system (i.e. the number of cells) is a complicating factor, so to begin with let's suppose that the system is a one-dimensional array of n cells. In this case, the salience of every string is the same, namely 2^{-n} , so that the complexity is n . Note that this result regards the target string (\mathbf{s} say) as distinct from its mirror image, $\mathbf{s}^{\bar{}}$. (I.e. $\mathbf{s}^{\bar{}}$ is just \mathbf{s} in reverse.) If we regard \mathbf{s} and $\mathbf{s}^{\bar{}}$ as identical, then the complexity of this object is $n - 1$.

This simple case also supposes that the system is *linear*, so that there are edge cells. This implies that each bit of \mathbf{s} has just one cell in which it can appear. Another possibility, however, is that the system is a *closed loop* of cells, so that there are no edges. In that case, the first bit of \mathbf{s} can appear in any of the system's n cells. There then are n products on each time step, so that \mathbf{s} is likely to appear much more quickly. The probability of \mathbf{s} being somewhere in the initial state, for example, is now $n.2^{-n}$. But since

there are n objects in that state, the n -salience is 2^{-n} . It is easily shown that the r -salience of \mathbf{s} is also 2^{-n} , for all r , so that the complexity of \mathbf{s} is still n .

What if the system is larger than n cells, however? In a larger system there is (one might say) more guessing going on, so that any given n -bit string (\mathbf{s} say) is likely to be guessed sooner. On the other hand, there are also more “products” at each time step, since each n -bit section of the state might be considered a product. What will the net effect on the salience of \mathbf{s} be, as the system size is increased? Actually there will be no change at all. Suppose there are m cells, for example, where $m > n$, then \mathbf{s} can appear in m different places in the system, so that there are m products on each time step. In that case the probability of \mathbf{s} in the initial state is $m \cdot 2^{-n}$, so that the m -salience (and indeed the r -salience) is still 2^{-n} .

The same situation obtains in two- and three-dimensional systems. But note that in a two-dimensional system we might allow the string to appear in a column, going up as well as down, as well as backwards in a row. Thus the salience of this set of four objects will be $4 \cdot 2^{-n}$, and its complexity $n - 2$. For large n this difference is rather trivial, however.

7. COMPLEXITY AND INFORMATION

In the previous section the “dynamical complexity”, or *Comp*, is defined in a merely arithmetical way from the salience of \mathbf{s} . The reader is therefore left in the dark as to what (if anything) dynamical complexity means. In this section I shall briefly explain why I regard it as measuring the “information content” of an object.

The term “information content” has become rather commonplace, although it has a number of different meanings. What does it mean here? To answer this it is best to begin with the epistemic context, of a thinker who has a particular *epistemic state*, or state of knowledge, at a given time. In that context one can define $Inf_K(A)$, the information content of a proposition A , relative to the epistemic state K , as $-\log_2 P_K(A)$, where $P_K(A)$ is the epistemic (or evidential) probability of A within K . Note that if A is believed with certainty in K then $P_K(A) = 1$ and hence $Inf_K(A) = 0$. Thus we see that *Inf*

measures the amount of “new” information in A , over and above what is already known. To understand the meaning of Inf more concretely, it helps to consider a decomposition of A into a conjunction of propositions that are mutually independent, and which each have probability $1/2$. If $Inf_K(A) = n$, then A will decompose into n such propositions. If we call such a proposition a “bit”, then it’s very natural to say that A contains n bits of information.

One may wonder what the value of introducing Inf is, however, since it allows one to say only things that can already be expressed in terms of P . In fact some relations, while they can be expressed perfectly precisely in terms of P , seem more natural and intuitive when expressed using Inf . This in turn renders some important logical facts easier to see. Consider, for example, the conjunction rule for epistemic probabilities:¹

$$P_K(A \wedge B) = P_K(B).P_K(A | B)$$

This same relation, in terms of information, is:

$$Inf_K(A \wedge B) = Inf_K(B) + Inf_K(A | B)$$

This is very intuitive – much more intuitive than the conjunction rule for probabilities. For consider that one way to learn the conjunction $(A \wedge B)$ is to learn B , and then learn A . The information gained in the first step is $Inf_K(B)$, of course, putting one in the new state $K+B$. Then, when one learns A , the extra information gained is $Inf_{K+B}(A)$, i.e. $Inf_K(A | B)$.

Since $Inf_K(A \wedge B)$ can also be expressed as $Inf_K(A) + Inf_K(B | A)$, we see here a kind of “path independence” involved in learning (this can be proved generally). The quantity of information gained along a “learning path” of expanding epistemic states depends only on the end points, and is independent of the path taken.

¹ Note that $P_K(A | B)$ is defined as $P_{K+B}(A)$, where $K+B$ is the epistemic state K expanded by adding full belief in the proposition B . In a similar way one can define $Inf_K(A | B)$ as $Inf_{K+B}(A)$. Also, $A \wedge B$ means “ A and B ”, i.e. the weakest proposition that entails A and entails B .

This is vaguely reminiscent of conservative force fields, where the work required to move a particle from point A to point B is independent of the path taken. One might therefore wonder whether the “information” thus defined, like energy, is some sort of conserved quantity. In fact there are some “conservation theorems” of this sort, involving *Inf*, that make the analogy useful. (Note that there are also many important disanalogies between energy and information.)

