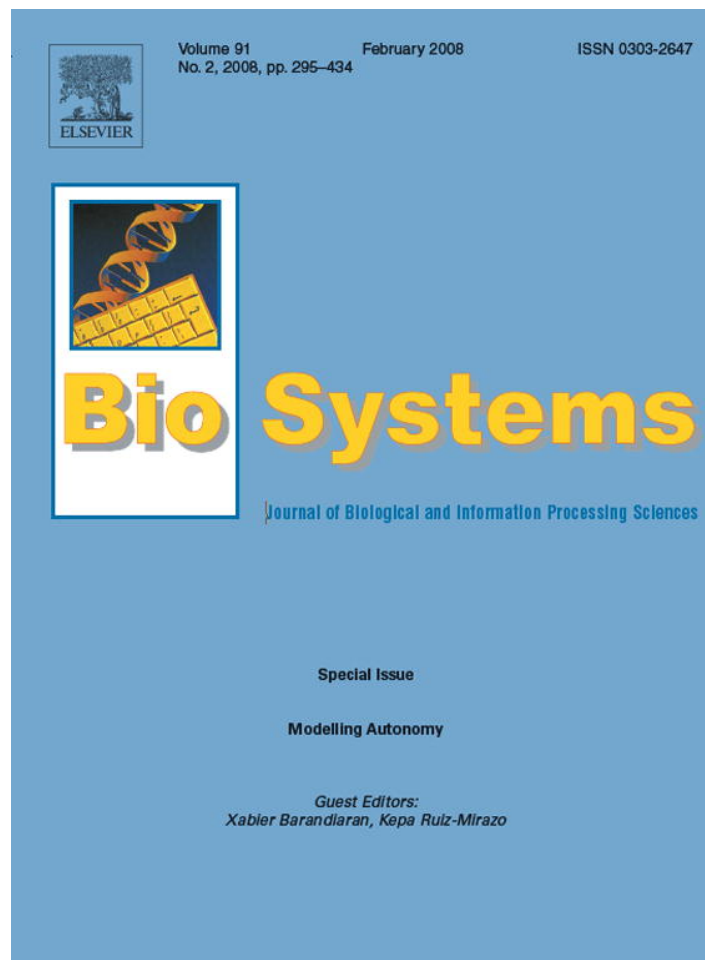


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article was published in an Elsevier journal. The attached copy is furnished to the author for non-commercial research and education use, including for instruction at the author's institution, sharing with colleagues and providing to institution administration.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



ELSEVIER

Available online at www.sciencedirect.com

BioSystems 91 (2008) 295–304

www.elsevier.com/locate/biosystems

Introduction

Modelling autonomy: Simulating the essence of life and cognition

The main concept which, I consider, underlies every observation of a living being, and from which we should never move away, is that it be autonomous in itself, that its parts hold a necessary relationship, that nothing be mechanic, so to speak built and produced from outside, even though the parts act towards the outside and are affected from outside.

JOHANN WOLFGANG GOETHE

1. Preliminaries

The story of this special issue goes back to early 2006, when we thought that it could be a good idea to organize a workshop for the ALifeX Conference (Bloomington, 3–7 June 2006) on the subject of ‘Artificial Autonomy’. The main goal was to analyse the state of affairs in the field, focusing on present modelling approaches and techniques, but also reviewing some foundational aspects of autonomy as a key concept in biological and cognitive sciences. Niels Bertschinger, Anthony Chemero, John Collier, Chrisantha Fernando, Takashi Ikegami, Keisuke Suzuki and ourselves contributed to it, and we were considerably satisfied with the results, in terms of the discussions we had, the quality of the work that was presented there and the shared conviction that autonomous systems research was in need of a thorough revision, assessing previous landmarks in the field but also incorporating some very promising new approaches. This gave us the necessary energy to engage in the project of elaborating further our contributions, putting them together and pushing a possible process of collective publication. The second important input that made this project actually realizable came from the editors of the journal *BioSystems*, who showed us very early their interest in a collection of novel articles on the topic of autonomy. As a result, we considered that it was worth the effort organizing a second meeting in the Basque country, with the aim of involving some other

researchers in the field (who could not attend the meeting in Bloomington) and, at the same time, coordinate and take most out of the reviewing process of the articles to achieve a more complete and integrated special issue.

