

# Autopoiesis and Life

**Margaret A. Boden**

*University of Sussex*<sup>1</sup>

Life is defined by Maturana and Varela as a type of self-organization: autopoiesis in the physical space. This resembles the concept of metabolism, which itself is typically included in definitions of life. Three senses of metabolism are distinguished. If life depends on either autopoiesis or metabolism (in the third sense), then strong A-Life is impossible. The theory of autopoiesis challenges concepts familiar in biology and cognitive science. While its use of informational language is too restrictive, its use of cognitive language is too liberal: life does not imply cognition.

Keywords: life, self-organization, autopoiesis, metabolism, strong A-Life

Notoriously, there's no universally agreed definition of life. Certain features are typically mentioned: self-organization, emergence, autonomy, growth, development, reproduction, adaptation, responsiveness, evolution, and metabolism. Whether all of these features are essential to life is disputed. Reproduction and evolution are omitted from the list by some biologists, as we shall see. And metabolism is omitted (or considerably watered-down) by others, whose biological research consists largely of work in artificial life, or A-Life (Ray, 1992, 1994).

There is general agreement, however, that the core concept here is self-organization: the emergence (and maintenance) of order, out of an origin that is ordered to a lesser degree, by means of fundamental, and autonomous, structural development. Self-organization is the core concept because it necessarily involves some of the other items on the list, and because the remainder are all special cases of it. Thus emergence, autonomy, and development are cognate concepts, included in the definition. Growth, reproduction, adaptation, responsiveness, evolution, and metabolism are instances.

Sometimes, self-organization is defined by reference to energy. For instance, it is said that (all) self-organizing systems have to be energetically open systems, and that "it is [the] continual flux [of energy and matter] that is the wellspring of new forms" (Thelen & Smith, 1993, p. 54). But some-

---

<sup>1</sup> School of Cognitive and Computing Sciences, Brighton, BN1 9QH. ENGLAND.  
E-mail: maggieb@cogs.susx.ac.uk

times, it is defined more abstractly – as I have done above. In that case, the particular sort of self-organization typical of, or essential to, life must be identified.

In this paper, I discuss Humberto Maturana and Francisco Varela's (1980) definition of life as a specific type of self-organization: "autopoiesis in the physical space". Their reference to the physical space does not mean merely that every living thing must have some sort of material base. Rather, it means that metabolism is an essential property of life.

The next section briefly discusses the more familiar concept of metabolism, and explains why it is problematic for proponents of "strong" A-Life. Supporters of strong A-Life believe that virtual creatures existing in computer memory, and manifested on the VDU-screen, can be genuinely alive. Practitioners of "weak" A-Life, by contrast, use computer models to express and test theories about living things, but with no claim that the models themselves are really alive. These terms are directly comparable to John Searle's (1980) distinction between weak and strong AI. To reject the possibility of strong A-Life, as I do in this paper, is not to deny the interest of weak A-Life. On the contrary, A-Life modelling has already illuminated a number of biologically significant problems, including the emergence of global behaviour such as flocking and some mathematical properties of evolution in the abstract (Boden, 1996).

Then, I present the closely related concept of autopoiesis. Autopoiesis as such is not a biological concept, but an abstract description of self-organizing systems in general. Its most fully developed application, however, is to biological organisms. Although it resembles the more well-known notion of metabolism, it also differs significantly from it. For instance, it commits one to denying that reproduction and evolution are philosophically fundamental, or even empirically universal, aspects of life. This section also outlines a computer model of autopoiesis, and shows that an autopoietic approach implies that virtual A-Life (strong A-Life) is impossible even though other kinds of artificial life are not. The prime reason why autopoietic biology rules out strong A-life is that it stresses the self-constitution of the boundary of the living thing, as a unitary physical system. The final part of the section discusses possible doubts about the importance of the bodily boundary.

In the penultimate section ("Autopoiesis, Biology, and Cognition"), I show that Maturana and Varela's unorthodox biological approach is problematic for cognitive science in general. They use autopoiesis to argue that some of the basic assumptions of (most) biologists and psychologists are mistaken, and that some widespread theoretical vocabulary is inappropriate. However, one can accept their autopoietic version of the concept of metabolism while rejecting their position on the relation of life and cognition.

Finally, I show that if life is necessary for cognition (as is claimed by proponents of autopoiesis), then the possibility of strong AI depends on the possibility of strong A-Life.

### **A-Life and Metabolism**

Proponents of strong A-Life generally place little stress on metabolism – if they do not ignore it entirely. Either they insist that it is unnecessary to “life as it could be,” while of course admitting that it is a universal feature of “life as we know it.” Or they interpret metabolism in a watered-down way that might be satisfied by A-Life models. (Even here, they sometimes overestimate the range of A-Life systems that could conceivably be covered: see below.)

The first of these alternatives is not acceptable. It would be intellectually perverse – or, at best, question-begging – to drop the criterion of metabolism from the definition of life (Boden, 1999, Section 4). Besides its ubiquity in life as we know it, there is a very good reason for keeping metabolism in the definition. Namely, its explanatory power in relation to one of the most fundamental questions in biology: how living bodies organize and maintain themselves as integrated physical systems. Moreover, there is no independent reason for dropping it, beyond the desire (of some people) to save the philosophical coherence of strong A-Life.

As for the second alternative, I have shown elsewhere (op. cit., Sections 2 & 3) that metabolism can be defined in three ways, of which only the first two apply to (actual or hypothetical) virtual A-Life systems. The third, strongest, sense does not. It is the third sense which we find in biology, and which people have in mind when they define life in terms of metabolism. This third sense would apply to Martian creatures with an alien biochemistry just as it does to terrestrial carbon-based life. But it does not apply to virtual creatures.

#### **The three senses of metabolism**

In the first sense of the term, metabolism denotes energy dependency, as a condition for the existence and persistence of the living thing as *that* particular physical unity. (Mountains and chairs depend on energy for their existence too, but not in a way that actively distinguishes one individual from another.)

If this were the (literally) vital criterion, then strong A-Life would be possible. For as Tom Ray – a professional botanist who has done influential work in the computer modelling of coevolution (Ray, 1992, 1994) – points out, A-Life “creatures” are utterly dependent on the energy involved in the electronic processes that constitute and manifest each individual creature. However, this sense of metabolism is somewhat artificial (the pun is intended). It is rarely if ever used except by people, such as Ray, trying to de-

fend the thesis of strong A-Life. *A fortiori*, it is not what people normally have in mind when they claim that metabolism is a criterion of life.

The two stronger senses of metabolism each involve ideas about using, collecting, spending, storing, and budgeting the energy on which the system is dependent. Thus the second adds the notion of individual energy packets used to power the activities of the creature, its physical existence being taken for granted. It assumes that each living system has assigned to it, or collects for itself, a finite amount of energy – which is spent as it carries out its activities. The third concept of metabolism, by contrast, does not take the physical existence of the system for granted. On the contrary, it defines metabolism as the use, and budgeting, of energy for bodily construction and maintenance, as well as for behaviour. (By “behaviour,” here, I mean all the activities of the organism, both internal and externally visible: in this very general sense, plants as well as animals – and single cells, too – engage in behaviour.)

The second type of metabolism could characterize some A-Life systems. In other words, someone who is content to define metabolism in this way must allow that genuine A-Life is conceivable. The relevant systems could include certain types of robot (Boden, *op. cit.*, Section 2), and also certain types of A-Life simulation. The former are irrelevant here, since robots are not candidates for strong A-Life. But the latter are more to the point: using this interpretation of metabolism, some merely virtual creatures could be said to be alive.

However, there is a catch. Simulation as we think of it today would not be enough. One would need to identify distinct (virtual) energy-packets for each individual – and perhaps sub-packets, differentially devoted to specific drives such as food-seeking, fighting, and mating. So far, so good: such simulations already exist. But to satisfy the sense of metabolism in question here, these programmed simulations would have to be complemented by distinct (and finite) sources of real energy in the computer.

Each real-energy source would have to be dedicated to a single virtual individual, perhaps even to the various types of simulated activity, or processing, involved. (The virtual individual would not need to be localized in computer-memory, since “physical existence” does not imply a body.) This is very different from the current situation, in two ways. First, in today's simulations the energy-source of the computer is available indiscriminately for every process specified by the program. Second, each process, or programmed “individual,” calls on energy which, for practical purposes, can be regarded as infinite. (In this context, we may ignore limitations of computer-memory, and the combinatorial explosion.) There is no way, given today's computers, of pulling the plug or zapping an energy-providing component so as to cut out a virtual “Tibbles” rather than a virtual “Fido.” Such selective death (or its converse, revivification) can be effected only by the programmer, not by the computer engineer.