To understand these information conservation theorems, it is essential to understand that epistemic probability is based on the idea of an ideal agent that is *logically omniscient*. This means that the agent believes all logical truths (such as tautologies) with certainty. His set of certain beliefs is also deductively closed, so that it contains all the logical consequences of its members. For such an agent, it is easy to see that no new information can be obtained by thought alone. Some sort of external information input is needed.

Gregory Chaitin (1982) said: “I would like to be able to say that if one has ten pounds of axioms and a twenty-pound theorem, then that theorem cannot be derived from those axioms.” This looks like a kind of information-conservation principle, and indeed it is a consequence of the theorem below.

Suppose that K represents one’s initial, “background”, knowledge, and that B is some theorem one would like to be able to prove. But B isn’t certain in K , having instead some non-zero information content $Inf_K(B)$. So let us add some new proposition A to K , in the hope that B will be provable (i.e. certain) in $K+A$. It is trivial to show that, for this to be possible, $P_K(A) \leq P_K(B)$, so that $Inf_K(A) \geq Inf_K(B)$. In other words, if the theorem B weighs twenty pounds (relative to K) then the axioms A must weigh at least twenty pounds as well.

The use of the term “conservation” to describe such results is not ideal, as it might suggest that the weight of the axioms always *equals* that of the theorems, which is obviously not the case. One can certainly prove a “light” theorem from a “heavy” set of axioms! What is ruled out is an *increase* in weight, from premises to conclusion. This result might therefore be better described as a non-amplification theorem. This is rather an awkward term, however, so I shall continue to call it a conservation theorem.

More generally we have the following (also trivial) conservation theorem.

Theorem If learning A reduces the information content of B by r bits, then the information content of A is at least r .

I.e. If $Inf_K(B | A) = Inf_K(B) - r$, then $Inf_K(A) \geq r$.

Similar conservation laws apply to the notion of dynamical complexity defined in Section 6. Before we examine these, I will explain why *Comp* can actually be thought of as a measure of information content.

Suppose you want to have the combination to a safe, say a 20-bit string \mathbf{s} . A certain machine \mathbf{M} will output the correct combination, with certainty, when a button is pushed. We might say that the machine “contains” this information. Intuitively we might say that $Inf_{\mathbf{M}}(\mathbf{s}) = 0$ in that case. But what if it spits out *two* strings, one of which is guaranteed to be correct? This is something a person might do, if they know only 19 of the 20 bits – they can then write down the two possible codes. It makes sense to say that $Inf(\mathbf{s}) = 1$, for such a person, as the true code supplies them with one extra bit of information.

What if \mathbf{M} outputs one string, which has a 50% chance of being correct? This is very similar to the previous case, as there the person could select one of the two strings at random, giving them a 50% chance of getting the right one. This suggests that the information content of the code \mathbf{s} , relative to the machine \mathbf{M} , can be defined as follows. We run the machine, so that it produces as many strings as it wants to and then stops. These outputs may be either deterministic or random, with whatever probabilities. Then we select one of the outputs at random (each having the same probability of selection). Let the random variable \mathbf{X} be the selected object. Then we define the salience of \mathbf{s} for the machine \mathbf{M} as the probability that $\mathbf{X} = \mathbf{s}$, i.e. the probability that \mathbf{s} is *both* produced by \mathbf{M} *and* selected from \mathbf{M} 's outputs. We define the information content of the string as minus the (base 2) logarithm of the salience. In other words:

$$Inf_{\mathbf{M}}(\mathbf{s}) = -\log P_{\mathbf{M}}(\mathbf{X} = \mathbf{s}).$$

This definition seems right in the case where \mathbf{M} produces a certain number of binary strings and then stops. But what if it keeps going, *ad infinitum*? Suppose it generates, eventually, every possible 20-bit string? In that case, we might still want to say that some strings are more salient than others, or that some strings have lower information content than others. If the machine always produces the code \mathbf{s} as its first output, for example, then it seems that it “contains” \mathbf{s} almost as much as if it produces \mathbf{s} as its sole output.

The best way I can see to satisfy this intuition is to define the r -salience of \mathbf{s} as the salience of \mathbf{s} for a slightly modified machine, one that is forced to halt (unplugged?) after producing r strings. One of these r objects can then be selected at random as before. The salience of \mathbf{s} is then the maximum of the set of r -saliences. Using this definition it’s clear that $Sal(\mathbf{s}) = 1$ and $Inf(\mathbf{s}) = 0$ in the case where \mathbf{s} is always the first output. I regard these numbers as putting an upper bound on the salience, i.e. a lower bound on the information content.

Finally, suppose we replace the machine with a dynamical system. It produces objects, all over the place, all the time. But we can use the same technique of defining the r -salience of an object by allowing the system to evolve freely, from a random initial state, and produce r objects before being halted. Then we select one of those objects at random. This approach yields the definitions of the previous section.

