Autonomy: an information theoretic perspective

Nils Bertschinger, Eckehard Olbrich, Nihat Ay, Jürgen Jost

Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, D 04103 Leipzig, Germany

Abstract

We present a tentative proposal for a quantitative measure of autonomy. This is something that, surprisingly, seems to be missing from the literature, even though autonomy is considered to be a basic concept in many disciplines, including artificial life.

We work in an information theoretic setting for which the distinction between system and environment is the starting point. We propose as a measure for autonomy the conditional mutual information between consecutive states of the system conditioned on the history of the environment. This works well when the system cannot influence the environment at all. When, in contrast, the system has full control over its environment, we should rather neglect the environment history and simply take the mutual information between consecutive system states as a measure of autonomy.

This discussion reveals some subtle facets of the concept of autonomy, ultimately leading towards the ambitious issue of autonomy as a fundamental (defining) concept of life itself. In our formalization, these aspects question the seemingly innocent system-environment distinction we took for granted and raise the issue of the attribution of control, i.e. the responsibility for observed effects. To further explore these issues, we evaluate our autonomy measure for simple automata, an agent moving in space, gliders in the game of life, and the tessellation automaton for autopoiesis by Varela et al.

Introduction

Autonomy is a central concept in many disciplines. Autonomy might mean the freedom of a system to set its own goals, to construct its own rules of operation, or to select the methods for achieving its aims according to some internal procedure or set of rules that is shielded from the control of the environment the system happens to be situated in. In the context of biological systems, autonomy denotes a qualitative difference between living and non-living systems in the framework of autopoiesis and organizational closure. For cognitive systems, in a constructivist framework, it refers to the ability to employ new distinctions and generate meaning in the system. Finally, in the context of psychic and social systems the concept of autonomy leads to more specific notions such as "intentionality" or "agency".

A slightly different meaning of autonomy is encountered in the context of "autonomous robots". In this case the desired feature is restricted autonomy of an artificial system which cannot be controlled directly by human operators — restricted, because the final goals should be given externally. But despite the fact that autonomy is a crucial notion in many fields, including artificial life, surprisingly, there does not seem to exist a good quantitative measure for the degree of autonomy that a given system possesses.

Our aim is to develop quantitative measures of system autonomy based on an information theoretic approach. For this we start with a tentative distinction between system and environment by selecting observables to describe states of the system and the environment, respectively.

By estimating the autonomy measure, one can then control this selection of observables — one has identified an autonomous system consistently if the measure becomes positive.

In this paper, after introducing the basic set up and the notation, we shall propose a tentative measure of autonomy using information theoretic quantities and discuss some examples including simple automata, an agent moving in space, gliders in the game of life (Beer, 2004), and the tessellation automaton for autopoiesis (Varela et al., 1974). We shall also discuss the relationship between autonomy and the different notions of closure which one encounters in systems theory: autopoiesis as organizational closure (Maturana and Varela, 1980), closure to efficient causation (Rosen, 1991) or operational closure (Luhmann, 1995) by relating these concepts to informational closure which can be defined using conditional mutual information and is particularly important for cognitive systems.

System and Environment

We consider the following setting: a system with states S_n interacts with its environment E_n through the channels $\hat{s}$ and $\hat{e}$ according to Fig. 1. $\hat{s}$ denotes the actions or the behavior of the system in its environment, and $\hat{e}$ the actions of the environment on the system. For the sake of simplicity, we restrict ourselves to discrete time and discrete states. Thus
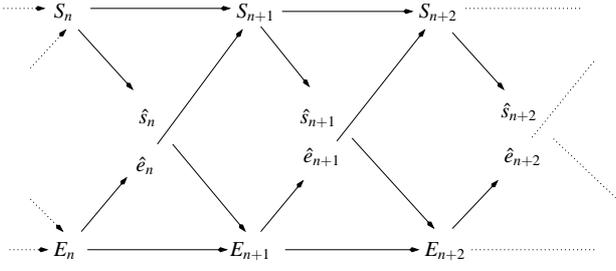
Figure 1: *The system S and the environment E interact through the channels ŝ and ê. The figure shows the temporal dependencies of this interaction.*

the system and the environment are represented by a pair of coupled hidden Markov models. In particular, S_n and E_n are conditionally independent given E_{n-1} and S_{n-1}. Here, we do not specify how to choose the observables for the environment and the system. Instead, we assume that subsequent analysis will lead to criteria for a useful choice.