This science-fiction scenario, sketched with the second sense of metabolism in mind, takes the physical existence of the virtual creature for granted. In other words, it offers a form of strong A-Life that shows material organization but not genuine self-organization. The electronic processes and (dedicated) energy-sources that constitute the physicality of the simulated Tibbles or Fido are organized, integrated, and maintained (and, if necessary, repaired) not by Tibbles or Fido themselves, but by the human engineer.

Clearly, then, the second sense of metabolism is not the biologist's concept of it. For no biologist takes the existence of a creature's body for granted. On the contrary, one of the prime puzzles of biology is to explain how organisms come into existence, and how they are maintained until the individual dies. We therefore need the third definition of metabolism – the use, and budgeting, of energy for bodily construction and maintenance (and behaviour) – if we are to capture what biologists normally mean by the term.

### **Metabolism and biochemistry**

Metabolism of this third type inevitably involves closely interlocking biochemical processes, to “engineer” the organism's self-maintenance, growth, and activity. Because of the unavoidable tendency to disorder (i.e. the second law of thermodynamics), metabolism must involve continual energy-intake from the outside world. And that, in turn, will engender further chemical processes within the body.

Very simple living (or lifelike) systems might take their energy directly from the environment whenever they needed it, thus satisfying only the first sense of metabolism defined above. But such systems would be highly vulnerable to environmental conditions: if no energy were immediately available, they would die.

If (by chance) they became able to store excess energy, so as to use it later when needed, their evolutionary fitness would be increased. And it might be best stored in some other form, perhaps because of higher chemical stability (green plants use light energy, but they don't store it as light). If so, then biochemical processes would be needed to convert the input energy into the storage medium, and to change it into a usable form again when needed. These processes would be most unlikely to work without any side-products, of which some might find other biochemical uses and some would need to be excreted. In short, a complex integrated system of biochemical conversions and internal energy-budgeting would evolve.

Even the proverbial Martians – intelligent or otherwise – would depend on such metabolic processes of synthesis and breakdown, for none of this mentions carbon. Significantly, the “laws of bio-energetics” are couched in highly abstract terms, referring for example to convertible “currencies” for energy-storage. This is due neither to biochemical ignorance nor to any need for abbreviation, for (in all cases that have been examined) only three chemical substances are used as energy currencies by terrestrial life (Moran, Mo-

reno, Montero, & Minch, 1997). Polysyllabic though they are (the most common is adenosine triphosphate, or ATP), they could be listed in a single breath. The point, rather, is that any living thing – more precisely: any system that metabolises in the third sense – will show a form of self-organization that satisfies these general principles.

Clearly, strong A-Life is not playing this game. When its proponents speak of the physical existence of virtual creatures, it is electronic – not metabolic – existence that they have in mind. To be sure, A-Life research includes a number of simulations of metabolism, modelling specific biochemical processes and/or general metabolic categories. One of these is Steve Grand's computer-world *Creatures* (Grand, Cliff, & Malhotra, 1996), whose coded inhabitants alter their behaviour as a result of simulated metabolites coded as having varying concentrations, and simulated metabolic exchanges of various types (fusion, transformation, breakdown, decay, and catalysis). Grand himself believes his virtual creatures to be primitive life forms (p.c.). But they are not.

Simulations of biological (third-sense) metabolism are no more than that: simulations. Even if *Creatures* were to be implemented in the science-fictional “dedicated energy” computer mentioned above, there would be no real bodily self-organization involved, just a cybercopy of it. An A-Life researcher working in evolutionary chemistry and using real chemicals in real test-tubes might conceivably create real, metabolising, life – wittingly or otherwise (see the sub-section on “Computer modelling of autopoiesis”). But virtual metabolism is not metabolism, for the same reason that a virtual Tibbles, even one existing within an energy-budgeting computer, is not a living cat. Metabolism (in the third sense) is the use of energy-budgeting *for* autonomous bodily construction and self-maintenance, and no actual body-construction goes on in simulations of biochemistry.

As we shall see in the next section, the concept of autopoiesis also highlights autonomous bodily construction and self-maintenance. It is therefore very close to the biological concept of metabolism defined above. So it, too, is incompatible with the project of strong A-Life. But autopoiesis is not identical with metabolism, and is described by Maturana and Varela in a very different theoretical vocabulary. Moreover, as the section on “Autopoiesis, Biology, and Cognition” will show, it leads them to some biologically – and psychologically – unorthodox conclusions.

## The Concept of Autopoiesis

The theory of autopoiesis – etymologically: self-making – has been developed by the physiologists Maturana and (later) Varela over more than thirty years. Its most sustained expression is in their book *Autopoiesis and Cognition: The Realization of the Living*, first published in Spanish in 1972 and translated a few years later (1980).

The idea of autopoiesis is highly reminiscent of the strongest sense of metabolism distinguished in the previous section. More precisely, since autopoiesis is a purely abstract concept, which Maturana and Varela apply to many different examples of self-organization, metabolism is close to the particular type of autopoiesis they regard as characteristic of life.

In their terminology, this is “autopoiesis in the physical space” – which is to say, autopoiesis effected in physical systems, by physical (metabolic) processes. Metabolism is not part of the definition of autopoiesis as such, since autopoiesis is a more general concept. But (using their characteristically opaque vocabulary) it is constitutive of its material structure when autopoiesis is realized as a living thing (op. cit., p. 88).

Maturana and Varela take this form of autopoiesis (metabolic self-organization) to be the real essence of life. They see it as logically and biologically prior to most of the other vital properties in the “typical” list given in the Introduction, including reproduction, evolution, growth, responsiveness, and adaptation (see below). Their approach has been adopted by some researchers in A-Life and cognitive science. Unlike many of their colleagues, these researchers – some of whom are involved in the computer modelling of biological processes – unequivocally deny the possibility of strong A-Life. (Since they see cognition as necessarily grounded in life, they also deny the possibility of strong AI: see the concluding section.)

Autopoietic systems in general are defined in terms of their organization, not of their components nor even the properties of their components. What is crucial is “the processes and relations between processes realized through components” (Maturana and Varela, 1980, p. 75). An autopoietic system, or “autopoietic machine,” is formally defined by Maturana and Varela as follows:

[...] a network of processes of production (transformation and destruction) of components that produces the components which: (i) through their interactions and transformations continuously regenerate the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in the space in which they (the components) exist by specifying the topological domain of its realization as such a network (op. cit., p. 79).

The “machine,” here, is abstractly conceived, as is a Turing machine; and the “concrete” unity is not necessarily physical. The concept of autopoiesis as just defined can be applied (for example) to inorganic chemistry, to business organizations, or to whole societies. None of these is classed by Maturana and Varela as a living system, although the last two are higher-level systems whose lower-level components are living organisms. For life, as they define it, embodiment is required: “autopoiesis in the physical space [is] a necessary and sufficient condition for a system to be a living one” (op. cit., p. 84). And embodiment, in turn, involves more than mere physical existence. It

requires the self-creation of a unitary physical system by the spontaneous formation of a physical boundary.

In plainer English, their claim is that living things are of necessity physical, with a bodily fabric – and boundary – produced and maintained by themselves. The outstanding feature of living organisms is a form of self-organization termed autopoiesis. This, they point out, is a special case of homeostasis, where what is preserved is not one feature (such as blood-temperature), but the organization of the system itself as a unitary whole (op. cit., pp. 78-9). Some autopoietic systems have a self-maintained identity that does not exist in the physical space. A society, for instance, consists of organisms closely coupled not only by physical relations, but also by semantic (i.e. linguistically grounded) communications. The self-organization of a society is constituted by a self-coherent and self-sustaining set of social practices, within which there may be sub-systems of intercommunication having their own autopoietic unity. Different legal systems, for example, shape and maintain themselves within the specific communities concerned, and help establish equilibrium between various social and economic institutions (Teubner, forthcoming).

Only human organisms can form part of a society, so defined. But for all living creatures, the very boundaries of the living system as a physical unity, as well as its bodily components, are continuously produced by its own activities.

A human body, or a tree, is an autopoietic unity in the physical space. But they are higher-level autopoietic systems, made up of many such systems at a lower level (op. cit., pp. 107-9). The basic phenomenon here is not the formation of a body with arms and legs, or leaves and boughs, but the self-organization of a single cell. The generation of the cell-membrane both bounds and constitutes the cell as an autonomous vital entity, distinguishable from its environment. Explaining how this can happen is universally acknowledged to be one of the core problems of biology (Maynard-Smith & Szathmary, 1995, ch. 7). Maturana and Varela unequivocally identify it as *the* philosophically and scientifically fundamental problem.