8. COMPLEXITY CONSERVATION THEOREMS

The idea of starting the system in a random initial state is that no initial information is provided. In general, however, some initial states will be more likely than others, and many states will be impossible. How will this affect things? In this section we will examine one kind of constraint on the initial state, where one restricts the initial state of the system to a subset of the full state space. It is easy to see that such restrictions can increase (as well as decrease) the salience of a given object, but a conservation theorem applies here. It is shown below that for a restriction of the initial state to reduce the complexity of \mathbf{s} by v bits, the restriction itself must contain at least v bits of information.

In cases where the initial state is restricted by a conditionalisation of the chance function, we can regard anything that emerges as a product of *conditional* self organisation. It is not, as it were, *absolute* self organisation, since the system had some outside help in getting started. But it is self organisation from that point onward.

In order to investigate conditional self organisation, we introduce the notion of a *program*, as follows.

Definition A *program* Π is a restriction of the initial state to a subset of the state space. (The subset is also called Π .)

Since the initial state is set at random, each program Π has a probability $P(\Pi)$, which in the case of a finite state space is simply the proportion of states that are in Π . The more restrictive the program, the lower its probability is. We also define the *length* of a program as follows:

Definition The *length* of Π , written $|\Pi|$, is $-\log P(\Pi)$.

We can then say that if Π reduces the complexity of \mathbf{s} by ν bits, then $|\Pi| \geq \nu$. Rather than prove this conservation theorem, however, it's more convenient to combine it with another that relates the complexity of an object with the time needed to produce it, with reasonable probability. Suppose, for example, that $Comp(\mathbf{s}) = n$, within a given system. How long will the system take to produce \mathbf{s} ?

In the case of a deterministic system it is easy to see that the system must produce \mathbf{s} in a set of no fewer than 2^n objects. For suppose the negation, that \mathbf{s} appeared reliably in a smaller set of 2^m objects, say, where $m < n$. In that case \mathbf{s} could be found by letting the system produce 2^m objects, and selecting one of them at random. Such a procedure would yield \mathbf{s} with probability 2^{-m} , giving a complexity of m for \mathbf{s} , contrary to the assumption that $Comp(\mathbf{s}) = n$. The conservation theorem below includes this result (see Appendix 2 for a proof). For reasons that will become clear, I call it the *random equivalence*, or RE theorem.

RE Theorem Suppose $Comp(\mathbf{s}) = n$, and $|\Pi^*| < n$. Then $P(F^{r^*}\mathbf{s} \mid \Pi^*) \leq r^* \cdot 2^{|\Pi^*| - n}$.

The RE theorem says that the probability of the system producing an object of complexity n among the first r^* products increases in proportion to r^* , but decreases exponentially with the difference between n and the length of the program. (It should be noted that this theorem is independent of Theorem 1.)

To see what this theorem really means, consider a case where a very intelligent person is asked to reproduce a hidden 20-bit string that has previously been produced by some purely random process, such as tossing a fair coin. The target string \mathbf{s} has epistemic probability 2^{-20} , of course, since all possible strings are equally likely. Then the information content of \mathbf{s} is 20 bits. What can this person do?

If he is allowed only one attempt, then he can do no better than flip 20 coins himself, submitting their output as his answer. Note that this corresponds, in the above theorem, to the case where $r^* = 1$ and $|\Pi^*| = 0$.

If the person is allowed multiple attempts at producing \mathbf{s} (suppose he's allowed r^* attempts) then what is the best strategy? There is actually nothing better than making a series of independent random guesses, being sure of course not to make the same guess twice. Using this method, the probability of success is of course exactly $r^* \cdot 2^{-n}$. Finally, we can consider the case where the person is allowed r^* attempts at the code, and is provided with additional relevant information in the form of the proposition Π , whose information content is then $|\Pi|$. By the conservation theorem of Section 7, we see that this can increase the probability of producing \mathbf{s} by a factor of $2^{|\Pi|}$ at most, giving a probability of $r^* \cdot 2^{|\Pi^*| - n}$.

In other words, the theorem says that *producing a given object with complexity n in a dynamical system is no easier than producing a given n -bit string in a completely random system*. This is a perfectly general result, applying to any system whatsoever. One obvious consequence is that an object whose dynamical complexity in a particular system is very large (a million bits, say) cannot be produced in that system in any reasonable length of time. One might as well wait for monkeys with typewriters to produce *Hamlet*.

9. COMPLEXITY AND IRREGULARITY

In Section 4 we defined the irregularity of a state s as $\log N$, where N is the number of states that are locally equivalent to s . In Section 6 we saw that the salience of such a state s is no greater than $1/N$, so that the dynamical complexity of an object always exceeds its irregularity. Then, in Section 8 we saw that states with very low salience, such as 2^{-n} (where n is reasonably large) effectively cannot be produced by a dynamical system. The question that remains concerns whether any of the objects we see around us have such low salience. In other words: How great is the value of N for real objects?

For a very simple case, consider a binary string s of n bits that is maximally irregular. In other words, the string contains all eight kinds of “local triple”, i.e. 000, 001, 010, 011, 100, 101, 110 and 111, in equal frequency. How irregular is s ?

I don't yet have a strict proof here, but given a very plausible assumption it is easy to show that the dynamical complexity of an irregular string is roughly the same as its length. (See Appendix 4 for a relative proof.)