Entropies and Information

Some notation: the entropy of the probability distribution of the random variable A, assuming the value a_i with probability $p(a_i)$, is

$$H(A) = -\sum_{i=1}^{N} p(a_i) \log_2 p(a_i). \qquad (1)$$

For two random variables A and B, the entropy of the joint probability distribution $p(A,B)$ is $H(A,B)$, and the entropy of the conditional probability of A given B is

$$H(A|B) = H(A,B) - H(B). \qquad (2)$$

The mutual information between A and B is then

$$MI(A;B) = H(A) - H(A|B), \qquad (3)$$

and the conditional mutual information between A and B given C is

$$MI(A;B|C) = H(A|C) - H(A|BC), \qquad (4)$$

measuring the reduction of the uncertainty of A given B under the condition that also C is known. This can be interpreted as the "information flow" from B to A given C (or in the other direction).

Non-Heteronomy

Our starting point in the discussion of autonomy is that an autonomous system should not be determined by the state history of the environment, i.e. that the system is not heteronomous. In our setting this would mean that

$$H(S_{n+1}|E_n, E_{n-1}, \ldots, E_{n-m}) > 0, \qquad (5)$$

i.e. there is a remaining uncertainty for the state of the system given the history of the environment. If we can observe only the behavior ŝ of a system one could alternatively consider

$$H(\hat{s}_{n+1}|E_n, E_{n-1}, \ldots, E_{n-m}) > 0. \qquad (6)$$

as a measure of non-heteronomy of the system's behavior. There are, however, some problems with (5) as a condition for non-heteronomy:

1. The condition depends on the history of length m. If the system came into existence at a definite time in the past, then extending the past beyond this time could reduce the entropy to zero even though we want to consider the system as autonomous on the basis of its present behavior.

2. The condition should quantify to what extent the system is not determined by the environment. If, however, the system can influence the environment this can also reduce the conditional entropy. In the extreme case of full synchronization between system and environment, one cannot tell within our information theoretic framework whether the system controls the environment or the environment controls the system. In order to make use of condition (5) we should consider environments that are, at least, not fully determined by the system. We shall return to this issue.

3. A related, but not identical, issue is the following: an autonomous system should still be adaptive, i.e. it should react to the environment in an in some sense "optimal" way. So, if one knows what is optimal for the system in a certain environment, one might predict the action, behavior or state of the system from the environment. Thus the condition Eq. (5) might not be fulfilled. Thus we require that the system is capable of pursuing different objectives in the same environment.

Self-Determination

As a second constraint on autonomy for a system, we should exclude completely random behavior from being considered autonomous. Therefore we require that not only is the state of the system NOT fully determined by the history of the environment, but also that the state of the system IS determined, to some extent, by the previous state of the system. In fact, this can be viewed as an abstract representation of the fact that the system sets its aims by itself. The aims have to be represented in some way in the state of the system.

These considerations lead to a tentative definition: a system is called autonomous if

$$A = MI(S_{n+1}; S_n | E_n, E_{n-1}, \ldots, E_{n-m}) > 0. \qquad (7)$$

Conditioning on the environment excludes that the "memory" (mutual information between subsequent states) is induced from the environment, reflecting only correlations

within the environment. Autonomy defined this way implies non-heteronomy (5), because

$$A = H(S_{n+1}|E_n, E_{n-1}, \ldots, E_{n-m}) - $$
$$H(S_{n+1}|S_n, E_n, E_{n-1}, \ldots, E_{n-m}) = 0 \quad (8)$$

if the system is heteronomous.

There is a problem with our tentative measure (7) that was already mentioned above: if the system can influence the environment then the state of the environment can be used to predict the state of the system. Thus, the measure indicates reduced autonomy, which clearly contradicts our intuition; control of the environment should increase and not decrease autonomy.

As discussed above, this problem comes from the strong assumption that the system cannot control the environment at all, i.e. all of the mutual information that an observation of the environment $E_n, \ldots, E_{n-m}$ provides about the system state S_{n+1} effectively reduces possible autonomy of the system measured by the entropy of S_{n+1}. We consequently use $H(S_{n+1}) - MI(S_{n+1}; E_n, \ldots, E_{n-m}) = H(S_{n+1}|E_n, \ldots, E_{n-m})$ as a measure for non-heteronomy (see Eq. 5).