### **Computer modelling of autopoiesis**

In the early 1970s, Maturana and Varela addressed the fundamental problem identified above by outlining a formal theory of autopoiesis (Varela, Maturana, & Uribe, 1974; see also McMullin & Varela, 1997). This formal account was further developed a few years later by Milan Zeleny, as a functioning computer model (Zeleny, 1977). Zeleny's computer model was an early exercise in (weak) A-Life – though certainly not intended as strong A-Life. It models the type of physical self-organization that Maturana and Varela claim is the (*sic*) essential property of life. Specifically, it simulates the emergence of a self-repairing membrane in a two-dimensional cellular automaton (Zeleny compared it with John Conway's "Game of Life"). That

is, it shows how a “cell” can be created, and simultaneously differentiated from its “environment”.

For understanding how such a system could in principle arise, one does not need to specify the chemicals involved in practice. Accordingly, Zeleny's molecules (like those simulated in *Creatures*, two decades later) are not identified as specific chemicals, but are defined in relational, functionalist terms. That is, they are classed as substrate molecules, catalysts, and three types of molecular link (free, single-bonded, and fully-bonded). Likewise, the chemical processes modelled are abstractly defined events such as molecular production, bonding, disintegration, and diffusion – whose rates can be varied by adjusting parameters in the rules.

The movements and interactions of the various components are governed by about twenty simple rules, although some molecular “meetings” occur partly by chance. The three molecular links have differing stability, but any link can (under certain conditions) disintegrate, forming two units of substrate. They also have varying potential for bonding. A free link can bond with a chain of bonded links; two chains of bonded links can be bonded into one, or re-bonded after their connecting link has disintegrated; and two free links can be bonded together to start a chain formation. A free link (and an empty “hole”) is produced whenever a catalyst interacts with two units of substrate. Disintegration and bonding can take place without catalysis, although bonding can occur only outside the catalytic neighbourhood. The role of the catalyst, given substrate molecules nearby, is to produce more complex component-linkages, which in turn are capable of bonding. Ultimately, they polymerize to generate a membrane whose self-closure both creates and bounds an enclosed space (Zeleny, 1977, p. 14).

This system is not only self-generating, but also self-maintaining. The (simulated) membrane is semi-permeable. It permits the diffusion of substrate molecules, because these are allowed to pass through a bonded link. Substrate can enter the space from the environment if there are holes in the space, adjacent to the membrane. By contrast, neither catalyst molecules nor free links can pass out through the membrane. These highly active (and, in the case of catalysts, relatively rare) components therefore stay trapped within the space. Moreover, the re-bonding rules cited above enable spontaneous repair of the membrane if it is ruptured at any point by disintegration. In other words, this is an autopoietic system.

More accurately, it is one of a large set of possible autopoietic systems, each produced by some variation in the rules determining potential interactions, or in the variables used in instantiating those rules (Zeleny, 1977, pp. 21-5). For instance, one can vary the number – and operational properties – of catalysts, their concentrations, or the size and shape of their “neighbourhoods”. One can also vary the rules regarding their interactions, with the result that catalytic activity in the space as a whole changes (perhaps rhythmically) through time. Similarly, one can introduce a wider variety of bonds,

or vary the properties, and amount, of the substrate(s). Different types of membrane, and different “cellular” behaviours, will be generated accordingly. (Some rule-variations, of course, will make autopoiesis less long-lasting, or even prevent its development altogether.)

### **Autopoiesis and A-Life**

Neither Maturana and Varela, nor Zeleny himself, regard Zeleny's computer simulation of life as real life. It is indeed an autopoietic system, but the autopoiesis does not take place “in the physical space”. To be sure, the system is actually implemented in the electronic processes of the computer. But at that level, there is no autopoiesis. The self-organization results only within an abstract representation of chemical relationships (e.g. catalysis) that are modelled by, not realized in, the computer. The same applies to Varela's more recent simulation of autopoiesis by means of an artificial chemistry (McMullin & Varela, 1997). In other words, the arguments given above to show that a simulation of metabolism is not really alive are paralleled by arguments showing that the type of autopoiesis characterizing (some) simulations is not the type of autopoiesis required for life.

That's not to say that Maturana and Varela deny the possibility of any conceivable sort of artificial life. On the contrary, they explicitly allow that one could in principle “design” and “make” a living system (op. cit., p. 114). They even remark that we may, unwittingly, already have done so. In saying this, however, they are not thinking of virtual creatures in computer memory, but of self-maintaining biochemical systems. Nor are they thinking of some marvel of nanotechnology, whereby individual bio-molecules are directly arranged by hand. Rather, they have in mind the human scientist's (perhaps unknowingly) “creating the conditions under which [autopoietic biochemical systems] form themselves” (Zeleny, 1977, p. 27).

(If nanotechnology were necessary to produce a particular biochemical system, even one that thereafter was self-sustaining, that system would not count as autopoietic. But suppose that a nanotechnologist deliberately aided the construction of a self-maintaining system by making chemical changes that could, indeed would, eventually have happened naturally. In that case, I would want to say that the resulting system was an autopoietic one, even though its self-construction had been aided by human intervention. Maturana and Varela, however, might view this intervention as adulterating – if not destroying – autopoiesis, not as aiding it.)

### **The importance of the boundary**

The similarity between the concept of autopoiesis and the third sense of metabolism is obvious. In particular, both refer to self-organized physical processes that generate, and maintain, the bodily fabric of living things. And Zeleny's computer model of autopoiesis could well have appeared in an A-Life journal, alongside other simulations of metabolic processes. Indeed, a

paper co-authored by Zeleny, on autopoiesis in (inorganic) chemistry, was published in the proceedings of the first A-Life conference (Zeleny, Klir & Hufford, 1989).

The main differences between metabolism and autopoiesis, and between accounts of life based on these concepts, result from the greater emphasis put by Maturana and Varela on the self-production of the organism's boundary as a unitary system. This is the core of their theory. They speak, for instance, of the „total subordination of [all the processes of change within] the system to the maintenance of its unity" (op. cit., p. 97). This perspective leads them to say a number of things about metabolism (which they call "energy-relations"), and about the concept of life in general, which most other researchers would not.

For example, because autopoietic systems are defined not by their components, but by their processes and the relations between their processes, the biochemical questions asked by Maturana and Varela are subtly different from those asked by most researchers into metabolism (Moran et al., 1997). They insist, for instance, that questions about the origin of life – the formation of the first cells – should be focussed not on molecules, but on the relations that molecules can have with one another (op. cit., p. 93). This is why the cell-membrane model described above does not specify any particular chemicals.

Admittedly, one does not need to cite autopoiesis in order to ask "relational" questions about the metabolism involved in the origin of life. For instance, Eric Drexler (1989) argues that even alien (non-carbon) biochemistries would have to share certain general properties with ours. They would have to employ general diffusion, not channels devoted to specific molecules; molecular shape-matching, not assembly by precise positioning; topological, not geometric, structures; and adaptive, not inert, components. These general characteristics might be instantiated in many different ways.

But even "relational" approaches may be radically criticized by theorists of autopoiesis. For example, Stuart Kauffman's (1969, 1971, 1992) abstract study of autocatalytic networks, widely cited in A-Life, is generally regarded as focussing on the metabolic origin of life. Maturana and Varela interpret this type of work differently (op. cit., pp. 93-4).

Seeing the self-organization of boundaries as an all-or-none phenomenon, they allow no intermediate stages between non-autopoietic (non-living) and autopoietic (living) systems. The first living system is the cell, brought into being by the formation of the cell-membrane. Autocatalytic networks, they insist, do not qualify as autopoietic systems, because they do not determine their own topology. The network's boundaries are set by something external, namely, the walls of the container in which the relevant chemical processes are taking place. When actual metabolic processes (some subset of an entire cell-metabolism) are reproduced *in vitro*, they do not constitute an

autopoietic system either. (This is not implausible: few people would say that a metabolic network functioning in a test-tube was an example of life.)

Similarly, even insofar as a virtual creature can be regarded as having a physical boundary, that boundary is not created and maintained by itself but (knowingly or not) by the computer engineer. The notion of boundary, in any event, is inappropriate here. The physical existence of most virtual creatures is not a unitary physical system, but a host of electronic processes scattered across an ever-changing set of memory locations. If a particular part of the memory-hardware were to be dedicated to an individual creature, that would be an arbitrary restriction, contributing nothing to the nature or unity of the creature's behaviour. To be the electronic implementation of a model of an autopoietic system is not to have a body, or (in other words) to be an autopoietic system in the physical space.