Conjecture If s is a maximally-irregular string of length n , where n is of the order one million or greater, then the irregularity (and hence dynamical complexity) of s is at least $0.999n$.

This conjecture entails that long, irregular strings have very low salience. An irregular string of a billion bits, for example, would have a dynamical complexity of virtually one billion bits. Thus, using the RE theorem, producing a particular billion-bit string of this kind is no easier in a dynamical system than in a completely random system. In other words, it is (effectively) impossible. This impossibility is quite independent of the actual dynamical laws, but depends only on the general features of locality and invariance in the dynamical laws. We thus have the following theorem:

Limitative Theorem A large, maximally irregular object cannot appear by self organisation in any dynamical system whose laws are local and invariant.

Proof:

Suppose that an object s is maximally irregular, and of size n bits. Then its irregularity is approximately n , by the above (practically certain) conjecture. Using Theorem 2, the dynamical complexity of s is at least (approximately) n as well. Then, using the RE theorem, to produce s with any reasonable probability requires that a total of about 2^n objects are produced. If n is large, say 10^6 or greater, then 2^n is ridiculous. ■

10. DID LIFE EMERGE SPONTANEOUSLY?

The appearance of the first self-replicating molecule (or system of molecules) may or may not have been by self-organisation. Some authors, Richard Dawkins for example, have appealed to the vast size of the universe to help explain this event. Having supposed that the probability of a self-replicator appearing on any single planet might be around 10^{-9} , Dawkins continues:

“Yet, if we assume, as we are perfectly entitled to do for the sake of argument, that life has originated only once in the universe, it follows that we are *allowed* to postulate a very large amount of luck in a theory, because there are so many planets in the universe where life *could* have originated. If, as one estimate has it, there are 100 billion billion planets, this is 100 billion times greater than even the very low [spontaneous generation probability] that we postulated.”

Other authors disagree with Dawkins here, however, claiming that the first self replicator self organised. Manfred Eigen seems to have held such a view. I shall steer clear of this issue, however, and assume only that the emergence of life *after a self-replicator exists* was by self-organisation. This view is very widely held. Dawkins, for example, expresses it as follows:

“My personal feeling is that, once cumulative selection has got itself properly started, we need to postulate only a relatively small amount of luck in the subsequent evolution of life and intelligence. Cumulative selection, once it

has begun, seems to me powerful enough to make the evolution of intelligence probable, if not inevitable.” (Dawkins 1986: 146)

Note that, by appealing neither to large amounts of luck, nor enormous times, nor external help, Dawkins is claiming that life self organised (in my sense, from Section 1).

I am aware that it is unusual to describe all of biological evolution as self-organisation (SO). It is more common to *contrast* SO with selection, identifying some biological structures as due to selection, and others to self-organisation, and see these as complementary processes. Blazis (2002) writes, for example:

The consensus of the group was that natural selection and self-organization are complementary mechanisms. Cole (this volume) argues that biological self-organized systems are expressions of the hierarchical character of biological systems and are, therefore, both the products of, and subject to, natural selection. However, some self-organized patterns, for example, the wave fronts of migrating herds, are not affected by natural selection because, he suggests, there is no obvious genetic connection between the global behavior (the wave front) and the actions of individual animals.

The term SO is applied, it seems, only to cases where the emergence of the structure is not controlled by the genome. Despite this usage, however, it is very important to see that biological evolution, proceeding by the standard mechanisms, satisfies the definition of SO in Section 1. Standard biological evolution is, therefore, a special case of conditional self-organisation (i.e. conditional on a self-replicator in the initial state).

The first self-replicator on earth must have appeared (very roughly) 4 billion years ago. Since that time, an enormous profusion of complex living organisms has appeared, as a result of the laws of physics operating on that initial state. Now, in view of the size of these organisms, 4 billion years is very much shorter than the time required to assemble such objects by pure chance. This is why I say that the emergence of life, given the first self-replicator, was by self-organisation. The three criteria from Section 1 are met.

Is this fact in conflict with the Limitative Theorem above, that large, irregular objects cannot emerge by SO? It may seem not. For, while living organisms are very large, containing trillions of atoms, they are far from maximally irregular, and the

Limitative Theorem applies only to maximally-irregular objects. However, the following two considerations should be born in mind.

First, while the Limitative Theorem applies only to maximally-irregular objects, there is unlikely to be too much difference in salience between maximally and highly irregular objects. An object has to be highly regular before its global structure becomes highly constrained by its local structure. A more general result, therefore, would surely find a similar situation with all irregular objects, not just maximally-irregular ones.

Second, one can apply the Limitative Theorem to the genome instead of the organism itself. Genomes are very small compared to phenotypes, of course, but they are still very large, and (I believe) much more irregular. The shortest bacterial genomes, for example, contain about half a million base pairs, or roughly a million bits. If these genomes are indeed highly irregular, as I suppose, then their production by SO is ruled out by the Limitative Theorem.

