Autopoiesis and Cognition

Paul Bourgine
CREA
Ecole Polytechnique
1 rue Descartes
75005 Paris
France
bourgine@poly.polytechnique.fr

John Stewart
CNRS
COSTECH
Centre Pierre Guillaumat
Université de Compiègne
BP 60649
60206 Compiègne
France
John.Stewart@utc.fr

Abstract This article revisits the concept of autopoiesis and examines its relation to cognition and life. We present a mathematical model of a 3D tesselation automaton, considered as a minimal example of autopoiesis. This leads us to a thesis T1: "An autopoietic system can be described as a random dynamical system, which is defined only within its organized autopoietic domain." We propose a modified definition of autopoiesis: "An autopoietic system is a network of processes that produces the components that reproduce the network, and that also regulates the boundary conditions necessary for its ongoing existence as a network." We also propose a definition of cognition: "A system is cognitive if and only if sensory inputs serve to trigger actions in a specific way, so as to satisfy a viability constraint." It follows from these definitions that the concepts of autopoiesis and cognition, although deeply related in their connection with the regulation of the boundary conditions of the system, are not immediately identical: a system can be autopoietic without being cognitive, and cognitive without being autopoietic. Finally, we propose a thesis T2: "A system that is both autopoietic and cognitive is a living system."

Keywords
Autopoiesis, cognition, life, tesselation automaton, random dynamical system, boundary conditions

Systems are not in Nature, they are in the mind of humans.—Claude Bernard

1 Introduction

The concept of *autopoiesis* is due to Maturana and Varela [8, 9]. Etymologically, "autopoiesis" means the capacity to produce oneself. It was proposed in order to characterize living organisms. In these publications, as in those that followed, Maturana and Varela continually associated the concept of autopoiesis with that of *cognition* [8, 10, 18]. The aim of this article is to revisit the concepts of autopoiesis and cognition in the hope of clarifying them and their mutual relation.

A rough preliminary definition of an autopoietic system is that of a network of processes that produce the components that reproduce those processes. This definition is not as complete or precise as those we shall propose below, but it clearly applies to the paradigmatic example of an autopoietic system: that of a living cell.

A rough preliminary definition of a cognitive system is that of a set of processes in structural coupling with its environment such that the system adapts to its environment and/or transforms that environment in such a way as to adapt the environment to the needs of the system. This definition is not very precise either, but it clearly applies to the paradigmatic example of animals with recognizable patterns of adaptive behavior.

The definition of autopoiesis is ambitious, in that it aims at specifying a property common to all living organisms that is the necessary minimum for life. The ideal would

be a definition of the living that was also a sufficient condition for life. However, on the basis of the preliminary definition given above, it is not trivial to deal adequately with the case of multicellular organisms and animals in particular. Conversely, the preliminary definition of cognition does not apply in an immediately obvious way to unicellular organisms or multicellular plants. In other words, the understanding of unicellular organisms focuses naturally on metabolism and autopoiesis, whereas the understanding of multicellular organisms refers rather to cognition. The object of this article is the tension between these two modes of understanding living organisms. Our aim is to examine the deep relations that exist among the concepts of the living, cognition, and autopoiesis. Maturana and Varela are happy to talk about cognition as soon as a system is autopoietic, but other people are not, whence the motivation of this article. As an example of discussion as to whether cognition can simply be equated to life, Wheeler [22] argues for a *continuity* rather than an equivalence between the two concepts.

The focus of the present article will be two theses that have been proposed by Maturana and Varela:

T-MV1: All living systems are autopoietic systems.

T-MV2: All living systems are cognitive systems.

Of course, these two theses (and their possible or dubious converses) are really meaningful only if the terms involved are rigorously defined. The strategy adopted in this article consists of making a distinction between *theses* (indicated notationally by T-), and *definitions* (indicated notationally by D-). Thus, we will attempt to define more precisely exactly what one means by "autopoietic systems" (D-Ap) and by "cognitive systems" (D-C). As Claude Bernard said [2] (and Maturana would readily agree), "systems are not pre-given as such in Nature but are rather constructed in the human mind"; we therefore have a certain degree of latitude in elaborating adequate definitions. We do however wish to remain as close as possible to Maturana and Varela, and to other authors inspired by them, such as those of [6, 7, 20].

The theses T-MV1 and T-MV2 have been put forward by Maturana and Varela, and our hope is that after the clarifications introduced by the definitions of key terms, they may form the object of a reasonable consensus. To the extent that the converse relations are valid, they would make it possible to identify certain classes with each other, two by two. For example, it is tempting to consider that all autopoietic systems are necessarily living systems, in which case we would have a necessary *and* sufficient condition for characterizing living organisms. However, it is not certain that all systems that may in the future be identified as autopoietic systems will appropriately be characterized as living. It is perhaps more convincing to follow Maturana and to say that all *molecular* autopoietic systems are living, although even that is not certain. It is not certain either that there is a consensus for considering that cognitive systems are necessarily living systems, or even that they are necessarily autopoietic. Key cases for T-MV1 and T-MV2 (can organisms as simple as bacteria properly be considered as cognitive? can there be a system that is minimally autopoietic but not cognitive?) will be discussed in some detail. We shall return to a consideration of these questions in our conclusions.

2 What is an Autopoietic System?

In Schrödinger's book "What is Life?" [14], the living organism is considered as an open system, continually traversed by a flux of matter and energy; this flux is a necessary condition enabling the system to maintain its organization. Thus, considered as a physical system, a living organism obeys both the first and second principles of thermodynamics (as of course it must).

Autopoiesis focuses on the network of processes that produce the components which reproduce the processes. The definitions of autopoiesis have evolved in the works of both Maturana and Varela:

D-Ap1: "*An autopoietic system is organized (defined as a unity) as a network of processes of production (transformation and destruction) of components that produces the components that*

(a) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produce them and

(b) constitute it (the machine) as a concrete unity in the space in which they exist by specifying the topological domain of its realization as such a network" [18].

Maturana recently[1] proposed:

D-Ap2: "*A molecular autopoietic system is a closed network of molecular productions that recursively produce the same network of molecular productions that produced them and specify its boundary remaining open to the flow of matter through it.*"

In his last book, Varela [21] proposed to simplify the previous definition into three basic criteria[2]:

D-Ap3: *A system is autopoietic if:*

(a) it has a semi-permeable boundary,

(b) the boundary is produced from within the system, and

(c) it encompasses reactions that regenerate the components of the system.

The last two definitions clearly indicate that autopoietic systems are "open systems" in the same sense as the physicist's. Nevertheless, the accent remains on the organizational closure and the recursivity that enables the system to produce itself. These two most recent definitions are sufficiently close to each other that we may adopt their common core meaning as a definition of autopoiesis for much of this article; more specifically, we retain the definition of Maturana, but removing the restriction that the components are necessarily molecular:

D-Ap4: "*An autopoietic system is a closed network of productions of components that recursively produce the components and the same network that produced them; the network also specifies its own boundary, while remaining open to the flow of matter and energy through it.*"

We may now ask what would be a *minimal* system that presents the property of autopoiesis as so defined. Such a system must have a boundary, with components produced in the volume enclosed by the boundary on the basis of a flow of matter across the boundary. The tesselation automaton has been proposed as an example of such a minimal system [6, 7, 18]. Here, we take a new look at this minimal system by developing a mathematical model. This will enable us to introduce our main argument, and will lead us to a revised definition of autopoiesis.