At the other extreme, if we make instead the assumption that all mutual information between the system and the environment is attributed to the system, then we only consider as non-heteronomy the reduction in the uncertainty about the system state arising from environmental fluctuations that cannot be explained from the past system state. We can then derive the following autonomy measure:

$$A^* = \underbrace{H(S_{n+1}) - MI(S_{n+1}; E_n, \ldots, E_{n-m}|S_n)}_{\text{Non-heteronomy under the assumption of S controlling E}}$$
$$-H(S_{n+1}|S_n, E_n, \ldots, E_{n-m})$$
$$= H(S_{n+1}) - H(S_{n+1}|S_n)$$
$$= MI(S_{n+1}; S_n) \quad (9)$$

This measure simply states that a system is autonomous if its next state can be predicted from the present one, i.e. it reflects how much the system is in control of its own dynamics. The dependence on the environment drops out since only un-controlled (actually unpredicted) influences from the environment are considered that appear as noise to the system. The influence from the environment therefore just leads to non-determinism of the system's internal dynamics. In general, we should expect an intermediate situation, i.e. a bidirectional interaction between system and environment. We shall see in the examples, however, that there exist situations where the assumption, of an environment that is not influenced by the system, is quite appropriate. At least the system–environment distinction can be made in such a way that this assumption is fulfilled to a large extent.

Autonomy and Closure

How are the notions of autonomy and closure related? To answer this question in our framework, we have to formal-

ize "closure". One possibility is to consider "informational closure" (Bertschinger et al., 2006), which refers to the "information flow" from the environment into the system. This information flow is defined as

$$IF(E \rightarrow S) = MI(S_{n+1} : E_n|S_n) \quad (10)$$
$$= H(S_{n+1}|S_n) - H(S_{n+1}|S_n, E_n) \quad (11)$$
$$= H(E_n|S_n) - H(E_n|S_n, S_{n+1}) . \quad (12)$$

and quantifies the amount of information that the environment additionally provides about the next state of the system given the previous state of the system. A simple possibility to achieve informational closure is to decouple system and environment, i.e. make them independent. We are interested in systems interacting with their environments, hence in non-trivial informational closure $NTIC$, which might be measured by

$$NTIC = MI(S_{n+1} : E_n, \ldots, E_{n-m})$$
$$-MI(S_{n+1} : E_n, \ldots, E_{n-m}|S_n) , \quad (13)$$

with $MI(S_{n+1} : E_n, \ldots, E_{n-m}|S_n) = MI(S_{n+1} : E_n|S_n) = IF(E \rightarrow S)$ due to the Markov property of our setting (Fig. 1).

This quantity is maximized when the information flow into the system is minimal but the mutual information between the system and the environment maximal. It measures the extent to which the system models its environment.

We should remark here that the concepts of organizational or operational closure are not captured by the informational closure.

There is an interesting relationship between the two measures of autonomy developed above, and the non-trivial informational closure (13):

$$A^* = A + NTIC . \quad (14)$$

If $A^* > A$ the difference can be interpreted as the amount of information about the correlations in the environment modeled by the system.

Examples

Simple Automata

The simplest autonomous system is a system that does not interact with its environment at all. This is not what we are interested in, however.

The simplest systems interacting with the environment are ones with two states, i.e. one binary observable $S_n = \{A, B\}$, and input from the environment likewise described by a binary observable $E_n = \{0, 1\}$. Hence there is no influence from the system on the environment and (7) applies.

For the deterministic automaton (a) in Fig. 2, the state S_n is determined by the state of the environment at the previous time step E_{n-1}. It follows that $H(S_{n+1}|E_n) = 0$, the system
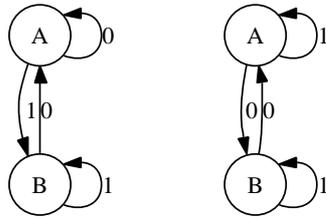
a) non-autonomous   b) autonomous

Figure 2: *Two deterministic automata*

is heteronomous and our autonomy measure is zero.