At this point we should recall, however, that I am assuming the existence of a self-replicating entity in the initial state. What difference does this make? The following theorem shows that, while the existence of a self-replicator might well reduce the dynamical complexity of life, such a reduction cannot exceed the complexity of the self-replicator itself. Hence, since the original self-replicator is assumed to be small, and relatively simple, its presence makes very little difference.

First we require some definitions.

Definition $Sal(s | s') = Sal(s)$ given that the initial state contains the object s' . In other words, one can use the definition of $Sal(s)$ above, but the probability function used is generated from the dynamics by applying a random initial state and then conditionalising on the presence of s' in that state.

Conditional dynamical complexity is then defined from conditional salience:

Definition $Comp(s | s') = -\log Sal(s | s')$

Theorem 3 Constraining the initial state of a dynamical system to include some object s' can reduce the complexity of other objects by no more than $\text{Comp}(s')$.

Proof: We shall first prove that $\text{Sal}(s) \geq \text{Sal}(s') \cdot \text{Sal}(s | s')$. Suppose that the value of r that maximises $\text{Sal}^r(s')$ is r_1 , and the value of r that maximises $\text{Sal}^r(s | s')$ is r_2 . Then one may try to generate s from a random initial state using the following method. One allows the system to evolve for some period of time, producing r_1 objects. One of these objects is then selected at random. The probability of the selected object being s' is exactly $\text{Sal}(s')$, according to the definition of salience. If s' is selected then the system is prepared in a random state containing s' , and allowed to evolve again to produce another r_2 objects. One of these r_2 objects is selected at random. Given that the first stage succeeds, the chance of selecting s at the second stage is exactly $\text{Sal}(s | s')$. The overall probability of getting s at the second stage is then $\text{Sal}(s') \cdot \text{Sal}(s | s')$. This is clearly less than $\text{Sal}(s)$, since $\text{Sal}(s)$ involves selecting just s in a history that begins with a random initial state, whereas here we are selecting s' as well as s in such a history. I.e. $\text{Sal}(s) \geq \text{Sal}(s') \cdot \text{Sal}(s | s')$. Using the definition $\text{Comp}(s) = -\log \text{Sal}(s)$, we immediately obtain the result that $\text{Comp}(s | s') \geq \text{Comp}(s) - \text{Comp}(s')$, as required. ■

We can roughly gauge the effect of introducing a self-replicator into the initial state by using a few very approximate numbers. Suppose we wish to make a relatively simple organism, such as a bacterium, whose complexity is about 10^6 bits. The complexity of the first replicator must be much less than this, for its appearance not to be a miracle. On Dawkins' view, for example, there might be around 10^{20} planets, which "pays for" a chance of about 10^{-20} per planet, or about 60 bits. Within each planet there might be many opportunities for the self-replicator to appear, over a billion years or more, which pays for perhaps another few tens of bits, or even as many as 100. In any case, the complexity must surely be below 1000 bits. But subtracting even 1000 bits from one million makes almost no difference. Hence the Limitative Theorem cannot be circumvented by imposing a self replicator in the initial state.

11. DOES THIS RESULT IGNORE NATURAL SELECTION?

Dawkins' view quoted above, concerning the high probability of intelligent life once a self replicator exists, is probably an extreme one among biologists. Nevertheless, it is very commonly supposed that the processes of genetic mutation and natural selection allow complexity to emerge far more quickly than pure chance would allow. This idea has been at the heart of the evolutionary thinking since Darwin and Wallace. My argument has, for this reason, been suspected of somehow ignoring natural selection, or assuming its absence. After all, my result finds very little difference between purely random processes and general dynamical systems in the time required to produce irregular objects. But surely we know that natural selection can produce complex objects much more quickly than pure chance can? It follows that I *must* somehow be assuming the absence of natural selection.

The short answer to this worry is that my Limitative Theorem cannot possibly make any assumptions about biological processes of any kind. It cannot assume that such processes are absent, since it does not engage with biological matters in any way! The argument is entirely at the level of physics, being based on symmetries in the dynamical laws. There cannot of course be any contradiction between biological and physical facts – any biological claim that violated the conservation of energy, or the second law of thermodynamics, for example, would be false. (If someone is convinced that some assumption *is* being made that rules out natural selection, then they are most welcome to identify it.)

While this short answer is correct, it sheds no light on what is going on. So let us examine the general idea of producing complex objects gradually, through a series of small modifications or changes. Theorem 3 perhaps suggests that a gradual approach might make a big difference to the time required. Consider, for example, an object s of complexity n relative to the first self-replicator. According to the RE theorem, it will require at least 2^n modifications to make s with high probability. But now suppose we consider an intermediate object s' , whose complexity is $n/2$ relative to the first self replicator. According to Theorem 3, the complexity of s relative to s' *may* be as little as $n/2$ as well. In that case, the production of s' from the random initial state might take

only about $2^{n/2}$ changes, and the same for the transition from s' to s . Hence the total could be a mere $2 \times 2^{n/2}$ changes, i.e. $2^{n/2+1}$, which is a tiny fraction of 2^n . It appears that the insertion of even just one intermediate stage has drastically reduced the time required to produce the object.

The “gradualist” argument of the previous paragraph must be a fallacy, since the proof of the Limitative Theorem is very general, and includes such cases as the one above. But what is wrong with the argument?