1 http://web.matriztica.org/1290/article-28335.html.

2 Quoted from Luisi [6].

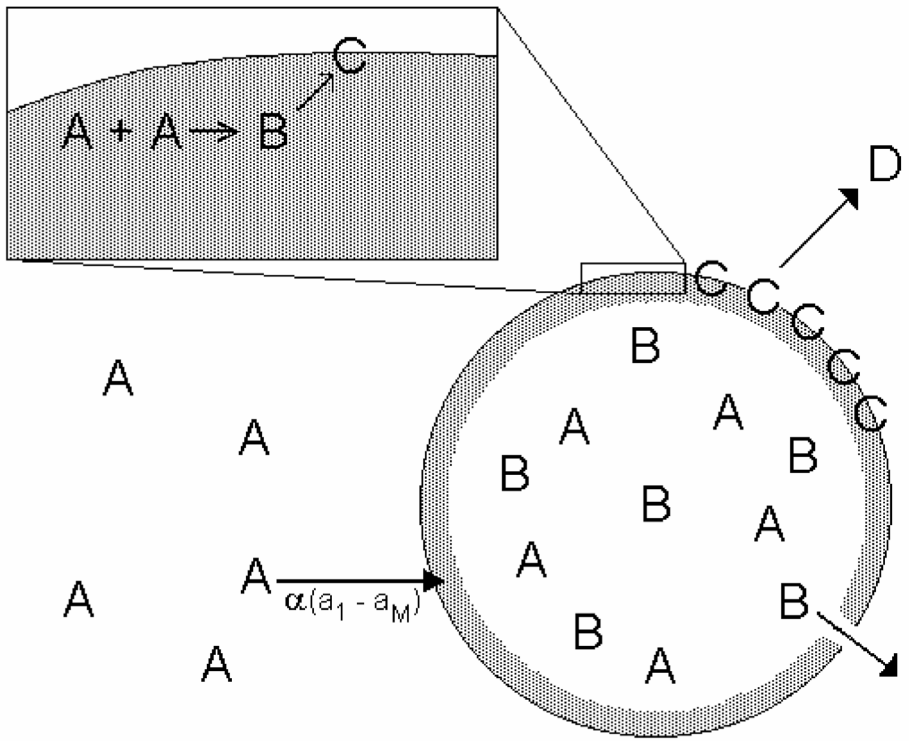

Figure 1. A schematic illustration of the tesselation automaton described in the text. The enlarged inset shows the processes occurring in a thin volume just under the membrane (catalytic production of *B*, and *B* components entering the membrane to become *C* components). The reaction $C \rightarrow D$ represents the disintegration of a *C* component, leaving a hole in the membrane. *B* components are normally confined by the membrane, but can be lost through holes.

2.1 A Minimal Autopoietic System: The Tesselation Automaton

Let us therefore consider the following tesselation automaton, which is characterized by a semi-permeable membrane and by the fact that the automaton as a whole is autocatalytic. We will pay especial attention to the local dynamics involved in repair of the membrane when holes are formed. The automaton consists of a spherical membrane enclosing an internal volume; it is schematically represented in Figure 1.

(a) *The structure of the membrane.* The membrane is formed of components *C*, assembled so as to form a two-dimensional sheet (the automaton considered is here three-dimensional[3]). With a rate-constant k_c per unit surface area, each of the *C* components can disintegrate: $C \rightarrow D$. The end product *D* cannot integrate the membrane: it escapes into the outside environment, leaving a hole in the membrane (or, if the *C* component that disintegrated was already on the edge of a hole, the hole becomes bigger).

3 In an early fragmentary discussion presenting the original tesselation automaton, ⟨http://www.eeng.dcu.ie/ alife/bmcm9701⟩, Appendix A), Varela noted: "It is evident that the system. . . could be extended to a three dimensional system. [However,] it is judged that the implementation of a three dimensional model would involve difficulties which are not worth considering at this stage, as the extension. . . would not involve any conceptual modifications." We prefer here a 3D version and a 2D membrane, because it enables us to model temporary holes in the membrane, an important feature of our model as we shall see. In a 2D automaton with a 1D membrane, two holes would immediately disrupt the membrane. The 2D membrane retains a topological continuity in spite of a substantial fraction of "holes" in it (although ultimately a 2D membrane can also fragment). In addition, with our mathematical treatment, a 3D automaton does not involve any additional difficulties compared to a 2D automaton.

(b) *Production of B.* The B components are formed by a reaction between two substrate molecules A: $A + A \rightarrow B$. This reaction only occurs inside the volume enclosed by the membrane. In the original formulation by Varela [18], this was attributed to the presence of a catalyst enclosed within the membrane. However, since the catalyst itself was not produced by the system, this system fell short of autopoiesis as defined. Here, we propose to consider that the reaction $A + A \rightarrow B$ is catalyzed by the inside surface of the membrane, that is, C catalyzes the production of B. There is an upper limit to the concentration of B that can be accumulated within the membrane.

(c) *Diffusion of A.* The substrate A exists in moderately high concentration in the environment of the system; it diffuses freely across the membrane. Since the concentration of A inside the membrane is depleted by the chemical reaction $A + A \rightarrow B$, there is a net flux of A into the system. Within the volume enclosed by the membrane, A diffuses freely.

(d) *Repair of the membrane and permeability to B.* The intact membrane is impermeable to the B components, which thus accumulate inside the membrane. The B components diffuse freely within the volume enclosed by the membrane. If two single B components collide, they do not combine. However, if a free B component collides with the edge of a hole in the membrane, it attaches to the surface and repairs the hole—completely if the hole was of size $1C$ (where C is the area on the membrane occupied by a C component), partially if the hole is bigger. A B component that integrates the membrane in this way thereby becomes a C component. If the hole is larger than $1C$, there is a finite probability that the B component will pass through the hole without attaching to any of the free edges. This probability increases with the size of the hole. Thus the membrane is completely impermeable to B if there are no holes, or the holes are of size $1C$; it is partially permeable for holes of size larger than $1C$. (In practice, in the discrete model, B components will attach if the size of the hole is less than about $10C$.)

It may be noted that this formulation focuses on the transformations of the *material* elements A, B (single molecules inside the volume), C (the same components when aggregated in the membrane), and the end product D. We have thus concentrated on fluxes of matter. However, since the sequential processes $A + A \rightarrow B \rightarrow C \rightarrow D$ are all considered to occur spontaneously,[4] the overall result $A + A \rightarrow D$ corresponds also to a dissipation of *energy*. This tesselation automaton is thus a dissipative structure.

Two mathematical models of this automaton are presented in the Appendix. In model I, the functional role of an intact membrane in confining the B components is not taken into account: these components remain within the internal volume, irrespective of the presence of holes in the membrane. This will be useful as a reference system, which in our view is not properly autopoietic. In model II, the loss of B components through holes in the membrane is taken into account; as we shall see, this system depends for its very existence on the repair and maintenance of the membrane.