Thus, during 22 and 23 March 2007, after the first stage of the reviewing process of the submitted contributions was over, we gathered again (with the only exception of A. Chemero, who unfortunately could not attend) in Donostia-San Sebastian, at the Department of Logic and Philosophy of Science of the University of Basque Country. This time, in addition to the original group, Ezequiel Di Paolo, John Stewart, Marieke Rohde, Eckehard Olbrich, Alvaro Moreno, Arantza Etxeberria, Jon Umerez, Tom Ziemke, Barry McMullin, Hugues Bersini, Jesús Ibañez, Matteo Mossio and Tom Froese were also able to attend. Under the title “Modelling biological and cognitive autonomy” the workshop had the same spirit and central goal as the one before, i.e., to address and assess the relevance of the concept of autonomy for living and cognitive systems and, particularly, for the ways in which it should be investigated by artificial means, with a special focus on current simulation modelling approaches to autonomy and novel mathematical attempts to formalize it.

This second workshop served, in fact, as a collective review of the candidate papers for the special issue. With the help of all participants, who made a remarkable effort to read in advance the materials, we managed to organize sessions in a way that promoted exchange of ideas and interactive, productive discussions, avoiding the traditional long oral presentation format. Three main sub-topics were chosen as the general scaffolding for our discussions, according to the three main type of contributions we were aiming at: (a) measuring and conceptualizing autonomy, (b) origins of biological autonomy and minimal agency and (c) autonomous cognitive/behavioural dynamics. Thus, apart from the suggestions coming from the standard review process, authors had the chance to share and comment their papers

with other specialists in the field, collecting feedback in private or group discussions, and incorporating it in the final versions of their work. Most of the articles that follow, putting together this issue, are thus the result of a process of approximately 1 year of development of the ‘embryo-manuscripts’ that were presented in our first meeting. They reflect, to a good extent, the present ways of conceiving and approaching the problem of autonomy, trying to establish the grounds for a new generation of autonomous systems research.

2. Re-modelling Autonomy 30 Years Later: On the Timeliness of This Special Issue

It is now more than 30 years since Varela, Maturana and Uribe published their autopoietic tessellation model in this same journal (Varela et al., 1974), as the first paper in which the autopoietic theory was presented to the international community. Soon afterwards, also in *BioSystems*, Ganti described, for the first time in English, his ideas on ‘chemoton’ systems (Gánti, 1975). These, together with Robert Rosen’s ‘M-R systems’ (e.g. Rosen, 1971), which followed Rashevsky’s mathematical approach to biological systems, were really pioneering models, aimed to grasp, against mainstream research at that time, the systemic nature of minimal living organization. They represented independent islands of ‘system thinking’ in a sea that was being covered by the immense amount of new data coming from molecular biology and the far more fashionable currents of thought focused on the development of the modern synthesis of evolutionary theory and the analysis of its implications. But apparently the scientific community was not ready yet to notice the significance of these works.

Later, half-way through to the present, in the Proceedings of the First European Conference on Artificial Life, Varela and Bourgine (1991) insisted: they envisioned a full research program on the artificial modelling and implementation of autonomous systems (from metabolic networks to immune systems and neural embodied agents). Since autonomy, in their view, was the fundamental property underlying both life and cognition, the challenge of the new artificial sciences (if they were to illuminate our understanding of those phenomena) consisted in the simulation or realization of systems with increasingly autonomous capacities. In addition, that very same year, the so-called “enactive cognitive science” (Varela et al., 1991; see also Varela, 1988) was launched, as a novel autonomous perspective on cognition. In short, artificial autonomy was proposed as a difficult but achievable research goal, whose pursuing could, furthermore, contribute to merge the new and old

paradigms of Artificial Life and Artificial Intelligence, and to uncover some of the most intricate properties of biological and cognitive organization.

However, the academic context was not yet ripe enough for such a challenge: biology was still under the strong influence of molecular reductionism, and cognitive science was just getting renewed by the newborn dynamicist and embodied approach to cognition. As a result, in many respects, the situation we witness 15 years later does not seem to be very different. Of course, there have been contributions during these years that may be considered as clear advances in the field, and a maturation of some early formulations of the theory. These have addressed specific aspects of autonomy, particularly in areas like situated robotics and adaptive behaviour (Steels and Brooks, 1995; Beer, 1997; Smithers, 1997; Ziemke, 1998; Di Paolo, 2003), the problem of the origins and minimal complexity of life (Walde et al., 1994; McMullin and Varela, 1997; Kauffman, 2000; Ruiz-Mirazo and Moreno, 2004; Luisi, 2006) or the philosophical implications that autonomy has for biological and cognitive explanations (Boden, 1996; Etxeberria et al., 2000; Weber and Varela, 2002). However, in their most fundamental and meaningful sense, autonomous systems still remain beyond our present experimental and theoretical capacities. Is it really a question of critical mass within the scientific community? Or is it, rather, that the problem of autonomy is too hard? Do we understand well enough all what the concept involves? Or have we just targeted some of the corollaries of a theory of autonomous systems, without addressing its full implications? Are the methodologies we employ unsuited for the purpose? Or have we not fully squeezed yet the analytic and synthetic tools that artificial life, dynamical systems and information theory provide to model autonomy?