In the second automaton (b), the state of the environment determines whether the system changes or retains its state. If we initialize one copy of the system with $S_0 = A$ and a second copy with $S_0 = B$ the two copies will remain in different states forever given the same input. Therefore $H(S_{n+1}|E_n) = 1 > 0$. However, if we know S_{n-1} then we can perfectly predict S_n, which implies $H(S_{n+1}|E_n, S_n) = 0$. Thus the system has positive autonomy $A = 1$ bits.
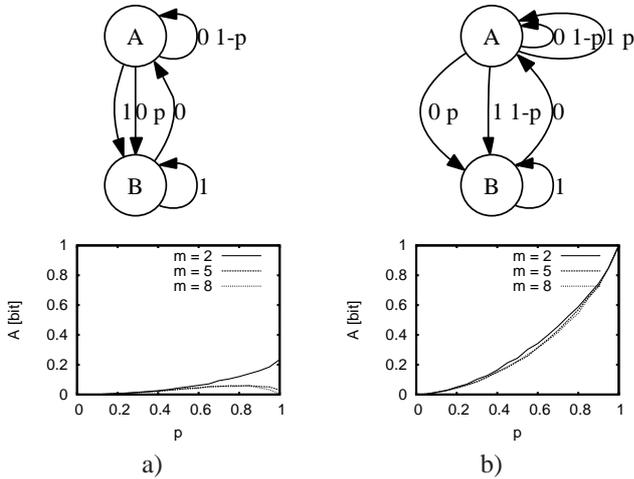

Figure 3: *a) Transition from a non-autonomous to a non-autonomous automaton; b) Transition from a non-autonomous (p = 0) to an autonomous (p = 1) automaton*

Up to this point, we have considered deterministic automata with $H(S_{n+1}|S_n, E_n) = 0$. What happens in the non-deterministic case? To investigate this question we consider two cases by varying a probability p between 0 and 1, where the limiting cases are deterministic automata: In the first case, both deterministic automata are non-autonomous (a) in Fig. 3), whereas in the second case, we started from a non-autonomous deterministic automata and changed it into an autonomous one (b) in Fig. 3). In the latter case, we observe, as expected, a monotonic increase of our autonomy measure. The autonomy measure seems to converge very fast with the history length m. The first case is more inter-

esting. In this case we know that the autonomy measure in the two deterministic situations has to vanish. For $p = 1$, however, it is exactly zero only in the limit $m \to \infty$. In the simulations, it effectively vanishes already for $m = 8$. For $0 < p < 1$, our simulations yield positive values for the autonomy measure with a maximum at $p \approx 0.8$. How can we understand this result? If the input is $E_n = 1$ then the state of the automaton is determined as B. Thus even after an input sequence $(\ldots, 1, 0)$, the state of the automaton is determined as A. Only after observing more than one 0, the state becomes undetermined and $H(S_{n+1}|E_n, \ldots, E_{n-m}) > 0$. So let us assume we have observed $(\ldots, 1, 0, 0)$ as inputs. Then $p(S_{n+1} = A) = 1 - p$ and $p(S_{n+1} = B) = p$. Does knowledge about S_n provide additional information about S_{n+1}? In fact, it does. If we know that $S_n = B$ then we can conclude that $S_{n+1} = A$ leading to a positive autonomy measure A. Thus we get the somewhat surprising result that "mixing" randomly two non-autonomous automata leads to an autonomous one. The introduction of a random "decision" in state *A* generates the autonomy of the automaton. When we introduced the autonomy measure however, we intended to exclude random behavior from being called autonomous. The main point is that the "decision" only occurs if the system is in state *A*, i.e. it is not totally random: whereas the outcome of the decision is random the system controls at which time the decision will be made. According to our autonomy measure Eq. (7) this is enough to generate autonomy. It is crucial, however, that the randomness is attributed to the system. If we introduced an additional observable in the environment that would determine the outcome of the "choice" between the two automata the system would be non-autonomous again.

A moving agent

Let us consider an agent moving on a grid space. Since the structure and the dimension of the space is not really important at this point, we take a 2-d rectangular lattice for the sake of simplicity. The agent can select one of the four adjacent points to move on.

First we have to define the environment observables E_n and the system observables S_n. The environment observables are the positions $E_n = (x_n, y_n)$, which emit $\hat{e}_n$, e.g. the concentration of some chemical. The movements of the system $\hat{s} = \{$up,down,left,right$\}$ can be viewed as the outputs of a hidden Markov model with internal states S_n that control the selection. Let us distinguish three cases:

A) Random selection: The system has only one state, the actions are selected according to some fixed probabilities.

B) Environment dependent selection: The action probabilities depend on the inputs from the environment $\hat{e}_n$ via the selection of different internal states, i.e. $S_{n+1} = F(\hat{e}_n)$. The movements $\hat{s}$ of the agent, however, determine partially the next inputs, for instance one can think of chemo-

taxis, i.e. the lattice sites might be occupied by some chemical and the agent prefers moving in the direction of occupied sites.