In both models, the key variable for understanding the structural dynamics of the system is $c^* = c_1/c_M$, that is, the proportion of the total surface area that is occupied by C components. In the case of model I, the analysis presented in the Appendix suggests that the system *always* has a single fixed-point attractor PF1. The value of c^* at this equilibrium depends on a number of parameters; but it can vary from 99% to less than 1% without any apparent discontinuity. This is not a very satisfactory model

4 This makes the model a bit unrealistic: the whole system is based on a "toy chemistry" invented to order. This is the price we pay for the simplicity of the model.

of autopoiesis (although it may provide a clue for a scenario concerning the possible *origin* of this sort of system, which is difficult in the case of the model II). It may be remarked that percolation analysis shows that there exists a critical value for c^*, below which the membrane no longer maintains topological continuity and fragments into two or more separate pieces. This may be taken as a model of death; but the separation of the phase space into two distinct regions remains rather arbitrary with respect to the actual dynamic functioning itself.

Model II, while mathematically very similar to model I, nevertheless appears to exhibit qualitatively different behavior. Here, there are two separate fixed-point attractors, PF1 and PF0. PF1 is similar to PF1 in model I, except that the value of c^* is necessarily above 50%. At PF0, $c^* = 0$ and the system has entirely collapsed. PF1 and PF0 are separated by a point of bifurcation, PF2, which is an unstable fixed point. It is worthwhile describing the dynamics of model II in qualitative terms. There are two quite different regimes in the phase space of the operation of the system. Above the point of bifurcation, the system maintains itself dynamically with a basin of attraction PF1 such that the rate of disintegration of the membrane due to $C \rightarrow D$ is balanced by the repair process $B \rightarrow C$. Note also that at PF1, since c^* is necessarily above 50%, it is guaranteed to be above the value at which percolation of the holes and qualitative fragmentation of the membrane can occur. Below the point of bifurcation, the dynamics of repair no longer manages to balance the formation of holes. It is not difficult to guess the consequences. The holes increase in size and number; the loss of B is accelerated; and this in turn makes it even more difficult to repair the holes. There is a vicious positive feedback: the system heads straight for disaster and its collapse as an autopoietic system.

This emphasizes the fundamental role of the membrane that confines the interactions within the intracellular space. These interactions reconstitute the components and the network of processes that produce them. But above all, these interactions make it possible to repair the membrane itself, and to maintain the vital property of semipermeability. The recursivity involved in the fact that the maintenance of the membrane itself depends on the functional integrity of the membrane decisively accentuates this separation into two qualitatively different sorts of functioning. Like a candle flame, the system is either "alive" or "dead." If it is "alive," it may waver at times (several holes may appear), but it can recover and go on as though nothing had happened. If it is "dead," however, nothing can resuscitate it; the system collapses and disintegrates entirely.

This *functional* separation into two qualitatively different sorts of functioning appears to us to be a necessary characteristic of systems that qualify as autopoietic. Varela [19] emphasizes that autopoietic systems can deal with a certain range of perturbations, but that perturbations beyond this range lead to the collapse of the system; and he identifies this feature as the basis of intentionality. This property is exhibited by model II, but not by model I. It may be noted that the original automaton proposed by Varela [18] also had this characteristic, because if the membrane was disrupted for too long, the catalyst inside the membrane could escape, which led to the total disintegration of the system. However, in Varela's model this result was obtained solely by computer simulation, which led to a certain vagueness in the interpretation (at which point, exactly, could one say that the membrane was decisively disrupted?). In addition, as we have already noted, Varela's model was not fully autopoietic (on his own definition) because the catalyst itself was not produced by the system. Here, we have overcome this limitation by supposing that the synthesis of B from two A molecules is catalyzed by the inner surface of the membrane itself; and our mathematical treatment, compared to computer simulation, opens up the possibility of characterizing more precisely the qualitative dynamics of the system.

2.2 Autopoietic Systems as Random Dynamical Systems

In this section, we aim to identify and generalize the key theoretical lesson from our analysis of the tesselation automaton. For this, we require two definitions.

The following definition of a *random dynamical system.* is drawn from Arnold [1].

Let X the be the state space of a system. Let $F = X \rightarrow X$ be a space of functions from X to X. Let T be the set of times with its additive group structure. T can be discrete or continuous, two-sided like R and Z, or one-sided like R^+ and Z^+.

Let us consider a differential equation if T is continuous, or a difference equation if T is discrete. Let $f_t(x)$ be the solution of this differential or difference equation, where x is the state at time 0. Geometrically, the whole set of trajectories is completely defined by this function f. This function has the two properties that $f_0(x) = x$ and $f_{s+t}(x) = f_s(f_t(x))$. Such a function is called a *dynamical system*: that is, a dynamical system is an application f: $T \rightarrow F$ such that

(1) $f_0 = \mathrm{Id}$,

(2) $f_{s+t} = f_s \circ f_t$.

Let us remark that the only assumption is that the same mechanism is used at each step. The main examples are differential and difference equations. Let us also remark that f describes the behavior of a system whatever its initial position is, and thus describes the behavior of a whole class of entities with the same behavior.

D-RDS: *Let* (Ω, F, P) *be a probability space, and let* θ *be a dynamical system in the space* Ω. *In other words, we have probabilities on the trajectories of* ω *in* Ω *through time. A random dynamical system is an application* f: $T^*\Omega \rightarrow F$ *such that*

(1) $f_{0,\omega} = \mathrm{Id}$,

(2) $f_{s+t,\omega} = f_{s,\theta(t)\omega} \circ f_{t,\omega}$.

There is an evolution of ω in Ω. The evolution of the dynamical system is perturbed at time t by the present position $\theta_t\omega$ of ω at the same time. Let us interpret $\theta_t\omega$ as the trajectory of the state of the environment.

Secondly, we require a definition of the *organized phase* of an autopoietic system:

D-OPh: *The organised phase of an autopoietic system is the autopoietic domain where it maintains its ongoing existence as a network and the regulation of its boundary conditions.*

We may now state our main thesis:

T1: *An autopoietic system can be described as a random dynamical system that is defined only within its organized autopoietic domain (RDS-OPh).*

Our discussion has two parts:

(i) What is involved in making it possible to describe an autopoietic system as a "random dynamical system"?

(ii) Conversely, since not all dynamical systems are autopoietic systems, what is special about certain dynamical systems that enables them to qualify as autopoietic systems? Can their special features be expressed mathematically, and if so, how?

(i): Within this first point, there are again two subaspects:

(a) since autopoietic systems are a priori spatially distributed, there are the difficulties involved in spatial integration and characterization of the emergent properties of the system as a whole; and

(b) since autopoietic systems are a priori locally stochastic, there are the difficulties involved in obtaining (again) a characterization of the emergent properties of the system as a whole, this time in the spirit of statistical mechanics.

We shall take these points in that order.