In order to investigate this, it will be helpful to consider a particular dynamical law, and some intermediate objects. Consider, for example, a 20-bit counter, that begins with random values, and then on each time step adds one to the number showing. (After 11111...1 it goes back to 00000...0.) To obtain the target number s will require, on average, about 2^{20} (about one million) iterations. Now suppose that the counter is, at some point in its evolution, showing 10 correct bits among the 20. Does this entail that s will be obtained in about 2^{10} , i.e. about 1000, further iterations? It does not, because it all depends on which 10 bits are correct! If the first ten are correct, then s is indeed close at hand. But if the last 10 are correct, then this means nothing, as those correct bits must be lost before the incorrect bits can change. We are still about a million steps away.

In this example we see that there is *some* intermediate state, namely where the first ten bits are correct, from where the goal s is very close. And this intermediate state has 10 bits of complexity, since it has salience 2^{-10} . (There is, after all, a probability 2^{-10} of getting it as the initial state.) Yet, interestingly, this state is very unlikely to be obtained in the first 1000 time steps. Also, there is another 10-bits-complex intermediate state that *is* quickly produced from the initial state. (Namely, where the last ten bits are correct.) But this is unlikely to evolve to s in less than about a million steps. So each intermediate object is far from either the initial state or from s .

At this point we should recall that the RE theorem is an inequality. It takes *at least* 2^n objects to have a good chance of producing an object with complexity n bits. Moreover, we know from modal logic that $\diamond A \ \& \ \diamond B$ does not entail $\diamond(A \ \& \ B)$, i.e. the possibility of A and B together isn't a consequence of the individual possibility of A together with the possibility of B . In this simple example we find that, while a 10-bit object can be close to the initial state, and can be close to the desired 20-bit object, it

cannot be close to *both* of them. Thus, through this counter-example, we have identified a serious fallacy in the gradualist argument. The individual possibility of each small step occurring in a short time does not entail the possibility of the entire sequence of steps each occurring in a short time.

We should therefore be wary of any general argument that seeks to show that a complex object can be produced gradually, by a cumulative process, far more rapidly than the Limitative Theorem allows.

12. CONCLUSION

I have argued that there is an important limitation on the kinds of object that can appear spontaneously in a dynamical system. Such systems, with laws that operate locally and invariantly across space and time, are able to control only the local structure of the state. The state as a whole is therefore uncontrolled, except insofar as it is constrained by the local structure. This led us to the Limitative Theorem, which says that an irregular object, i.e. one that is largely undetermined by its local structure, cannot easily be produced in a dynamical system. Indeed, it was shown that its production is no easier than the appearance of an object of very similar size in a purely random system.

This result, while relevant to biology, does not of course contradict the theory of evolution in its most general form, i.e. that life evolved through a process of descent with modification. This is just as well, since the historical process of phylogeny is very well supported by the evidence. Nevertheless, the Limitative Theorem does suggest that the currently recognised processes driving evolutionary change are incomplete.

APPENDICES

1. Proof of Theorem 1

First consider a case where one can, in some manner, choose at least the local structure of the next state of the grid. In that case one would, at each iteration, choose a local

structure that equals the one of the target. (Otherwise one is *bound* to fail!) It is also clear, on the other hand, that one can do no better than that. Hence in such a case, one's task would be exactly as hard as the old problem.

In the new problem, one has strictly less control over the grid than this, since one cannot (directly, in one step) choose even the grid's local structure. Hence the new problem is at least as hard as the old. ■

Appendix 2: Proof of Theorem 2:

(Note that this result applies to the old problem.) We previously defined $Sal^r(\mathbf{s}) = P(F^r \mathbf{s})/r$. Now let $I(F^r \mathbf{s}_i)$ be the indicator function for the proposition $F^r \mathbf{s}_i$, so that $I(F^r \mathbf{s}_i) = 1$ when \mathbf{s}_i is among the first r products, and $I(F^r \mathbf{s}_i) = 0$ otherwise. Note that, for all i , $P(F^r \mathbf{s}_i) = E[I(F^r \mathbf{s}_i)]$, where $E[\]$ is the expectation operator.

Since there are no more than r members of S among the first r objects, we see that:

$$\sum_{i=1}^N I(F^r \mathbf{s}_i) \leq r.$$

And also:

$$E \left[\sum_{i=1}^N I(F^r \mathbf{s}_i) \right] \leq r.$$

Then, since the expectation operator is linear, it follows that:

$$\begin{aligned} \sum_{i=1}^N E[I(F^r \mathbf{s}_i)] &\leq r \\ \Rightarrow \sum_{i=1}^N P(F^r \mathbf{s}_i) &\leq r. \end{aligned}$$

But now, since (for all i) $P(F^r \mathbf{s}_i) = P(F^r \mathbf{s})$, in the old problem, it follows that $P(F^r \mathbf{s}) \leq r/N$. Hence, for every r , the r -salience of \mathbf{s} is no greater than $1/N$. From this it follows that $Sal(\mathbf{s}) \leq 1/N$. ■

Appendix 3: Proof of the RE Theorem

First we prove this useful lemma.