C) State dependent selection: In the most general case, the agent adopts an internal state that depends both on the previous state and the inputs from the environment.

Whereas in case A) no autonomy is possible because $H(S_{n+1}) = 0$ and in case C) all situations are possible, case B) is of particular interest for our purposes. Because the state of the agent is a function of the inputs of the environment, the non-heteronomy measure (5) vanishes and therefore also the autonomy measure A is zero. But the measure A^* might be non-zero, which would reflect some structure in the environment, e.g. a concentration gradient. But is this autonomy? If we could observe two agents of the same type (with the same internal structure) - one moving in the direction of the gradient and a second one performing a random walk, we would attribute to the first one the goal of finding a cite with high concentration, whereas the second one obviously has some other goal. This, however, corresponds already to case C), because the different goals have to correspond to different internal states of the agent.

Gliders in the Game of Life

In (Beer, 2004) Randall Beer presented an instructive discussion of the concepts from autopoiesis theory with the example of gliders in the game of life. The game of life is a two-dimensional deterministic cellular automaton where the cells can have two states, live and death, and gliders are specific moving patterns in this automaton. If we assign the position of the glider to the environment as in the case of a moving agent, the minimal organization of the glider as considered in (Beer, 2004) has four internal states. If we additionally identify states that only differ by mirror and rotation symmetries the number of internal states is reduced to two. These two states correspond to two different configurations of "living" cells. If there are no living cells in the environment of the glider these two or four states, respectively, are run through cyclically introducing a phase like internal degree of freedom.

Applying our measure of autonomy (7), we have

$$A = H(S_{n+1}|E_n, \ldots, E_{n-m}) . \tag{15}$$

The second term vanishes because the game of life is a deterministic automaton. To measure the autonomy of the glider we have to study to what extent the state of the glider can be controlled by the environment. The main problem, which one immediately encounters, is, however, that interaction with almost all stable structures that are common in the game of life is destructive for the glider, at least for one internal state of the glider. Thus no system observables remain to estimate the autonomy. This leads us to an interest-

ing conclusion: whereas our measure quantifies the **behavioral autonomy** of an existing system, it cannot quantify the **basic autonomy**[1] that is manifested by the mere existence of a system. By including, however, the situation of disintegration as a "death" state in the set of system observables we can use our autonomy measure (7) also to quantify basic autonomy. In the simplest case one considers only two states of the system, live and death, and two possible environments, deadly and non-deadly. Thus the instability of the glider is expressed as a very low basic autonomy.

Varela's Tessellation Automaton

The tessellation automaton presented in (Varela et al., 1974) was intended to illustrate the basic ideas of the theory of autopoietic systems in a simulated model system. The model is a cellular automaton living on a square lattice. In contrast to the game of life, this cellular automaton has stochastic update rules. It represents the following situation: substrate particles diffuse freely on a lattice. In the neighborhood of a catalyst, two of them might form a new particle called "link". These "links" have the ability to bond to each other. A chain of bonded particles forms then a "membrane" that is permeable for the substrate, but not for the "links". The links, either bonded or free, can decay into the substrate again with some probability. If the bonded links form a cavity including the catalyst this is regarded as an autopoietic unity in (Varela et al., 1974) because the higher concentration of links inside the cavity allows for a high probability of spontaneously occurring "self-repair" of the membrane. A more recent reevaluation of this model (Mullin and Varela, 1997) showed that this interpretation depends crucially on the details of the model and one has to include an interaction (chain based bond inhibition) not mentioned in the original paper. Moreover, one could also criticize that for real autopoiesis also the catalyst should be reproduced by the system. This is, however, not our concern in this paper. We are interested to what extent the model is autonomous according to our autonomy measure (7). We here present a qualitative discussion only.

The main difficulty, which exemplarily occurs in this case, is to define a suitable system – environment distinction with the corresponding observables. At this moment we are not able to provide a general method to perform this task, because this would require to identify the organization of the system (i.e. the autopoietic organization for an autopoietic system) algorithmically, which, at least implicitly, would solve the problem of a formal and operational definition of autopoiesis, which is not available by now.