(a): We take as a starting point the postulate that an autopoietic system is a spatially distributed system of components. To each component we can attribute a mathematical function that represents the way the state of the component changes as a result of its local interactions with neighboring components. This gives rise to a description of the system by a set of partial differential equations (PDEs), for example, Equations 2 and 3 in the Appendix. The (approximate) theoretical treatment, which highlights the separation of the phase space into two qualitatively different regions, is based on the dynamic system defined by Equations 1, 4, and 5. What is the fundamental procedure that makes it possible to pass from a system of partial differential equations to a system of ordinary differential equations? Our main argument rests on the idea that it is, very precisely, the *functional* role of the membrane that makes this passage possible. *It is because the membrane makes it possible for the distributed system to control its own boundary conditions that the (re)modeling of the system as a random dynamical system is rendered possible.*

Let us take a closer look at what is involved in passing from a PDE system to an ODE system. This transition relies on a spatial integration; and spatial integration can only be undertaken against specific, spatial boundary conditions, which must be *constant* in time. This poses an immediate difficulty for any system whose spatial demarcation is, in fact, variable. For that particular class of systems it will not *generally* be possible to transit from PDE to ODE. However, within that general class there is—arguably—a particular subclass for which this transformation can still be achieved: namely, those systems with *potentially* dynamic spatial boundaries where the boundary is, nonetheless, somehow *stabilized* and thus rendered constant after all. Insofar as such boundary stabilization (one may even say "homeostasis") may be precisely characteristic of autopoietic systems, a route to the thesis T1 is opened.

However, there is still some way to go. Autopoiesis may well inherently involve active stabilization of an intrinsically unstable membrane; but that is very far from rendering such a membrane "constant" in the manner that would be required to permit a classical transformation to ODEs. Rather, membranes would normally be extremely spatially dynamic—growing, shrinking, changing shape, and, of course, changing in detailed permeability. Thus, the ability to integrate out (in a time-independent way) a time-varying spatial boundary condition is problematical. The idea we wish to pursue is that autopoietic systems are characterized by a somewhat weaker constraint: that although the system boundary is not constant, it is, at least, *determined* and lies within a basin of attraction around a fixed point. This is the case for the mathematical model presented in the Appendix, and we speculate that this is sufficient to (recursively) permit the transformation from PDEs to ODEs.[5]

5 We are grateful for the perceptive comments of one of the referees on this point, and for permission to incorporate these remarks in our discussion.

The thesis T1 is not meant to imply that modeling an autopoietic system as a random dynamical system will be an easy task. In the minimal autopoietic system modeled here, the task was greatly facilitated by the postulate that the surface S of the membrane remained fixed in space, so that the only dynamical variable was the permeability to B. It is a mathematical challenge to develop a formalism in which the fact that (spatial) boundary conditions are part of a random dynamical system is sufficient to mandate a transformation from PDEs to ODEs *in general*. In each case, as we have schematically outlined for the tesselation automaton, it will be necessary to propose multiple models hierarchically embedded at different levels of organization, and approximate models in order to render the complex dynamic phenomena intelligible. Many of the requisite mathematical tools do not yet exist. If the thesis T1 is indeed correct, an immense task awaits generations of researchers in order to substantiate it.

(b): The second difficulty is related to the stochastic nature of certain processes. This difficulty is well recognized in the field of artificial life: if individual entities and/or events play a crucial role, it is not possible to obtain accurate ODEs by statistical aggregation, so that the only recourse is computer simulation. The drawback of this is that the results are often ad hoc and difficult to validate or to generalize. In the case of the tesselation automaton as modeled in the Appendix, there are sorts of stochastic processes: *internal*, that is, concerning formation and repair of holes in the membrane; and *external*, concerning the relationship to fluctuations in the external environment. With respect to the *internal* processes, we have proceeded by formulating a continuous approximation and validating this *locally* by Monte Carlo simulations. With respect to the external processes, we have proceeded by distinguishing *rapid* fluctuations (which are arguably buffered by the accumulation of B components within the whole volume of the automaton), and *long-term* fluctuations, for which we can assume that the system is in "adiabatic" dynamic equilibrium. It remains to be shown that these procedures are valid, even in the deliberately hyper-simplified case of the tesselation automaton; the difficulties will be obviously greater for more realistic, but inevitably more complicated, models.

(ii): The converse of the thesis T1 is clearly not true: not all dynamic systems are autopoietic systems. This is the flip side of T1: if it is true that all autopoietic systems can be described as random dynamical systems, then what is special about autopoietic systems? We may take as a simple example the movement of a satellite in orbit around a central mass (a planet around the sun, or the moon around the earth). Such systems can be described with an ODE, namely, $dX/dt = f(X)$; it was Newton's achievement to show that in order to achieve such a description, the state vector X must include not only the position but also the instantaneous velocity of the satellite. Why is such a system not autopoietic?

A part of the answer lies in the semantics of the interpretation of the mathematical equations. In order for a system to count as autopoietic, it must be possible to interpret the functioning of the system as the decay and continuous regeneration of its components. In addition, it must be possible to interpret the equations as involving the active regulation of certain critical boundary conditions. The tesselation automaton, as modeled, does allow such interpretations, whereas the Newtonian equations for planetary motion do not. It is however difficult to give a mathematical definition of such semantic constraints.

A mathematically more tractable part of the answer lies in the principle of Laplacian determinism: for the planetary system, given X at any point in time and given $f(X)$, the state of the system is fully determined for all future times (and all past times as well). Thus, in a sense, there is nothing that can happen to the system. This was the case for model I in the Appendix, and for this reason we considered that it did not conform to the spirit of autopoiesis. What seems to be necessary in order to consider the system

as autopoietic is that it should have (at least) *two* attractors (or classes of attractors), separated by a point of bifurcation. In addition, one of the attractors must correspond to the complete collapse and disintegration of the system (this is a mathematical condition for semantic interpretation as death); and the other attractor must correspond to a state in which the variables all maintain nonzero values (this is a mathematical condition for semantic interpretation as a system that actively regulates its own boundary conditions in such a way as to dynamically regenerate its own boundary and its existence as a system, i.e., life). This was the case for model II, which is why we propose it as a minimal model of autopoiesis; it was not the case for model I, and it is not the case for the planetary system either.

2.3 Autopoietic Systems at Several Levels of Organization

The living cell is a paradigmatic model of an autopoietic system. Living cells are first-order autopoietic systems. We can then ask about the possibility of higher-order autopoietic systems.

Multicellular organisms appear to be societies of cells, differentiated into different cell types, engaged in a complex network of coordinated processes, with a very sophisticated degree of organization. If we consider that individual cells are the components of a second-order autopoietic system, it is clear that a large number of these components disappear every day and are constantly renewed. This is the case, in particular, for the epidermal cells. Many of the features of autopoietic systems are thus present. What is necessary in order to consider multicellular organisms as autopoietic systems, not just because they are made up of first-order autopoietic systems, but as second-order autopoietic systems in their own right? We are looking for something analogous to the membrane of unicellular organisms, which will make it possible for the distributed system to control its own boundary conditions, and thus render the (re)modeling of the system as a random dynamical system possible.