These are still open questions (see also Etxeberria et al., 2000), but we feel the time is coming to start giving more definite answers. Biological and cognitive sciences are undergoing profound changes at the turn of the century, driven by the development of new tools and techniques that allow more holistic approaches to complex dynamical systems. In biology, especially in theoretical biology, there seems to be an ever increasing awareness of the limitations of reductionist approaches, like those exclusively centred on the gene concept (Keller, 2000; Westerhoff and Palsson, 2004), and this is inducing a revival of the idea of ‘systems biology’ (Kitano, 2001; Boogerd et al., 2007), as well as the realization that traditional mechanistic approaches in biology need to pay attention to the notion of organization (Bechtel, 2007). In particular, the emerging field of ‘synthetic biology’

(Benner and Sismour, 2005; Forster and Church, 2006; Solé et al., 2007) can provide some important insights into the minimal organization of life and make possible to deal, in the near future, with empirical (*in vitro* and *in silico*) assessments of the autopoietic theory. Similarly, advances in cognitive neuroscience are progressively approaching the holistic and embodied nature of cognitive processes, partly due to the rapid improvement of neuroimaging techniques and complex modelling tools. On the one hand, the large-scale integration of brain activity is closer to scientific scrutiny (Varela et al., 2001; Freeman, 2001; Tsuda, 2001; Buszaki, 2006—to mention just a few), while simulations of embodied autonomous agents are penetrating the field of neuroscience (Rupin, 2002). On the other hand, the embrained body and the link between internal bioregulatory processes and cognitive behaviour (the intimate relationship between life and mind) are gaining increasing attention (Damasio, 1999, 2003; Moreno and Lasa, 2003). It seems that the scientific community is somehow getting ready, finally, to take on board the challenge of autonomy, a system concept by definition. This is the main reason why we considered that it was a suitable time to publish a special issue on the subject, with the objective to evaluate the present state of the art and the potential results that research in this field will provide in the near future.

This special issue, however, does not stand alone. We are witnessing a revival of the notion of autonomy. Francisco Varela's passing away triggered a number of reviews and publications (Di Paolo, 2004; Bacigalupo and Palacios, 2003), of which the present one could be taken as a late tribute. At the same time, Rosen's ideas on closure are being revisited (Letelier et al., 2005), together with Ganti's (2003, 2004), while the notion of autonomy and enaction in cognitive science are regaining attention (Thompson, 2007; Stewart et al., 2007).

3. Autonomy: Its Characterization and Modelling

All organisms appear endowed with the capacity to continuously generate and regenerate their components, to set their own goals without external control and to adaptively and selectively modify their internal and interactive processes according to those goals. It is their *autonomy* that strikes as a characteristic – yet elusive – property: i.e. their capacity to create their own norms, constraints and regulation in a recurrent way, creatively playing with the laws of physics and chemistry against thermodynamic inertia, while continuously reproducing their own identity. Since early Aristotelian conceptualization of nature, organisms have been distinguished by a

holistic interdependence between matter, form and function in them; by their unity as tightly integrated systems, endowed with an intrinsic teleology, a continuous regeneration of their *raison d'être*, of their causes. Spinoza actually built a good part of his philosophy of nature on top of the notion of *conatus*: organisms' intrinsic drive towards self-assertion and self-maintenance. Yet, Kant was among the first to recognize the real difficulties that the understanding of such systems involved for the human mechanistic mind set-up (especially after its shaping by Cartesian rational dualism and Newtonian corpuscular physicalism). Nevertheless, following the Kantian tradition, the *naturphilosophie* put organisms and their autonomy at the centre of early biological research (Goethe, Steffens and so on).

However, it was not until the development, during the 1970s, of earlier cyberneticist ideas that the concept of autonomy (closely related to the notion of network, self-organization and closure) came to be formulated so as to be scientifically approachable. As a result, today it stands (free of its vitalistic connotations) at the centre of the sciences of complexity, systems biology and artificial life for quite a few of us. Autonomy is a kind of complexity, a type of dynamic organization that is taken to be constitutive of life and mind: i.e. the type of mechanism that distinguishes living and cognitive systems from the non-living. Yet, and despite the effort made by some researchers, there is no mainstream discipline that has seriously taken up the challenge of further developing and empirically grounding the concept.