### **Some doubts about the boundary**

Someone not already committed to autopoietic theory – in particular, someone committed to the alternative approach of neo-Darwinism – might ask why one should accept that the maintenance of the bodily boundary is the fundamental achievement of life. Why should we follow Maturana and Varela when they say that everything that goes on in a living thing is “totally subordinated” to this task? After all, altruistic behaviour is explained by sociobiologists in terms of preserving gene pools distributed across many organisms, not the living body of the individual organism itself. And suicide hardly seeks to maintain the organism's bodily boundary.

Part of the reason for regarding the maintenance of the bodily boundary as fundamental to life is that (as shown in the next section) one can conceive of living things that are self-bounded but not subject to evolution. If one allows this as a theoretical possibility, then some organisms may lie outside the scope of orthodox evolutionary biology. That's not to deny that all the life we know about is subject to evolution, nor to reject neo-Darwinism outright. In contrasting autopoietic and evolutionary biologies, we are dealing with a difference of theoretical emphasis, not wholesale contradiction.

Evolutionary biology, one could say, regards all living things as totally subordinated to fitness constraints. This does not mean that no other type of explanation is relevant: biochemistry, for example, is crucial. But the survival or extinction of a species, and the adaptiveness of metabolism, bodily organs, and behaviour are to be explained in terms of fitness. To be sure, some aspects of an organism may be fitness-neutral, maintained because of the chemistry of DNA rather than any adaptive contribution to survival or reproduction. But neutral features, by definition, do not destroy the organism's life-chances. Sometimes, genetic mutation results in something so seriously maladaptive that it does destroy the possibility of life (and therefore of reproduction). In that case, the maladaptive structure is not so much a biological feature as a chemical phenomenon – one that happens to have oc-

curred within a (previously) living creature. In short, fitness on this view is necessary for the existence of living organisms. One might put this by saying that, according to evolutionary biology, vital processes are totally subordinated to the maintenance of fitness.

Autopoietic biology, by contrast, allows that some life need not be subject to evolutionary pressures. (As explained in the next section, this is a substantive empirical hypothesis, not merely a logical implication of the definition of autopoiesis.) It follows that fitness constraints cannot be the overarching consideration for life as such, despite their undeniable importance for all the life we know about. If fitness is not the key to life, what is? For Maturana and Varela, it is the maintenance of the organism's bodily unity.

The formation of a bounded physical unity is a necessary prerequisite of other properties that are normally taken as criteria of life. In reproduction, something is reproduced; in responsiveness, something responds ... and so on. Even to speak of the environment is to make a distinction between an organism and its (sic) biologically relevant surroundings. Every object located in space has surroundings, but only living things have an environment.

What of growth, which might seem to apply to a crystal, or to an autocatalytic network? This property, too, can be ascribed only if we can identify some individual entity. This is unproblematic where crystals are concerned. An autocatalytic network, by contrast, can be identified as an "entity" (as opposed to a mass of non-individuated physical stuff) not by any inherent characteristic or boundary, but only by reference to some arbitrary physical object, the container. This is not part of the network's "environment": network and container are not mutually coupled, but mutually inert.

If the autocatalytic chemical mixture were to be spilt on the floor, rather than confined within the test-tube, it would be most unlikely to expand across the room and out into the corridor. For the raw materials needed for this "growth" would not normally be present. If they were, the network would admittedly seem more like a living thing. Even so, Maturana and Varela would insist that – in the absence of a self-constructed and self-maintaining bodily boundary – the network would not really be alive.

Someone tempted to dismiss their position here as biologically ungrounded should take note that this imaginary example is analogous to the all-too-real phenomenon of fire. Some leading orthodox biologists are content to say that a fire grows, multiplies, varies, and even metabolises (Maynard Smith & Szathmary, 1999, p. 5). However, they refuse to say that it is alive. Their reasons are that a fire has no inner organs whose function is to ensure its survival and reproduction, and that (having no heredity) it does not evolve.

Maturana and Varela, if questioned about the vitality of fire, would be unpersuaded by the fire's lack of evolution. But they, too, would stress the absence of internal organs. Their definition of autopoiesis, as we have seen,

highlights the role of the system's (self-produced) constituents in maintaining the system. ("Maintaining" the system, not "functioning to ensure its ... reproduction.") In most living things, these constituents are either bodily organs or intracellular organelles. Only the simplest prokaryotes lack such constituents. Even those organisms, however, contain enzymes. Enzymes are regarded by evolutionary biologists as "the chemical equivalents of [legs and eyes], organs that function to bring about the growth of the system as a whole" (Maynard Smith & Szathmary, 1999, p. 5). Maturana and Varela would agree. So they have good reason to deny life to fires – and, by analogy, to autocatalytic networks.

What of self-destructive behaviour? This seems to go against the autopoietic theorist's stress on bodily self-maintenance. Altruism and suicide are commonly explained in terms of "selfish genes," whose interests take precedence over those of the organism concerned (Dawkins, 1976). But genes can have only survival-chances, not interests. Concepts such as selfishness – and altruism and suicide too – can be properly applied only to whole organisms. Indeed, they can apply in the full sense only to some human beings, namely, language-using adults. Even so, suicide does sometimes happen. How can autopoietic theory allow for this?

Maturana and Varela do not say that organisms never behave in such a way as to destroy their self-maintenance. They can admit that a deer may trip when fleeing from a lion, or that a moth may fly too close to a fire, and that the deer and the moth will then die. More pertinent than these examples, which concern accidental death, are cases where the self-destruction is systematically predictable or even deliberately planned.

Deliberate suicide is not excluded by autopoietic theory, because the autopoietic processes that enable it to happen occur not in the physical space but in the space of linguistic communication (see above). In other words, suicide is a human-psychological phenomenon, not a biological one. People who sacrifice themselves for the sake of their country or religion are seeking to maintain a particular communicative system, or culture. Even a mother laying down her life for her son is acting so as to maintain (as a psycho-cultural system to some degree independent of her own existence) her notion of family values and/or her beliefs and goals with respect to her relationship with her son. This behaviour may be causally over-determined, in that it has also been selected by biological evolution. But that does not mean that the psychological explanation is irrelevant. Considered as an intentional phenomenon, suicide cannot be explained in purely biological terms.

Processes going on in the communicative space might even result in the deliberate destruction (mass suicide) of the entire human species. Maturana and Varela would see this as a failure of autopoiesis, an unfortunate phenomenon – comparable to the moth's fiery death – caused by lack of foresight or some other communicative maladaptation. But such a disaster is conceivable only because of the potentially all-encompassing scope of hu-

man communication, by means of powerful communicative technologies which in this hypothetical situation would have to reach even to the depths of the Amazonian forests. (Biological autopoiesis could not universally self-destruct, but it could all be near-simultaneously destroyed by some cosmic catastrophe.)

Animal “suicide” and “altruism” cannot be explained by autopoiesis in the communicative space. For, as the scare-quotes suggest, they are not self-consciously deliberate actions. (I ignore charming anecdotes about dogs welcoming death in order to save their masters.) According to sociobiology, they are genetically-based behaviour patterns whose occurrence predictably favours the individual's gene pool, which is not a living organism but an abstraction. In principle, an autopoietic biology can allow for these behaviours – though it is unlikely to choose them as research topics. For autopoiesis allows that self-organizing activities may result in the emergence of new phenomena, which can then provide feedback to affect the originating systems. A prime example of such a phenomenon is evolution, which Maturana and Varela acknowledge to be a feature of all the life we know about. Once evolution gets started, it will affect the behaviour of living things – perhaps even causing some individuals to destroy their autopoietic unity earlier than would otherwise have happened.

A further doubt about the importance of the bodily boundary concerns its relevance for A-Life. In discussing “A-Life and Metabolism” (above), I showed that virtual A-Life creatures do not exhibit metabolism in the strong sense, and – despite being implemented in physical computers – do not have bodies. The similarity of metabolism to autopoiesis in the physical space means that such creatures don't satisfy Maturana and Varela's criterion of life, either. An autopoietic biology therefore rules out strong A-Life as impossible.