It is tempting to cast the skin (of those animals that possess one) in this key role. There are, however, certain problems with this, both empirical and theoretical. Empirically, not all multicellular organisms have a skin (plants, many invertebrate animals). And even for animals such as vertebrates that do have a skin, the whole intestinal tract is topologically external to the animal. However, the gut harbors essential symbiotic microorganisms, so that it is difficult to dismiss the common-sense view that gut and stomach contents are "part of" the animal. The theoretical reasons are more profound. The key passage is that from a system of discrete, spatialized partial differential equations to a system of ordinary differential equations where the sharp bifurcation between two qualitatively distinct regions of phase space—"alive" and "dead"—clearly appears as such. What is required for this is a "boundary" defined in *functional* terms; there is an uneasy suspicion that identifying this functional boundary with a structure such as the skin may be an over-facile and misleading reification. (This criticism may also apply to the membrane of unicellular organisms, although in that case there is an overwhelming case for supposing that, if only for fundamentally contingent reasons, the functional boundary and the cell membrane do at least happen to coincide in all known living organisms.)

These considerations are even more compelling if we consider the possibility of a third-order autopoietic system including among its components both first- and second-order autopoietic systems. One possible candidate here is the whole ecosphere of the planet Earth, considered in the light of Lovelock's Gaia hypothesis [5]. This is certainly a system consisting of a network of processes that continually produce the components (including first- and second-order autopoietic systems) that reproduce those processes. Lovelock has pointed out that the biosphere as a whole actively regulates its boundary conditions (the composition and temperature of the atmosphere, the oceans, and the

soil) in just such a way that the continuation of life on the planet is possible.[6] It must be an open question at present whether the terrestrial ecosphere actually is a *bona fide* autopoietic system; to answer this question, it will be necessary to put our mathematical definitions of autopoiesis in full working order, and then to carry out the considerable task of applying them to the case of the ecosphere. However, we can already say at this stage that we do not want to rule out this possibility simply because the ecosphere does not have a single clearly reified membrane. Another set of interesting candidates is provided by the colonies of insects such as bees and ants. Does a colony of this sort really form a super-organism in its own right? Our point here is not to argue particularly for or against the hypothesis that Gaia, or an insect colony, is a third-order autopoietic system. Our point is rather that for these to become tractable questions, we require a renewed definition of autopoiesis that does not depend on an excessively reified definition of "membrane" or "boundary."

These considerations lead us to propose a modified, more general definition of autopoiesis:

D-Ap5: *An autopoietic system is a network of processes that produces the components that reproduce the network, and that also regulates the boundary conditions necessary for its ongoing existence as a network.*

The difference from the preceding definitions, in particular D-Ap4, is that the emphasis on the semipermeable boundary, which lends itself to over-reification of a membrane, is replaced by a more dynamical emphasis on the regulation of boundary conditions. However, the requirement for a *functional* boundary, which must be produced by the system itself, is maintained. It is interesting to note that Zaretzky and Letelier [23], in an article devoted to an articulation between the concept of autopoiesis and the concept of "closure under efficient cause" due to Rosen [13], also come to the conclusion that a distinction must be made between a physical boundary and a functional boundary.

This new definition changes nothing in the thesis T1—on the contrary, it comes from a consideration of some of its implications. With this definition, for the same reasons as before, an autopoietic system can be modeled as a random dynamical system.

3 Cognition

We will turn now, somewhat more briefly, to the question of cognition. We have seen that autopoietic systems are necessarily thermodynamically open systems, continually traversed by a flux of matter and energy. Thus living organisms are clearly *dissipative structures* of the sort studied by Prigogine and Stengers [12]. However, as Simondon [15] has pointed out very perceptively, purely physical (or physicochemical) dissipative structures (such as the flame of a candle, or a cyclone) are intrinsically *ephemeral*; they last only so long as certain external conditions, over which they exert no control at all, are maintained. By contrast, there is something special about living organisms considered as dissipative structures: they function in such a way that although they can be disrupted and killed at any moment, they are *potentially immortal*. In other words,

6 It is for this reason that Lovelock [5] insisted that a planet is *globally* either "alive"—as in the case of the Earth, with an atmosphere far from chemical equilibrium—or "dead," as in the case of all the other planets in our solar system, whose atmospheres are all virtually at chemical equilibrium. On this view, there is no chance that a few isolated living organisms, bacteria for example, could be found on the other planets (much to the chagrin of NASA, which gave Lovelock one research contract but not two; however, this is a digression).

their functioning extends to include a regulation of their own boundary conditions.[7] As we have seen above, this is a key feature that distinguishes autopoietic systems from dynamical systems in general.

Cognition, in this perspective, focuses on exactly this aspect: management of the interactions between an organism and its environment. This clearly prefigures an intimate relationship between cognition and autopoiesis, because in the new definition of autopoiesis (D-Ap5) central emphasis is placed on the regulation of the system's own boundary conditions. However, as we noted in the introduction, autopoiesis focuses naturally on the internal functioning of the organism, notably its metabolism; whereas cognition naturally thematizes the interactions between an organism and its environment.

3.1 Sensory Input and Actions

Analytically, the interactions between a system and its environment can be subdivided into two sorts.[8] Firstly, there are those interactions that have consequences for the internal state of the organism: we may call these *type A* interactions. Secondly, there are those interactions that have consequences for the state of the (proximal) environment, or that modify the relation of the system to its environment: we may call these *type B* interactions. This terminology allows us to propose a definition of "cognition":

D-C1: *A system is cognitive if and only if type A interactions serve to trigger type B interactions in a specific way, so as to satisfy a viability constraint.*

If (but only if) a system is cognitive in this sense, type A interactions can be termed "sensations," and type B interactions can be termed "actions." As we shall see, in order to qualify as cognitive, the type A interactions in question will generally be mediated by more or less specialized *sensory organs* situated in the boundary of the system, and type B interactions by more or less specialized *effector organs* also situated in the boundary of the system. It may be useful to illustrate this by examples of interactions such as falling down stairs, eating, or breathing (including the breathing of a poisonous but odorless gas). Ordinarily, such interactions are not considered as "cognitive." On the definition proposed here, they will not be cognitive unless the consequences for the internal state of the system are employed to trigger specific actions that promote the viability of the system. Thus, falling down stairs will be cognitive if but only if the fall triggers reactions such as a modification of muscle tone that limits the damage; and this does require specific sensory and motor organs. Similarly, eating is cognitive if but only if it triggers a reaction of satiety that prevents damage from overeating; breathing a poisonous gas is cognitive if but only if it triggers evasive action, which will require a specific sensory organ to detect the poison, and the resulting sensation to trigger an appropriate, coordinated motor response.

The definition DC-1 employs a term, "viability constraint," that is deliberately rather vague. It clearly includes the case of autopoietic systems, for whom the viability constraint consists intrinsically in maintaining their autopoiesis. However, we did not wish

7 The appearance and disappearance of dissipative structures corresponds to a phase shift, so that they are all-or-nothing phenomena. We suggest that the distinction between intrinsically ephemeral and potentially immortal systems is likewise qualitative, although this point requires substantiation by a sufficient number of concrete studies aimed at classifying borderline candidate systems (rivers? stars? viruses?).