Basic Lemma $Comp(\mathbf{s}) = \min_{r, \Pi} \{\log r + |\Pi| - \log P(F^r \mathbf{s} | \Pi)\}.$

Proof of Basic Lemma:

Let the O_i^r be all the possible output sets of length r . Then

$$P(F^r \mathbf{s}) = \sum_i P(F^r \mathbf{s} | O_i^r) P(O_i^r).$$

Now $P(F^r \mathbf{s} | O_i^r) = 1$ if $\mathbf{s} \in O_i^r$, and is 0 otherwise. Thus:

$$P(F^r \mathbf{s}) = \sum_{\mathbf{s} \in O_i^r} P(O_i^r); \quad \text{Also } P(F^r \mathbf{s} | \Pi) = \sum_{\mathbf{s} \in O_i^r} P(O_i^r | \Pi).$$

Further, $P(O_i^r) = P(O_i^r | \Pi)P(\Pi) + P(O_i^r | \neg\Pi)P(\neg\Pi)$, so $P(O_i^r) \geq P(O_i^r | \Pi)P(\Pi)$. But, if

Π is the entire state space, then $P(O_i^r) = P(O_i^r | \Pi)P(\Pi)$. Hence

$$P(O_i^r) = \max_{\Pi} \{P(O_i^r | \Pi)P(\Pi)\}.$$

Substituting this in the previous equation gives:

$$P(F^r \mathbf{s}) = \sum_{\mathbf{s} \in O_i^r} \max_{\Pi} \{P(O_i^r | \Pi)P(\Pi)\}.$$

$$\begin{aligned} \text{Thus } Sal(\mathbf{s}) &= \max_r \left\{ \frac{1}{r} \sum_{s \in O'_i} \max_{\Pi} \{P(O'_i | \Pi) P(\Pi)\} \right\} \\ &= \max_r \left\{ \max_{\Pi} \left\{ \frac{P(\Pi)}{r} \sum_{s \in O'_i} P(O'_i | \Pi) \right\} \right\}. \end{aligned}$$

$$\text{But } \sum_{s \in O'_i} P(O'_i | \Pi) = P(F^r \mathbf{s} | \Pi),$$

$$\begin{aligned} \text{So } Sal(\mathbf{s}) &= \max_r \left\{ \max_{\Pi} \left\{ \frac{P(\Pi)}{r} P(F^r \mathbf{s} | \Pi) \right\} \right\} \\ &= \max_{r, \Pi} \left\{ \frac{P(\Pi) P(F^r \mathbf{s} | \Pi)}{r} \right\}. \end{aligned}$$

$$\begin{aligned} \text{Then } Comp(\mathbf{s}) &= -\log \left[\max_{r, \Pi} \left\{ \frac{P(\Pi) P(F^r \mathbf{s} | \Pi)}{r} \right\} \right] \\ &= \min_{r, \Pi} \left\{ -\log \left(\frac{P(\Pi) P(F^r \mathbf{s} | \Pi)}{r} \right) \right\} \\ &= \min_{r, \Pi} \{ \log r - \log P(\Pi) - \log P(F^r \mathbf{s} | \Pi) \}. \end{aligned}$$

Then, putting $|\Pi| = -\log P(\Pi)$, we get:

$$Comp(\mathbf{s}) = \min_{r, \Pi} \{ \log r + |\Pi| - \log P(F^r \mathbf{s} | \Pi) \}. \blacksquare$$

Proof of the RE Theorem:

From the basic lemma, $n = \min_{r, \Pi} \{ \log r + |\Pi| - \log P(F^r \mathbf{s} | \Pi) \}$. Then consider some particular program Π^* and some value r^* . It is then clear that:

$$n \leq \log r^* + |\Pi^*| - \log P(F^{r^*} \mathbf{s} | \Pi^*),$$

And therefore $P(Fr^*s | \Pi^*) \leq r^*.2^{|\Pi^*|-n}$. ■

4. Counting Irregular States

To get a rough estimate on the number of (maximally) irregular strings of n bits we first define the “1-triple form” of a binary sequence. Consider, for example the 24-bit state 000001010011100101110111. We can break this into three-bit strings, or triples, as follows:

000 001 010 011 100 101 110 111

There are of course 8 possible triples, which we can call 0 (000), 1 (001), etc. up to 7 (111) in the obvious way. We then obtain the 1-triple form of this sequence as:

01234567

I call this the 1-triple form because the first triple begins on bit #1. We could begin on bit #2, and get the 2-triple form, namely:

02471356, i.e. 0 000 010 100 111 001 011 101 11

(Note that I am treating the sequence as a closed loop here.)

In a similar way, the 3-triple form is:

05162734, i.e. 00 000 101 001 110 010 111 011 1.

Now: an irregular state is one where each triple occurs with frequency $1/8$. This doesn't require, of course, that the triple has frequency $1/8$ in the 1-triple form, 2-triple form and 3-triple form individually, but only that it has frequency $1/8$ overall. Nevertheless, if the triple *does* have frequency $1/8$ in each of those forms, then it will have frequency $1/8$

overall. Note that, in this contrived example, each triple form is irregular, so that the whole state is irregular as well.