But what of robots? We have seen that a plant or animal is a higher-level autopoietic system, made up of lower-level self-sustaining systems, or cells. Many aspects of an animal or vegetable body can be explained by the fact that they make the lower-level autopoiesis possible. For instance, tubes for transporting air, blood, or sap enable oxygen and/or nutrients to be carried to cells that need them but cannot absorb them directly from the environment. Could a colony of self-maintaining robots be essentially similar, so genuinely alive?

Certainly, there could be “colonies” of self-replicating robots. These could involve automatically functioning robot-factories, used for the production of robot-parts. There could be assembler-robots dedicated to building new robots out of those parts, and transplant-robots capable of replacing some of their own (or other robots') damaged parts. In principle (though this would probably be more difficult to engineer), there could even be tinker-robots, which could mend some types of damage rather than substituting a new robot-part for an old one. Indeed, John von Neumann's early ideas on

mechanical self-assembly and self-replication (Burks, 1966) led to several years of NASA-sponsored discussions about how these things might actually be done on the surface of the moon (Levy, 1992, pp. 32-42). And his suggestion (soon afterwards) that copy-errors could make artificial evolution possible might be applied to such self-replicating robots, so that the final generation was different from, and fitter than, the first. The "fingers," for instance, might be longer (giving more leverage), or the "feet" broader (more stability). This example is not totally fanciful: robot controllers ("brains") and sensory systems have already been evolved by the use of genetic algorithms (Cliff, Harvey & Husbands, 1993; Husbands, Harvey & Cliff, 1995).

Notice, however, that I have been careful here to speak of robot-parts, not of body-parts. For these hypothetical robots do not really have bodies. It's true that, unlike virtual creatures (the candidates for strong A-Life), they have clearly identifiable physical boundaries. Their behaviour involves the use of physical organs (jointed supports or "legs", manipulators or "hands", cameras or "eyes") to interact with the physical world. Moreover, brand new robots are assembled by other robots, and each robot (with or without assistance from others) can change, or renew, parts of itself as time goes by. Analogously, one might say, every cell in a human body is renewed at least once within seven years, and depends on many other specialised cells for that to be possible.

But analogy is all that is on offer here. And the analogy is weak. These robots are not genuinely self-constituting, or autopoietic. The raw materials for their construction may have been collected on the moon by robot-miners and purified by robot-smelters, rather than being sent up ready-made from Earth. But the miners and/or the factories had to be rocketed up by us in the first place. And the general nature of their activities (whether part-programmed or fully mechanical) was decided by human engineers. Once the process gets started, the robot-colony is in a sense self-sustaining. But it is not autopoietic. For some of the constituents of the system were not autonomously generated, but manufactured by alien beings (namely, humans) on Earth. This is engineering, not life.

### **Autopoiesis, Biology, and Cognition**

Autopoietic theory exemplifies "biology as it could be." Indeed, since it already exists (the first publications appeared in the 1960s), one could say that it exemplifies "biology as we know it". But this would be misleading. The vocabulary of autopoiesis is unfamiliar to most biologists, and a minority taste among those biologists who have encountered it. Moreover, there are a number of differences between the theory of autopoiesis and more orthodox biology, which prevent its wider acceptance.

Some differences are arguably matters of emphasis, which need not drastically affect the choice of research questions. These include what the autopoietic approach says about the nature of death, and about the relations between life, reproduction, and evolution (see below). Other differences are more likely to arouse genuine puzzlement, or even exasperated rejection. For Maturana and Varela have fundamental reservations about some concepts widely used in biology – and in psychology and cognitive science, too. On the one hand, they reject theoretical terminology that is common in both biology and psychology. On the other hand, they speak of knowledge and cognition in many contexts where psychologists and most cognitive scientists would not.

These more unorthodox aspects of autopoietic theory will be addressed in the following two sub-sections (“No information, or representation” and “All life is cognition”). Here, let us consider three aspects of the autopoietic concept of life that distinguish it from most other definitions – namely, its implications regarding death, reproduction, and evolution.

One unusual characteristic of this approach is that it insists on the continuity of metabolism. The third definition of metabolism given above allows the possibility, in principle, of non-continuous metabolism in living things. It allows, for instance, that metabolism in a frozen or “dormant” spore might not merely be very slow (as it is in a hibernating animal), but actually suspended. By contrast, no possibility of interrupted autopoiesis is granted by Maturana and Varela (op. cit., p. 98). This is not primarily because they believe – as most proponents of metabolism perhaps do too – that the physical processes concerned are, as a matter of fact, continuously dynamic. Rather, their conception of autopoiesis as the fundamental source of the unity of the living thing forbids any suggestion that it might be interrupted without thereby destroying the vital integrity of the system in question. As they put it: “in a living system, loss of autopoiesis is disintegration as a unity and loss of identity, that is, death” (op. cit., p. 112).

In less philosophically self-conscious writing, what counts as death and what as dormancy, or what as life and what as mere viability, are slippery issues. A science-journalist reporting a technique for freeze-drying the sperm of transgenic mice speaks of the “dead” sperm being stored in vacuum-sealed jars, but goes on in the next paragraph to describe them as having “remained viable” (Boyce, 1998). (When water is added, and the sperm heads are removed and injected directly into mouse eggs, which are then implanted in the wombs of female mice, about one-third grow into normal adults.) If this is death, it is clearly not the loss of life but only its intermission (the copy-editor's sub-title refers to “Sperm that come back from the dead”).

The scientists responsible for this work (Ryuzo Yanagimachi and Teruhiko Wakayama) are quoted as saying in their technical report on it: “Although they [the sperm] are dead in the conventional sense, they can sup-

port normal development when injected directly." The sense of "support," here, is metabolic: the sperm heads are not being described as sympathetic friends or cheerleaders, but as developmentally essential metabolites (specifically, DNA). Since it is even more unlikely that dynamic metabolism is going on in freeze-dried sperm than in naturally frozen spores (which still contain water molecules), it would be interesting to know what Maturana and Varela would say about this admittedly highly artificial case.

Besides insisting that life requires uninterrupted physical autopoiesis, Maturana and Varela insist that reproduction is not part of the definition of life (op. cit., pp. 105-7). They argue that the notion of self-reproduction assumes the pre-existence of an identifiable unity. That unity, for them, is explained by autopoiesis. Thus autopoiesis alone is the definition of life. Reproduction is a secondary property.

As expressed in the previous paragraph, this may sound like a merely semantic point. Interestingly, however, it is not. On the contrary, it suggests a substantive biological hypothesis.

The first living things (cells) would necessarily have satisfied the criterion of autopoiesis, but might have been incapable of reproduction. The earliest reproduction could have happened accidentally. The cause of this form of "reproduction" (more accurately: increase in numbers) might have been mere mechanical splitting, without any mechanism for active self-copying. If autopoietic powers were distributed across the whole system, and the cell were to be broken into two or more parts by some external force, then each part (subject, perhaps, to a condition of minimal size) would be capable of persisting as a self-coherent unit. Conceivably, there might even be significant differences between the successor-cells. Suppose that some substructure of the original cell were located within only one of the cell-fragments. Provided that autopoiesis of some type or other could persist in the absence of this structure, the fragments would constitute organisms of various kinds: different from their "ancestor-cell" and also different from each other.

If, by chance, any structural change were to happen which made accidental breakage easier, or which made some non-accidental (internally generated) splitting possible, then systems capable of reproduction would gradually outnumber non-reproducing systems. Given evolutionary pressures, they would eventually exclude them. Hence, say Maturana and Varela, the ubiquity of reproduction in life as we know it.

The concept of evolution, of course, presupposes the existence of reproduction – where this implies not mere increase in numbers but self-copying, or heredity. According to the biologist John Maynard Smith, this, in turn, requires some digital mechanism of inheritance (1996, p. 117). In other words, evolution cannot get off the ground until some particulate mechanism of inheritance has emerged. Mere increase in numbers of autopoietic organisms is not enough. So evolution too, for Maturana and Varela, is a

secondary characteristic of life rather than an essential one. The suggestion mooted by Maynard Smith (e.g. Maynard Smith & Szathmary, 1999, p. 3) and by the philosopher Mark Bedau (1996), that evolution be included within the concept of life, is unambiguously rejected.

(One does not have to be a champion of autopoiesis to be troubled by the logic of evolutionary definitions of life, for these are counterintuitive in a number of ways. One problem is especially relevant here. In the everyday sense of the term, and also in biologists' research on metabolism, the paradigm case of life is an individual living thing. An oak tree, a lion, an ant, an amoeba, even an individual bacterium ... these things exemplify life. But an individual organism cannot evolve: only populations, consisting of successive generations linked by inheritance, can do that. Bedau explicitly admits that an evolutionary philosophy of life can regard single organisms as "alive" only in a derivative sense, namely, in that they form part of an evolving population. This implies that artefacts, such as the moon-robots described above in "Some doubts about the boundary," could not be alive if they had been manufactured to match a blueprint, rather than evolved over time – not even if their behaviour was highly adaptive with respect to their lunar environment.)