8 Maturana and Varela state clearly that because of their circular organization, autopoietic systems are not "input-output" systems. However, they also state that for analytical purposes, it is always possible to cut a circle at two points and to regard each "half" as having inputs and outputs. Note that concerning the circular relationship between an organism and its environment (which is the question here), what are "sensory inputs" for the organism are "outputs" for the environment, and what are "output actions" for the organism are "inputs" for the environment. Putting the two halves back together, it is clear that what the environment "is" *for the organism* is neither more nor less than the consequences of its actions for its subsequent sensations.

to *define* the satisficing of a viability constraint as the maintenance of autopoiesis, as this would have amounted to identifying cognition with autopoiesis rather trivially, by definitional fiat. The sort of counterexamples we have in mind are the autonomous robots of artificial life [3]. By metaphorical extension, we can consider (for example) that a robot that navigates on the surface of a table may be required *neither* to simply remain immobile, *nor* to fall off the edge of the table and crash. The extension is useful also for other putatively cognitive systems, such as the immune system [16].

This definition, if accepted, has certain implications for cognitive science. It considers that cognition (or intelligence) is rooted in low-level sensory-motor loops—a view consonant with Piaget's developmental perspective [11]. However, this perspective is not limited to low-level cognition [17]. It becomes more natural to speak of "cognition" when we consider that the sensory input must not only be used to guide the actions in an intelligent way, but that, conversely, the actions of an organism also have consequences for its subsequent sensory inputs. It follows that what the environment "is" *for the organism* amounts to neither more nor less than these consequences of its actions for its sensory inputs. Thus, in this perspective, the "objects of cognition" are quite inseparable from the organism itself; in a certain sense, the organism itself specifies what there can be for it to perceive; without the organism, the objects of its cognition would not even exist.[9]

3.2 Bacterial Cognition

We noted in the introduction that our preliminary definition of cognition did not seem to apply in an immediately obvious way to unicellular organisms or multicellular plants. We now return to this question, using the definition in D-C1. It is only analytically that we can separate sensory inputs and actions; since the sensory inputs guide the actions, but the actions have consequences for subsequent sensory inputs, the two together form a dynamic loop. Cognition, in the present perspective, amounts to the emergent characteristics of this dynamical system.

There is no difficulty in applying the scheme just described to bacteria. For example, certain bacteria have molecular receptors in their membranes that can take two conformations:

- A if the concentration of sugar in their local environment is constant or decreasing;
- B if the sugar concentration is increasing.

In our terminology, A and B count as *sensory inputs*.

These bacteria also have cilia on the membrane, and these cilia can move in two ways:

- X: The cilia vibrate in a coordinated fashion, in smooth waves (an external observer can see that this causes the bacterium to advance more or less in a straight line, but of course the bacterium itself does not know that).
- Y: The cilia vibrate in a chaotic, uncoordinated fashion (the external observer sees that this causes the bacterium to re-orient itself randomly, but again the bacterium cannot know that).

In our terminology, X and Y count as *actions*.

9 This is in sharp contrast to the computational theory of mind in classical cognitive science. Here, the objects of perception are pre-given entities that exist "in themselves" quite independent of their relation to a cognitive subject; in this perspective, the task of cognition is to "represent" these objects correctly, that is, the representation should be suitably isomorphic to its pre-given referent.

In a normal, viable bacterium the sensory inputs are used to guide the actions in the following way. The sensory input A triggers the action Y; the sensory input B triggers the action X. If the bacterium is in moderate proximity to a source of sugar, so that there is a radial concentration gradient, this mode of guiding the actions by the sensory inputs gives rise to the following emergent behavior. Suppose that the bacterium is initially in the action state X, but the orientation of the bacterium is such that the quasi-straight line leads away from the source of sugar. In this case, the local sugar concentration will decrease in time, triggering the sensory input A. The bacterium will therefore shift to the action Y, and re-orient. It can of course happen that this new orientation also leads away from the sugar source, but in this case a similar sequence will occur and the bacterium will rapidly re-orient again. Sooner or later, the new orientation will lead generally *towards* the sugar source. This will cause an increase in local sugar concentration, sensory input B, and action X. The bacterium will persist in this movement (generally towards the source) until it reaches a col in the gradient, when the concentration will start to decrease again, leading to another re-orientation. The overall result is that the bacterium will tend to move towards the source, albeit in a highly zigzag fashion.[10] Since a sufficiently high concentration of sugar in the local environment is a necessary boundary condition for the maintenance of the bacterium's metabolism and autopoiesis, this emergent behavior is indeed an example of cognition as we have defined it here.[11] Note that if the sensory inputs were used to guide the actions in an *inappropriate* way—for example, if the sensory input A triggered the action X and B triggered Y, or if the sensory inputs had no effect at all on the actions—the bacterium would move away from a source of sugar, or drift aimlessly. Neither of these behaviors would be viable. Satisficing the viability constraint—in the case of the bacteria, maintaining their autopoiesis—thus *requires* cognition. Now the different modes of using sensory inputs to guide actions are all equally compatible with the basic laws of physics and chemistry. The laws governing viable cognitive behavior are thus specifically biological, beyond the realm of the laws of physics and chemistry.

3.3 Cognition and Autopoiesis

There is little difficulty in extending this sort of analysis to the case of multicellular plants. Plants grow upwards, turn their leaves towards the sun, and open and close their stomata to regulate their gaseous and aqueous exchanges with the air; they send their roots downwards, which both enables exchanges of water and mineral nutrients in the soil, and anchors the plant so as to maintain its overall orientation. These are all actions that are appropriately guided by sensory inputs, and thus clearly correspond to cognition as we define it here. Equally clearly, this behavior contributes decisively to regulating the boundary conditions of the organism so as to maintain them within the narrow range compatible with the maintenance of autopoiesis. Thus, notwithstanding the apparent tension between the concepts of autopoiesis and of cognition that we noted in the introduction, it appears that the two concepts are fundamentally related.

10 It is interesting to note that bacteria are so tiny that the difference in concentration between one end and the other of a single bacterium is quite undetectable, either by bio-molecules or by the most refined methods of human chemists. The bacteria are only able to perceive this relevant aspect of their environment because they are *acting* within that environment. In a certain sense, we may say that the bacteria transform their environment into an "ecological niche" with Gibsonian "affordances" for their actions. As Gibson [4] would have said, perception is "direct" and does not (necessarily?) involve computational representations; but what is thus perceived is not pre-given referential objects, but these "affordances" themselves that cannot be defined and do not even exist independently of the strategies of actions deployed by the organism.

11 Varela [19] considers that bacteria are "cognitive" because they are clearly autopoietic, and on his definition autopoiesis and cognition are identical. We do not presuppose that autopoiesis and cognition are immediately identical, and so although the somewhat counterintuitive conclusion that organisms as simple as bacteria should be regarded as cognitive is the same, our argument is not quite the same.

Finally, we shall return to question of the relation between autopoiesis and cognition. It follows from our definition D-C1 that for those cognitive systems whose viability constraint takes the form of maintaining their autopoiesis, cognition is necessary for autopoiesis. But this is not sufficient to prove a relation of identity between autopoiesis and cognition. By allowing metaphorical extensions of the term "viability constraint," we have already allowed the possibility that certain systems may be cognitive without being autopoietic. Conversely, is it possible for a minimal autopoietic system to exist that would *not* be "cognitive" as defined here?