Therefore we have to start with some plausible, but not necessarily unique, distinction. If there exists a closed chain of bonded links, we would call the inside of this "membrane"

---

[1] This notion was inspired by (Ruiz-Mirazo and Moreno, 2004). They used, however, basic autonomy in a more specific and elaborate way.

including the membrane itself the system and the outside the environment. The corresponding states are determined by the type and position of the different particles in- or outside the membrane. In particular the membrane has to enclose the catalyst, otherwise we do not expect the system to be stable, i.e. to possess a minimal basic autonomy. The fuzziness of the system–environment distinction results from the fact that the membrane can decay which, in 2-d, results in non-connected parts. It is not clear whether such a configuration should still be considered as a single system with an inside and an outside. One possible criterion would be the ability of the membrane to maintain a concentration gradient of unbonded links between its inside and outside.

In order to make some qualitative statements about the autonomy of the model according to our autonomy measure we can adopt a pragmatic position because these statements do not depend on the details of the state definition or system definition, respectively.

(1) There are no obvious feedback loops through the environment affecting the state of the system, therefore Eq. (7) should be applicable.

(2) The state of the environment does not, at least not fully, determine the state of the system, i.e. the system is not heteronomous.

(3) There are correlations between subsequent system states that are not caused by correlations in the environment, i.e. knowing the previous state of the system gives additional information about the actual state of the system compared to only knowing the environment history.

Consequently, our autonomy measure of the system is positive. In fact, we see here a slight extension of the basic autonomy discussed above because the system states can be classified according to their viability into highly viable (intact membrane), less viable (small rupture, unbonded links nearby) and non-viable ones (large defects in the membrane) (see also the discussion of the simulations in (Mullin and Varela, 1997)).

Discussion

We proposed a measure that quantifies some important aspects of the intuitive notion of autonomy: that (1) an autonomous system should not be determined by its environment and that (2) an autonomous system should determine its own goals. We conceive of our measure as an interesting tool to quantitatively investigate simulations of models in artificial life. We found that if there is mutual information between the system and its environment the appropriate measure of autonomy depends on whether this mutual information is considered as caused by the environment (A) or by the system (A^*). This led to the problem of the system–environment distinction. In addition to the problem of defining the appropriate observables, i.e. state spaces, there is the problem of how to attribute the processes constituting the system dynamics (formally described by the transition kernel of the Markov process) between the system and the environment. This problem occurred already in the discussion of the randomly "mixed" simple non-autonomous automata giving rise to an autonomous one. There, the result depended on the attribution of random selection which was part of the transition kernel of this system. The problem, however, is more profound than that. The concepts of autopoiesis, operational closure or closure to efficient cause and the related concepts of autonomy are all concepts of self-referential closure and therefore self-maintained autonomy. That means that the systems achieve closure and autonomy with their own means. In our setting, however, the autonomy is essentially a property of the transition kernel of the coupled Markov models. How self-referential closure could be incorporated in our information–theoretic approach is still an open problem. Therefore it remains to be seen whether the quantitative notion of autonomy developed in this paper is compatible with the qualitative notions of autonomy used in systems theory by many authors e.g. (Varela, 1979; Rosen, 1991).

References

Beer, R. D. (2004). Autopoiesis and cognition in the game of life. *Artificial Life*, 10:309–326.

Bertschinger, N., Olbrich, E., Ay, N., and Jost, J. (2006). Information and closure in systems theory. to appear in the proceedings of the 7th German Workshop of Artificial Life, 26.-28.07.2006.

Luhmann, N. (1995). Probleme mit operativer Schließung. In *Soziologische Aufklärung 6*, pages 12–24. Westdeutscher Verlag, Opladen.

Maturana, H. R. and Varela, F. J. (1980). *Autopoiesis and Cognition. The Realization of the Living*. Reidel, Dordrecht.

Mullin, B. M. and Varela, F. J. (1997). Rediscovering computational autopoiesis. In Husbands, P. and Harvey, I., editors, *Proceedings of the Fourth European Conference on Artificial Life*, pages 38–47, Cambridge, MA. MIT Press.

Rosen, R. (1991). *Life itself: A comprehensive enquiry into the nature, origin and fabrication of life*. Columbia University Press, New York.

Ruiz-Mirazo, K. and Moreno, A. (2004). Basic autonomy as a fundamental step in the synthesis of life. *Artificial Life*, 10:235–259.

Varela, F. J. (1979). *Principles of Biological Autonomy*. North Holland, New York.

Varela, F. J., Maturana, H. R., and Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems*, 5:187–196.