Despite their substantive hypothesis about the very earliest living things, a biologist might be content to disregard Maturana and Varela's remarks about reproduction and evolution. For they don't deny the universality of reproduction and evolution in practice, even while denying their philosophical necessity.

Similarly, one need not be too troubled over their definition of death, since questions about the continuity of metabolism rarely arise in practice. Biologists know that hibernation is not the complete cessation of metabolism. Even in cryogenics, where the situation is less clear, what is important is the phenomenon (the effect of injecting freeze-dried sperm, or the biochemical changes – if any – that can happen while it's stored in the jar), not the words used to describe it. Admittedly, the identification of death in human beings is a real problem, biologically as well as legally. But this is largely because it concerns the breakdown of a hugely complex autopoietic system. Individual humans are conceptualized at a much higher level than their various constitutive sub-systems (think of dementia, the heart-beat, "brain-death," and "persistent vegetative state"), never mind the metabolism of single cells.

Some other theoretical eccentricities, however, cannot be brushed aside so lightly. For some of Maturana and Varela's claims challenge fundamental assumptions that are widespread in the scientific community. Specifically, their critique of familiar concepts such as information, function, and representation, and their unusual gloss on the concept of cognition threaten to undermine – or at least to reinterpret – a huge variety of work in biology and cognitive science.

**No information, or representation**

Maturana and Varela's emphasis on self-organized unity as the constitutive principle of the living system leads them to reject informational concepts in biology, including neuroscience. Such concepts apply, they say, only when describing the passing of messages ("the reduction of uncertainties") between two independent unities, where the messenger acts as "an arbitrary non-participant link" (op. cit., p. 102). Where every constituent of the system is an essential participatory element of it, talk of information processing is out of place.

According to autopoietic theory, then, life is not (definable as) information processing, nor do living things – as opposed to artefacts – do information processing. More accurately, only a tiny subset of living things – namely, adult human beings with access to information technology (including pencil and paper) – can do information processing. And they do not do so in virtue of being alive, but in virtue of possessing a language. Language engenders complex autopoietic systems of communication, involving symbolic codes and self-description. Language-users can indeed state, transmit, and process information, if they wish. Even so, this is not what they are usually doing when they use language: usually, they are communicating meanings. By contrast, living systems as such, even including human brains, can neither communicate meanings nor process information.

Nor do Maturana and Varela allow that processes within a living body, as opposed to processes in a computer, can really have a function. "Function", they say, applies only to the domain of purposeful human design. Moreover, function defined (metaphorically) in relation to evolutionary fitness is, like evolution itself, necessarily secondary to autopoiesis. This puts them at odds not only with many biologists, but also with most workers in cognitive science and A-Life.

They would certainly reject von Neumann's suggestion (lauded by most A-Life workers, and a fortiori by supporters of strong A-Life) that reproduction can be glossed as an informational notion – namely, self-copying (Burks, 1966, 1970). They distinguish between "self-reproduction" and "self-copying", saying that only the first involves autopoiesis.

They even describe the genetic "code" as a fundamentally misleading metaphor (op. cit., p. 102). Since DNA is a constitutive part of the autopoiesis of the organism, not a mere arbitrary non-participant link, its role cannot be informational. Admittedly, it may be arbitrary in the sense that some other biochemical carrier of heredity might have evolved (as other molecules might have been the energy currencies used by all known life: see the discussion of metabolism, above). But, once DNA has taken on this autopoietic role, it is integral to the self-reproduction of the organism.

Maturana and Varela's avoidance of informational concepts leads them to deny also that organisms have any inputs from or outputs to the environment. They speak only of "perturbations" of the system itself. They grant

that an observer may find it useful to distinguish between “internal” and “external” perturbations, but insist that for the autopoietic system itself these are indistinguishable. A state of the system is merely a state of the system: it does not carry a label announcing its causation. In reality, then, they are all internal perturbations. (Moreover, they are all perturbations in the present tense: an observer may say that the system has “learnt” something, but the system merely does what its state at that moment leads it to do.)

Equally, they refuse to describe bodily processes, even the activities of nervous systems, as involving internal representations of the environment. Such language, they say, is “metaphorically useful, but inadequate and misleading [in revealing] the organization of an autopoietic system” (op. cit., p. 99; and see pp. 22-6). According to Maturana and Varela, only human autopoiesis can produce representations. For only human beings can act as observers of their own cognitive interactions, and treat representations of them as though they were independent entities (op. cit., p. 14). Most of these representations are words or sentences in the person’s language. A picture is a representation too – not because of any physical similarity between it and what it depicts, but because of the communicative context that gives it meaning. And this context is human language. Likewise, human beings can sometimes distinguish between “inputs” and “outputs” in their own case, because they can reflect on the causal history of their own states.

The eschewal of terms such as input, output, information, function, and representation – and the insistence that “the neuron cannot be considered as the functional unit of the nervous system” (op. cit., p. 19) – is especially interesting, given that Maturana was a co-author of the famous paper “What the Frog’s Eye Tells the Frog’s Brain” (Lettvin, Maturana, McCulloch & Pitts, 1959). He and his colleagues interpreted their research as having found single-cell feature detectors in the frog’s retina. Their experiments provided the first neurophysiological evidence of such things. But the idea that perceptual feature detectors might exist had come from an information-theoretic context. At the end of their paper, the authors specifically acknowledged their intellectual debt to Oliver Selfridge, whose pioneering model “Pandemonium” had simulated a series of increasingly complex visual feature detectors (Selfridge, 1959).

Maturana’s seminal work encouraged David Hubel and Torsten Wiesel to look for feature detectors in the visual cortex of cats and, later, monkeys (Hubel & Wiesel, 1959, 1962, 1968). Since then, many others have been identified in the brains of humans and other species. Some of these appear to detect fairly complex inputs, such as faces, hands, or paws. One, found in certain wading birds that use their beaks to search for food in the mud, fires only when the lower surface of the upper beak-half and the upper surface of the lower beak-half are stimulated simultaneously (Pettigrew & Frost, 1985). Most neuroscientists and cognitive scientists would say that the cell in ques-

tion carries the information that, or even represents the fact that, the bird has picked up something potentially edible. Indeed, Maturana and his co-authors made similar remarks about the “convexity detectors” or “bug perceivers” that they found in the frog’s retina. Now, however, Maturana avoids such language.

Rather, he says that the nervous system – of which the relevant cell is an integral part – continuously couples the organism with its environment so that it (the organism as a whole) is perturbed in certain ways, given certain conditions in the environment. These perturbations sometimes involve specific types of behaviour (conduct), such as nicely directed movements of a frog’s long, sticky tongue. An observer may find it convenient to describe this behaviour as an appropriate “output” caused by “input” to the relevant “feature detector”, and may regard this causation as part of the “function” of the cell concerned. But for Maturana and Varela, the proper description of this sensory-motor circularity as a biological phenomenon is in terms of the ongoing autopoiesis of the whole animal. As they put it: “Only conduct itself [not the neuron, nor any fixed group of neurons] can be considered as the functional unit of the nervous system” (op. cit., p. 19).

It is not only creatures with nervous systems that are credited by (most) biologists with internal representations. The widespread phenomenon of the “body-clock” is often seen as involving internal representations of diurnal or seasonal rhythms. Maturana and Varela would say, instead, that it is an autopoietic phenomenon, enabling the organism to preserve its unity by making internal compensations for rhythmically varying conditions of light and temperature.

It seems, then, that Maturana and Varela are fundamentally at odds with most work in biology and cognitive science (A-Life included). But it’s important here to distinguish two different claims. The first is that certain terms cannot be applied strictly literally to living systems, although it may be convenient – and usually not seriously misleading – to use them metaphorically. The second is that certain terms should not be used at all, since doing so inevitably leads one down research-paths that are ultimately empirically sterile.

The first claim is exemplified by Maturana and Varela’s remarks about “input,” “output,” and “function,” which even they allow are in practice helpful. Their position on “representation,” that it should be applied only when self-reflexive language-based thought is involved, is an example of the second.