The test case that springs to mind is the tesselation automaton that we have discussed at length in Section 2. This automaton produces its own functional boundary (the membrane), and that boundary is crucial for maintaining the boundary conditions for the metabolic processes that take place within the cell. However, nowhere in the scheme outlined in Section 2.1 is there any question of "actions" guided by "sensory inputs"; so that, unlike bacteria and plants, the tesselation automaton is not cognitive in the sense defined here. In the end, it comes down to a question of definition for which we have some latitude. If we were to *postulate* that cognition and autopoiesis are identical, then it would follow that the tesselation automaton is not autopoietic under that definition. However, we prefer to say that *minimal* autopoiesis does not necessarily require cognition.

Of course, as we have already noted, the scheme in Section 2.1 is only a model with a "toy chemistry" invented to order; it is an open question whether a tesselation automaton could actually exist materially with the physics and chemistry of the real world. This is the area where Luisi has been working for decades and is now joined by others, and the final answer is not yet in. However, it seems likely that in order to exist, such an automaton would have to be built artificially; it is not a very plausible candidate for the spontaneous origin of life on the planet Earth. This brings us back to the question of *life*, which was of course from the outset the principal motivation behind the concept of autopoiesis itself.

4 Summary and Conclusions: What is Life?

We have seen that on the basis of the definitions we have proposed, an autopoietic system is not necessarily cognitive; and that a cognitive system is not necessarily autopoietic. Where does this leave us with respect to Maturana and Varela's theses, T-MV1 and T-MV2?

Let us consider first T-MV1: "All living systems are autopoietic systems." The modified definition of autopoiesis that we have proposed in D-Ap5 does not affect the validity of this thesis. On the contrary, the modification was made precisely in order to include second- and third-order autopoietic systems (multicellular organisms and Gaia), which, intuitively, we do wish to consider as being alive. Thus T-MV1 is only consolidated.

Now for T-MV2: "All living systems are cognitive systems." If we accept T-MV1, it follows that the viability constraint that living systems must satisfice in order to qualify as cognitive takes the form of maintaining their autopoiesis. The examples we have studied, including bacteria and plants, strongly suggest that for all known living organisms of the planet Earth, the maintenance of autopoiesis does indeed require cognition as defined in D-C1. Thus T-MV2 also emerges quite unscathed, and rather strengthened, from our examination.

It follows, logically, that all living systems lie within the intersection between autopoietic systems and cognitive systems. Being both autopoietic and cognitive is thus a *necessary* condition for being a living system. But is it a *sufficient* condition? If the class of autopoietic systems were fully identical to the class of cognitive systems, as Maturana

and Varela originally proposed, then being autopoietic and/or cognitive would be not only a necessary but also a sufficient condition for being a living system. However, because (on our definitions) an autopoietic system is not necessarily cognitive (e.g., the tesselation automaton), and a cognitive system is not necessarily autopoietic (e.g., robots that satisfy a viability constraint other than autopoiesis), it is not logically necessary that a system that is both autopoietic and cognitive should thereby be a living system. We propose to consider this question in the form of a thesis T2:

T2: A system that is both autopoietic and cognitive system is a living system.

This thesis is meant in a spirit analogous to the thesis of Church: "Any calculation is formally equivalent to a recursive function." The two terms in this thesis do not have the same degree of conceptual precision: it was an attempt to give the rather vague and intuitive term "calculation" a more precise definition. It could not qualify as a "theorem," just because of this vagueness in one of the terms; it was meant rather as a challenge to come up with a counterexample of a process that was intuitively a "calculation" but was nevertheless not formally equivalent to a recursive function. In T2, it is the term "living system" that is vague and intuitive, and the terms "autopoiesis" and "cognition" that are precise (or rather, hopefully can be made precise, as we have attempted to do in this article). T2 is thus also a challenge, to come up with a counterexample in the form of a system that is both autopoietic and cognitive, but intuitively is not really living. We do not mean to commit the hubris of implying that such a counterexample is definitively impossible; we remark only that in order to substantiate a counterexample, it will be necessary to make explicit a fuller definition of exactly what we mean by "living." If no counterexamples can be found, then T2 will hold up as a scientific definition of "living"; if a counterexample is found, T2 will have served its purpose by giving way to an improved definition.

References

1. Arnold, L. (1998). *Random dynamical system*. New York: Springer-Verlag.
2. Bernard, C. (1947). *A introduction to experimental medicine (1858–1877)*. Mineola, NY: Dover.
3. Bourgine, P., & Bonabeau, E. (1998). Artificial life as a synthetic biology. In T. L. Kunii & A. Luciani (Eds.), *Cyberworlds*. New York: Springer-Verlag.
4. Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.
5. Lovelock, J. (1988). *The ages of Gaia*. New York: Norton.
6. Luisi, P. L. (2003). Autopoiesis: A review and a reappraisal. *Naturwissenschaften, 90*, 49–59.
7. McMullin, B., & Varela, F. (1997). Rediscovering computational autopoiesis. In P. Husbands & J. Harley (Eds.), *Proceedings of the fourth ECAL*. Cambridge, MA: MIT Press.
8. Maturana, H. (1970). Biology of cognition. BCL Report 9. University of Illinois.
9. Maturana, H., & Varela, F. (1973). *De máquinas y seres vivos*. Santiago: Editorial Universitaria.
10. Maturana, H., & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Boston: D. Reidel.
11. Piaget, J., & Inhelder, B. (1966). *La psychologie de l'enfant*. Paris: PUF.
12. Prigogine, A., & Stengers, I. (1979). *La nouvelle alliance*. Paris: Gallimard.
13. Rosen, R. (1991). *Life itself: A comprehensive enquiry into the nature, origin and fabrication of life*. New York: Columbia University Press.

14. Schrödinger, E. (1948). *What is life?* Cambridge, UK: Cambridge University Press.

15. Simondon, G. (1989). *L'individuation psychique et collective.* Paris: Aubier.

16. Stewart, J. (1993). Cognition without neurones: Adaptation, learning and memory in the immune system. *CC-AI, 11*, 7–30.

17. Stewart, J. (1996). Cognition = life: Implications for higher-level cognition. *Behavioural Processes, 35*, 311–326.

18. Varela, F. (1979). *Principles of biological autonomy.* New York: North Holland/Elsevier.

19. Varela, F. J. (1997). Patterns of life: Intertwining identity and cognition. *Brain and Cognition, 34*, 72–87.

20. Varela, F. J., & Frenk, S. (1987). The organ of form: Towards a biological theory of shape. *Journal of Social Biology and Structure, 10*, 73–83.

21. Varela, F. (2000). *El fenómeno de la vida.* Santiago: Ensayo.

22. Wheeler, M. (1997). Cognition's coming home: The reunion of life and mind. In P. Husbands & I. Harvey (Eds.), *Proceedings of ECAL '97.* Cambridge, MA: MIT Press. ftp://ftp.cogs.susx.ac.uk/pub/ecal97/online/F035.ps.gz.

23. Zaretzky, A. N., & Letelier, J. C. (2002). Metabolic networks from (M,R) systems and autopoiesis perspective. *Journal of Biological Systems, 10*, 265–284.