But why bother with these triple forms? It's because it's easy to calculate the number of n -bit sequences that are irregular *in the 1-triple form* (and similarly for both of the other triple forms, of course). This will give us a ballpark estimate of the number of irregular sequences, I think.

Suppose you have a bag of $n/3$ triples, containing $n/24$ of each type of triple (n is a multiple of 24). By arranging these into a sequence, you're sure to generate a state of n bits that is irregular in its 1-triple form, and moreover you can generate every such (1-triple irregular) sequence this way. So how many ways are there to arrange this bag of triples? Let the number of arrangements be N . We then have:

$$N = \frac{\left(\frac{n}{3}\right)!}{\left(\frac{n}{24}\right)!^8}$$

Applying Stirling's approximation to the factorial, namely:

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

We obtain:

$$N \approx 2^n \frac{\sqrt{8}}{\left(\frac{\pi n}{12}\right)^{\frac{7}{2}}}$$

To get an idea of how big a fraction of 2^n this is, let's consider $\log N$, i.e.

$$\log_2 N \approx n + \frac{3}{2} - \frac{7}{2} \left(\log \left(\frac{\pi n}{12} \right) \right)$$

I am interested in cases where n is in the rough interval from one million to one billion.

Let us plug in these values for n .

For $n = 10^6$, $\log N = 999,939$, approx.

For $n = 10^9$, $\log N = 999,999,904$, approx.

In other words, while the 1-triple irregular states are a tiny subset of the whole (being tens of orders of magnitude smaller) *on the logarithm scale the sizes are roughly equal*, to a tiny fraction of 1%.

I think that the number of states that are 1-triple irregular is a very rough estimate of the number of irregular states, but I'd guess that it's an overestimate. To get a (fairly firm but not rock solid) *lower bound* on the number of irregular states, I think we can take the *cube* of the proportion of this set in the total.

To see this, consider the fact that:

(i) If (but not only if) a state is irregular in all three triple forms, then it is irregular.

Also,

(ii) I think we can assume that the proportion of 2-triple irregulars among the 1-triple irregulars is at least as great as the proportion of 2-triple irregulars among the total. And similarly, the proportion of 3-triple irregulars among those that are irregular in both the 1-triple form and the 2-triple form is at least as great as the proportion of 3-triple irregulars among the total.

Given (i) and (ii), the product of these three (equal) proportions will be at least as great as the proportion of the intersection of the three sets.

In this way we obtain the lower bound N' for the number of irregular states of length n .

$$N' \approx 2^n \frac{8^{\frac{3}{2}}}{\left(\frac{\pi n}{12}\right)^{\frac{21}{2}}}$$

This gives similar results to the previous estimate. We now have that $\log N'$ is roughly $n - 21/2 \log n$, so that for $n = 1,000,000$, $\log N'$ is roughly 999791.

Bibliography

- Bennett, C. (1996) "Logical depth and physical complexity", unpublished manuscript.
- Blazis, D. E. J. (2002) "Introduction", *Biological Bulletin*, Vol. 202, No. 3 (Jun., 2002), pp. 245-246
- Chaitin, G. (1975) "A theory of program size formally identical to information theory", *Journal of Assoc. Comput. Mach.* 22, 329-340.
- (1982) "Gödel's Theorem and Information", *International Journal of Theoretical Physics* 21, pp. 941-954
- Darwin, C. (1859) *On the Origin of Species by Natural Selection*, London: John Murray.
- Dawkins, R. (1986) *The Blind Watchmaker*, Reprinted by Penguin, 1988.
- Gärdenfors, P. (1988) *Knowledge in Flux*, Cambridge, Mass: MIT Press.
- Hinegardner, R. and Engleberg, J. (1983) "Biological complexity", *Journal of Theoretical Biology* 104, 7-20.
- Howson, C. and Urbach, P. (1989) *Scientific Reasoning: The Bayesian Approach*, 2nd ed., La Salle: Open Court, 1993.
- Kampis, G. and Csányi, V. (1987) "Notes on order and complexity", *Journal of Theoretical Biology* 124, 111-21.
- Kolmogorov, A. N. (1968) "Logical basis for information theory and probability theory", *IEEE Transactions on Information Theory* IT-14, No. 5, 662-664.
- Lewis, D. (1980) "A subjectivist's guide to objective chance", Reprinted in Lewis (1986b), 83-113.
- (1986b) *Philosophical Papers Volume II*, New York: Oxford University Press.
- Livingstone, D. N. (1987) *Darwin's Forgotten Defenders*, Edinburgh: Eerdmans and Scottish Academic Press.
- McShea, D. W. (1991) "Complexity and evolution: what everybody knows", *Biology and Philosophy* 6, 303-24.
- Ramsey, F. P. (1931) "Truth and probability", in *The Foundations of Mathematics and Other Logical Essays*, London: Routledge and Kegan Paul.

Shannon, C. E. (1948) “The mathematical theory of communication”, *Bell System Technical Journal*, July and October.

Solomonov, R. J. (1964) “A formal theory of inductive inference”, *Information and Control* 7, 1- 22

Von Neumann, J. (1966) *Theory of Self-Reproducing Automata*, Urbana Illinois: University of Illinois Press.