The latter objection to orthodox biology and (especially) cognitive science is less eccentric, and more substantive. Indeed, one does not have to be a theorist of autopoiesis to raise doubts about the propriety and methodological fruitfulness of ascribing representations to people, animals, or computers. Both classical and connectionist cognitive science (whether as AI or as

computational psychology) have been widely criticized for making such ascriptions.

Here, philosophical and empirical problems are intimately mixed. For just what it might mean to say that a creature employs a (certain sort of) representation is only gradually becoming clear. In the cases where such criticism is most helpful, it offers an empirical hypothesis. Namely, that (some) cognitive processes do not involve mechanisms of a certain type, and/or do involve mechanisms of some other specified type. These mechanisms may be defined psychologically or neurophysiologically (cf. Clark, 1997).

By the same token, dynamical theorists must be able to distinguish different types of "perturbation" if their approach is to inform us about the specifics of human and animal autopoiesis. Compare atomic theory: Democritus and Dalton said something important when they said that all matter is composed of atoms. But they needed to say more than this. They had to explain the diversity of material things in terms of specific properties of the various types (and arrangements) of atoms. In biology and psychology too, whether one favours "representations" or "perturbations," one must be prepared to give specific chapter and verse.

### **All life is cognition**

If Maturana and Varela are unusually niggardly in their use of informational and representational language, they are unusually prolific in their ascriptions of cognition. Knowledge, for them, has no need of a nervous system, still less of anything one could call a brain. Human knowledge, they admit, is more complex than the knowledge embodied in oak trees and amoebae. This is partly because our sense organs are more advanced, but mostly because we have language with which we can go beyond our sensory perceptions. But according to them all living things, without exception, are cognitive systems: "living as a process is a process of cognition" (op. cit., p. 13).

This claim is unpersuasive. It is unnecessary and confusing to widen the scope of cognition so as to include all living things -- including algae and flowering plants (Boden, in press). To be sure, different autopoietic systems have evolved to be closely coupled with different environments. Bacteria, barnacles, and buttercups naturally inhabit (fit closely into, are intimately moulded by) worlds very different from those of bison and butchers. But one can express this biologically important fact without using tendentious terms such as knowledge and cognition.

Where animals are concerned, these terms (which in their primary use concern propositional attitudes) are admittedly tempting. For, at least outside the contexts of theology and AI, ascriptions of knowledge and cognition normally presuppose the possession of perceptual and motor capacities, integrated in adaptive ways. And animals, by definition, have such capacities. So, for instance, the early ethologist Jacob von Uexkull (1957) described

the “worlds” of animals – such as ticks, house-flies, and dogs – as being constructed by their innate repertoires of action and perception. These were worlds of sensory-motor potential, which delimited the range of behaviour that was possible for the species concerned.

Von Uexkull was right to point out that the potential of distinct species differs, and that an aspect of the environment that is crucial for one species may be wholly inaccessible, or invisible, to another. In that sense, animals are subjective creatures who construct their (species-specific) environments from an objective world – a claim that Maturana and Varela would strongly endorse.

But despite his vocabulary of subjectively constructed *Umwelten*, von Uexkull was not arguing that ticks and flies have knowledge, or cognition, in anything like the sense that humans – or even dogs – do. On the contrary, he was taking pains to stress the radical differences between the “worlds” of these creatures. And even he stopped short at ticks. To ascribe cognition to buttercups as readily as to bison, as theorists of autopoiesis do, is to obscure the significant differences (borderline cases notwithstanding) between the living organisms we distinguish as animals and plants.

In effect, Maturana and Varela conflate “cognition” and “adaptation.” Bacteria and buttercups are indeed naturally adapted to their environments, and (within limits) individuals of these species respond adaptively to fleeting environmental changes of the relevant kinds. The possession of knowledge enables adaptive response too, of course. But that’s not to say that all adaptation involves knowledge.

Likewise, one can hold that cognition is necessarily grounded in life without claiming that all life is cognitive. This philosophical belief dates back at least to Aristotle, who taught that humans share many vital properties with plants and (non-human) animals but that only humans possess rational knowledge – or cognition, in the full sense. Commentators disagree over whether Aristotle saw rationality, or *nous*, as a naturally emergent extension of lower vital capacities (such as perception and adaptive action) or as something metaphysically distinct (Matthews, 1996). And later philosophers, including Herbert Spencer and John Dewey, have argued for both these ways of interpreting the claim that cognition requires life (Godfrey-Smith, 1994). But however we interpret it, this claim can be intelligibly made without conflating the concepts of life and cognition as theorists of autopoiesis do.

It would be better, then, to use the term “cognition” more strictly, so as to avoid the implication that autopoiesis necessarily involves cognition. (This would not preclude the claim that only autopoietic systems can have knowledge; nor would it affect the autopoietic attack on informational approaches.) If that restriction is made, the concept of autopoiesis in the physical space comes even closer to the biologist’s concept of metabolism.

## Conclusion

We have seen that the concept of autopoiesis is significantly similar to the strong (third) sense of metabolism. But whereas accounts based on metabolism may take for granted the formation of the cell boundary, or regard it as just one important metabolic phenomenon among others, the theory of autopoiesis explicitly identifies this as the origin of life. One or other of these concepts, each of which focuses on the self-organization of the bodily fabric, should be recognized as the most fundamental feature of life. For without a self-maintaining bodily organism, however simple it may be, the other vital phenomena – growth, development, responsiveness, adaptation, reproduction, and evolution – cannot emerge.

But we have seen also that the two concepts (autopoiesis and metabolism) have different implications for the definition of life. In particular, reproduction and evolution are conceptually secondary to autopoiesis, but are typically listed on the same level as metabolism in non-autopoietic definitions of life. Accordingly, autopoietic theory implies substantive biological hypotheses that are not taken seriously, if they are considered at all, by most biologists. Specifically, it allows the possibility that the very earliest living things were incapable of reproduction, and therefore incapable of evolution too.

The autopoietic approach is unusual also in its choice of theoretical vocabulary for describing behaviour. Orthodox biologists, neuroscientists, and most cognitive scientists are happy to speak in terms of function, information processing (including input, output, computation, instruction, translation, execution, and code), representation, and learning. Maturana and Varela use autopoietic arguments to criticize each of these concepts. Although admitting that they can be useful metaphors, they also see them as potentially misleading. They prefer to speak in literal terms of intimately coupled dynamical systems, connected in a continuous process of mutual perturbation.

To take their philosophy of autopoiesis seriously, then, would be to undermine many concepts and theories familiar within cognitive science. A few cognitive scientists, with or without reference to autopoiesis, have likewise argued against information processing accounts and in favour of dynamical systems theory (Port & van Gelder, 1995; Wheeler, 1996). But this is still a minority view. Nor is it clear to what extent the phenomena that interest cognitive scientists – language, for example – can be described in this format. Much as Dalton's followers had to identify many different types of atom and combinations thereof, so dynamical theorists must specify many different types of cognitive perturbation.

Maturana and Varela have spelt out a number of ways in which their theory has implications for cognitive science (Varela, Thompson & Rosch, 1991; Maturana & Varela, 1992). Quite apart from any points of detail, their

general approach explicitly identifies life with cognition. Put less provocatively, it sees cognition as co-extensive with life. I have argued above that this is unjustified. It is not helpful, for instance, to ascribe cognition to lilies, whose adaptation to their lily-environments is best described in other ways.

That's not to deny that life may be an essential prerequisite of cognition. I have not argued either for or against this view (but see Boden, in press). But if life is indeed a necessary ground of cognition, then it follows that strong AI would have to be grounded in strong A-life. And if strong A-Life is impossible, then strong AI is impossible also.

This paper has shown that strong A-Life is indeed impossible, because of the lack of (third-sense) metabolism and a self-constituted bodily boundary. Although artificial life of some (broadly biochemical) kind is possible, virtual life is not.

Whether this has methodological implications for the enterprise of weak AI (using computer models to implement and test psychological theories) and/or computational psychology is problematic. On the one hand, certain much-criticized failures of traditional AI, such as the frame problem, may be due to the use of "unnatural" (symbolic) methods of information processing. If the methods evolved by biological organisms to enable situated action and flexible pattern-recognition had been used instead, perhaps today's AI-technology would be more successful. On the other hand, even some champions of connectionist AI have suggested that the human brain needs to simulate a von Neumann machine in order to engage in logical, sequential, and/or hierarchical thinking (Norman, 1986). If so, then weak AI can legitimately attempt to simulate such thinking without using methods drawn from connectionism or A-Life.