Appendix

The tesselation automaton, both in its original version [18] and in the 3D version presented here in Section 2.1, is a discrete cellular automaton with purely local rules that specify computer simulations. Assuming that the numbers of all components and events are large, an *approximate* theoretical treatment in terms of differential equations is possible as follows. Let us consider first the dynamics of the quantity $c_M(t)$ of the component C in the membrane. $c_M(t)$ belongs to the interval $[0, c_1]$, where c_1 is the maximal quantity of C when there are no holes. We have

$$\frac{dc_M}{dt} = -k_c c_M + k_b b_M (c_1 - c_M) \tag{1}$$

The first term represents the constitution of holes by $C \rightarrow D$. The second term represents the reconstitution of the membrane by $B \rightarrow C$, because the probability of an encounter of B with C is proportional both to the concentration b_M of B in the layer just under the membrane, and to $c_1 - c_M$, the "concentration" of holes. Note that we suppose here that the encounter between a free B molecule and a hole in the membrane always succeeds in reconstituting the membrane, which is equivalent to supposing that the membrane is unconditionally impermeable to the escape of B: we call this *model I.* If the holes become bigger [with a distribution $f(D, c_M)$ for the sizes D of the holes], then the membrane becomes partially permeable to B, and k_b becomes dependent on c_M. This situation, *model II*, can be modeled by modifying Equation 1:

$$\frac{dc_M}{dt} = -k_c c_M + k_b b_M (c_1 - c_M) p_r(c_M) \tag{1'}$$

where $p_r(c_M)$ is the probability that a B component that hits the region of a hole will attach to the edge of the hole (i.e., the membrane is partially repaired), rather than going straight through (and being lost).

The volume enclosed by the membrane is a distributed system of discrete elements $\{A, B\}$, which move by random walk in the manner of Brownian motion. It is known that random-walk processes can be approximated by partial differential equations for the local concentrations $a(x, t)$ of A and $b(x, t)$ of B, the latter belonging to $[0, b_1]$, where b_1 is the maximal concentration of B inside the membrane:

$$\frac{\partial a}{\partial t} = \Delta a - 2k_a a_M^2 c_M (b_1 - b_M)\delta_M \tag{2}$$

$$\frac{\partial b}{\partial t} = \Delta b + k_a a_M^2 c_M (b_1 - b_M)\delta_M \tag{3}$$

In these equations, the first term represents the diffusion due to random walk (with the Laplacian operator Δ). The second term represents the reaction $A + A \rightarrow B$ catalyzed by C. The probability of this reaction is (i) proportional to a_M^2 (where a_M is the concentration of A under the membrane) corresponding to the encounter of two A molecules, (ii) proportional to c_M corresponding to the encounter with a free catalyst C, and (iii) proportional to $b_1 - b_M$, because we suppose that there is an upper limit b_1 to the possible concentration of B, corresponding to saturation. δ_M is a characteristic function for a thin layer just underneath the membrane [$\delta_M(x) = 1$ if but only if x is at a distance less than δR; otherwise $\delta_M(x) = 0$].

By integration of the partial differential equations over the whole space of the automaton, we obtain the macrodynamics of the total quantities of A and B, that is, $V\langle a\rangle$ and $V\langle b\rangle$, where $\langle a\rangle$ and $\langle b\rangle$ are the mean concentrations of a and b. Note that the integration of the Laplacian in Equation 2 gives the net flux of A across the membrane, $\alpha(a_1 - a_M)S$ in Equation 4; note also that the integration of the Laplacian in Equation 3 gives the net output flux of B, $k_b b_M(c_1 - c_M)$ in Equation 5, which corresponds exactly to the creation of new C molecules in Equation 1.

We thus have

$$V\frac{d\langle a\rangle}{dt} = \alpha(a_1 - a_M)S - 2k_a a_M^2 c_M(b_1 - b_M)\ \delta V \tag{4}$$

$$V\frac{d\langle b\rangle}{dt} = -k_b b_M(c_1 - c_M) + k_a a_M^2 c_M(b_1 - b_M)\ \delta V \tag{5}$$

where

S is the total surface of the membrane (and is proportional to c_1),

V is the total volume enclosed by the membrane,

$\delta V = S\ \delta R$ is the volume of the thin layer just below the membrane,

a_1 is the concentration of A in the external environment and is supposed to be fluctuating with time around a mean value $\langle a_1\rangle$,

a_M and b_M are the concentrations of A and B in the thin layer and can also change with time.

In order to characterize the qualitative behavior of this system mathematically, the procedure we propose is to examine the equilibrium conditions obtained by putting Equations 1, 4, and 5 equal to zero, and assuming that a_1 (the concentration of A in the external environment) is constant. On this basis, it is then possible to infer the effects of statistical fluctuations in a_1 and in the local processes occurring in the membrane (formation and repair of holes).

A very rough preliminary treatment involving additional simplifying assumptions (not shown here) tentatively suggests the following conclusions. With model I, the system appears to have two fixed points: PF0 ($b_M = c_M = 0$) and PF1, with nonzero values for both b_M and c_M. However, PF0 is an unstable equilibrium; if the initial conditions are such that b_M and c_M are nonzero, however small, the system will move to the stable attractor at PF1. In other words, depending on the parameter values and the value of a_1, the system can exist whatever the value of c_M. We comment on this result in the main part of the article.

With model II, the system appears to have three fixed points:

- PF1, with b_M and c_M both high. This is a stable attractor, quite analogous to PF1 in Figure 2.
- PF0, with b_M and c_M both equal to zero. This is similar to PF0 in Figure 2, but with the difference that it is now also a stable attractor.
- PF2, with b_M and c_M at critical intermediate values. This new fixed point PF2 is an unstable point of bifurcation, any perturbation leading either to PF1 or to PF3.

The robustness of the system clearly depends on the distance between PF1 and PF2 in phase space, and this depends on the parameters. The distance increases with k_b and decreases with increasing k_c; it also increases with k_a and a_1. For unfavorable values of the parameters, and/or low values of a_1, the points PF1 and PF2 both disappear and the system inevitably disintegrates to PF0. We comment on the qualities of this model in the main part of the article.

We are now in a position to discuss the effect on the system of fluctuations in a_1. It is convenient to distinguish between rapid fluctuations and long-term changes. In the case of rapid fluctuations, because of the Laplacian terms in Equations 2 and 3, the effects of fluctuations in a_1 are buffered by the reserves of A and B molecules accumulated in the whole internal volume of the automaton. Thus, to a very good approximation, a_1 can be replaced by its mean value $\langle a_1 \rangle$, and the analysis is not affected; this is the case for both models.

In the case of slow long-term changes, the system as a whole will come into equilibrium at each point in time. The effects of fluctuations are now different for the two models. In the case of model I, which does not take account of the loss of B molecules through holes, a prolonged reduction in a_1 will of course lead to a reduction in b_M and c_M; but the system will never disappear, and whenever a_1 returns to more favorable levels, the system will be restored. In the case of model II, with a loss of B molecules through holes, the situation is not the same. Here, if the prolonged reduction in a_1 leads the system below the bifurcation at PF2 (or to the conjoint disappearance of both PF2 and PF1), the system will collapse entirely on the basis of its own dynamics, and cannot be resuscitated whatever the subsequent increase in a_1. These results are also discussed in the main part of the article.