We have seen that if life is defined in terms of autopoiesis in the physical space, then strong A-Life is out of the question. Suppose we were to define life, instead, in terms of autopoiesis in general. If we did that, then societies and cultural institutions could be counted as living things. And so could some computer simulations (such as Zeleny's, outlined in the sub-section on "Computer modelling of autopoiesis"). That is, strong A-Life would be achievable.

This suggestion should be resisted, for the same reason that one should resist dropping metabolism from the typical definition of life. There is no independent ground, other than the wish to save strong A-Life, for deleting reference to "the physical space." Doing so would obscure the fundamental distinction between starfish and societies, or butterflies and businesses, or columbines and (some) computer simulations. All these systems are autopoietic in the most general, abstract, sense. But only self-organizing systems in the physical space can originate real, metabolising, life.

In sum, Maturana and Varela's concept of autopoiesis in the physical space – that is, biological self-organization – is both intriguing and subversive. Since its cognate concepts include life, metabolism, and cognition, it is

relevant not only for biologists but for cognitive scientists too. In taking self-organization as the basic concept of theoretical biology, autopoietic theory respects what is so special about living things: their victory over the second law of thermodynamics. It celebrates metabolism – and in doing so implies that strong A-Life is impossible. But it reinterprets the role of reproduction and evolution in life as such. Not least, it challenges orthodox cognitive science.

Specifically, it criticizes much of the everyday vocabulary of cognitive science, including related aspects of neuroscience. Strong A-Life falls by the wayside, and strong AI too. But they may be accompanied by some of our deepest theoretical assumptions.

### References:

- Bedau, M. A. (1996). The nature of life. In M. A. Boden (Ed.), *The philosophy of artificial life* (pp. 332-357). Oxford: Oxford University Press.
- Boden, M. A. (Ed.) (1996). *The philosophy of artificial life*. Oxford: Oxford University Press.
- Boden, M. A. (1999). Is metabolism necessary? *British Journal for the Philosophy of Science*, 50(2), 231-248.
- Boden, M. A. (in press). Life and cognition. In J. Branquinho (Ed.), *The foundations of cognitive science at the end of the century*. Lisbon.
- Boyce, N. (1998). Freeze-dried mice. *New Scientist*, 4 July, p. 4.
- Burks, A. W. (Ed.) (1966). *Theory of self-reproducing automata*. Urbana: University of Illinois Press.
- Burks, A. W. (Ed.) (1970). *Essays on cellular automata*. Urbana: University of Illinois Press.
- Clark, A. J. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, Mass.: MIT Press.
- Cliff, D., Harvey, I., & Husbands, P. (1993). Explorations in evolutionary robotics. *Adaptive Behavior*, 2, 73-110.
- Dawkins, R. (1976). *The selfish gene*. Oxford: Oxford University Press.
- Drexler, K. E. (1989). Biological and nanomechanical systems: Contrasts in evolutionary complexity. In C. G. Langton (Ed.), *Artificial Life* (pp. 501-519). Redwood City, CA: Addison-Wesley.
- Godfrey-Smith, P. (1994). Spencer and Dewey on life and mind. In R.A. Brooks and P. Maes (Eds.), *Artificial life IV* (pp. 80-89). Cambridge, Mass.: MIT Press. Reprinted, with minor revisions, in M. A. Boden (Ed.), *The philosophy of artificial life* (pp. 314-331). Oxford: Oxford University Press.
- Grand, S., Cliff, D., & Malhotra, A. (1996). *Creatures: Artificial life autonomous software agents for home entertainment*. Research report CSRP 434. Brighton, University of Sussex School of Cognitive and Computing Sciences. Available from: <ftp://ftp.cogs.susx.ac.uk/pub/reports/csrp/csrp434.ps.Z>.
- Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, 148, 579-91.
- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160, 106-54.

- Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey Striate cortex. *Journal of Physiology*, 195, 215-43.
- Husbands, P., Harvey, I., & Cliff, D. (1995). Circle in the round: State space attractors for evolved sighted robots. *Robotics and Autonomous Systems*, 15, 83-106.
- Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly connected nets. *Journal of Theoretical Biology*, 22, 437-467.
- Kauffman, S. A. (1971). Cellular homeostasis, epigenesis, and replication in randomly aggregated macro-molecular systems. *Journal of Cybernetics*, 1, 71-96.
- Kauffman, S. A. (1992). *The origins of order: Self-organisation and selection in evolution*. Oxford: Oxford University Press.
- Lettvin, J. Y., Maturana, H. R., Pitts, W., & McCulloch, W. S. (1959). What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers*, 47, 1940-59. Reprinted in W. S. McCulloch, *Embodiments of mind* (pp. 230-255). Cambridge, Mass.: MIT Press, 1965.
- Levy, S. (1992). *Artificial life: The quest for a new creation*. New York: Pantheon.
- McMullin, B., & Varela, F. J. (1997). Rediscovering computational autopoiesis. In P. Husbands and I. Harvey (Eds.), *Fourth European Conference on Artificial Life* (pp. 38-47). Cambridge, Mass.: MIT Press.
- Matthews, G. B. (1996). Aristotle on life. In M. A. Boden (Ed.), *The philosophy of artificial life* (pp. 303-313). Oxford: Oxford University Press.
- Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realisation of the living*. London: Reidel.
- Maturana, H. R., & Varela, F. J. (1992). *The tree of knowledge: The biological roots of human understanding*. Boston: Shambala.
- Maynard Smith, J. (1996). Evolution – natural and artificial. In M. A. Boden (Ed.), *The philosophy of artificial life* (pp. 173-178). Oxford: Oxford University Press.
- Maynard Smith, J., & Szathmary, E. (1995). *The major transitions in evolution*. Oxford: W. H. Freeman.
- Maynard Smith, J., & Szathmary, E. (1999). *The origins of life: From the birth of life to the origin of language*. Oxford: Oxford University Press.
- Moran, F., Moreno, A., Minch, E., & Montero, F. (1997). Further steps towards a realistic description of the essence of life. In *Artificial life V* (Proceedings of the Fifth International Workshop on the Synthesis and Simulation of Living Systems) (pp. 255-263). Cambridge, Mass.: MIT Press.
- Norman, D. A. (1986). Reflections on cognition and parallel distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models* (pp. 531-546). Cambridge, Mass.: MIT Press.
- Pettigrew, J. D., & Frost, B. J. (1985). A tactile fovea in the scolopacidae. *Brain, Behavior, and Evolution*, 26, 185-186.
- Ray, T. S. (1992). An approach to the synthesis of life. In C. G. Langton, C. Taylor, J. D. Farmer & S. Rasmussen (Eds.), *Artificial life II* (pp. 371-408). Redwood City, CA: Addison-Wesley. Reprinted in M. A. Boden (Ed.), *The philosophy of artificial life* (pp. 111-145). Oxford: Oxford University Press.
- Ray, T. S. (1994). An evolutionary approach to synthetic biology: Zen and the art of creating life. *Artificial Life*, 1, 179-210.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3, 417-457.

- Selfridge, O. (1959). Pandemonium: A paradigm for learning. In D. V. Blake & A. M. Uttley (Eds.), *Proceedings of the Symposium on Mechanization of Thought Processes* (pp. 511-529). London: H. M. Stationery Office.
- Teubner, G. (forthcoming) The evolution of legal systems. (Paper given at the British Academy, April 1999.) To appear in M. Wheeler & J. Ziman (Eds.), *The evolution of cultural artefacts* (provisional title). Publisher under negotiation.
- Thelen, E., & Smith, L. B. (1993). *A dynamic systems approach to the development of cognition and action*. Cambridge, Mass.: MIT Press.
- Uexkull, J. von (1957). A stroll through the worlds of animals and men. In C. H. Schiller (Ed.), *Instinctive behavior: The development of a modern concept* (pp. 5-82). New York: International Universities Press.
- Varela, F. G., Maturana, H. R., & Uribe, R. B. (1974). Autopoiesis: The organisation of living systems, its characterisation and a model. *Biosystems* 5 (4), 187-96.
- Varela, F. G., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, Mass.: MIT Press.
- Wheeler, M. (1996). From robots to Rothko. In M. A. Boden (Ed.), *The philosophy of artificial life* (pp. 209-236). Oxford: Oxford University Press.
- Zeleny, M. (1977). Self-organisation of living systems: A formal model of autopoiesis. *International Journal of General Systems*, 4, 13-22.
- Zeleny, M., Klir, G. J., & Hufford, K. D. (1989). Precipitation membranes, osmotic growths, and synthetic biology. In C. G. Langton (Ed.), *Artificial life* (pp. 125-139). Redwood City, CA: Addison-Wesley.