Here is where the present special issue finds its place, as an attempt to fill in the gap left by standard disciplines around this idea of autonomy, so often neglected and misunderstood, which nevertheless holds the key, in our view, to the future sciences of life and cognition. Mimicking the structure of Varela et al.'s article, as if we were actually harvesting the – now diversified – produce of that seminal work, the contributions to this special issue are organized in two main parts: characterization and modelling. The first block aims to *characterize* autonomy at a conceptual level, providing theoretical schemes and mathematical tools for the task. The second block contains specific examples of current *modelling practices* of autonomous systems, ordered in a bottom-up way: from chemical or (proto-)metabolic autonomy to higher level, behavioural–cognitive approaches to it. The emphasis is put on modelling practice rather than on the model objects themselves (mostly computer simulations) to stress their use as epistemic tools, whose interpretation and manipulation, even if partial and incomplete, can throw new light on the nature of autonomous phenomenology.

Despite the diversity in the motivations, tools and backgrounds of the contributions, they integrate a rather coherent ensemble, which covers a good deal of what is being (or has been) done in the field. Each article of the issue can be regarded as a particular entry in the matrix of open theoretical problems and modelling approaches/techniques that define the current network of autonomous system's research (from philosophy to robotics, through supramolecular chemistry, epistemology, mathematics and neuroscience). But, as we outline below, there is a very natural complementary relationship between the different articles of the collection, which contributes to their overall integration.

3.1. *Autonomy: Attempts to Define and Formalize the Concept*

Right at the beginning, in the launch article of this issue, **Margaret Boden** sends us the warning that the notion of autonomy still remains a buzzword that stands somewhere in-between (or across) the muggy lands of self-organization and freedom. Denoting the capacity to act without external control (including, for some, the programming or engineering action of an external designer), the concept needs to face the paradoxes that stir from the origins and definition of that *self* or *auto*, which is to be the source of the rules, mechanisms or constraints that govern its behaviour. Yet, this is just one of the difficulties that the concept entails. The different degrees and types of self-determination that constitute autonomy would be another. According to Boden, three aspects define a system as autonomous: (a) the degree of indirect (internally mediated) response to environmental changes, (b) the degree of self-generation of the behaviour mediating internal mechanisms and (c) the degree of selective self-modification of such mechanisms (both sensitive to specific contexts and wider concerns). Holism, dynamical systems, emergence theory and a certain dose of anti-representationalism permeate the field but still, Boden suggests, a GOFAI approach (the Good Old Fashioned AI, symbolic and disembodied, traditionally neglected by a life and situated robotic practitioners) seems unavoidable if we are to understand human freedom, which remains as the epitome of autonomy and would be best understood in terms of symbolic reflexive reasoning.

Therefore, conceptually speaking, autonomy still requires further clarification work. **Moreno, Etxeberria and Umerez**, in the following paper, contribute to that work, assessing the value and limitations of the classical autopoietic approach. Regarding the minimal and more basic form of autonomy (autopoiesis, or biochemical

self-production) they stress that the abstract formulation of a recursive organization needs to be complemented with a materially and thermodynamically grounded perspective, capable of accounting for what makes the system an active source of its own behaviour (over and above the laws of physics and chemistry that determine the system). Three open issues come out as a consequence: (a) interactive aspects of autonomy should be taken as definitory of autonomous organizations (a complementary integration between identity and agency), (b) the past focus on spontaneous “one-shot, order-for-free” self-organized patterns should be shifted to the hierarchical integration of differentiated functions that constitutes natural autonomous systems and (c) the different degrees and levels of autonomy that evolution has unfolded require more attention than what it previously attracted (how does a system become “more” autonomous, which are the crucial transitions, how do levels interrelate, etc.). Finally, Moreno et al. deal with some of the paradoxes and difficulties that the very idea of artificial autonomy encloses: in particular, the paradox of creating something that needs to remain self-created. The field, they conclude, has much more to gain from the use of artefacts as models (as incomplete and revisable tools for understanding autonomy) than from the aspiration to create (and philosophically revolve around) “genuine” autonomous artefacts.

Certainly, what remains central to some of the most influential accounts of biological autonomy is the notion of closure (**Van der Vijver and Chandler, 2000**). Autonomous systems appear circularly networked, all its component processes relate to each other in an interdependent manner. Varela's notion of *operational or organizational closure* (1979), Rosen's *closure to efficient causation* (1991), **Kauffman's closure in the space of catalytic tasks** (1986; 2000) or **Pattee's statistical and semantic closure** (1973; 1982) stand as important contributions in this direction. But, although these – now classical – conceptual frameworks can be regarded as some of the most interesting or solid formulations of autonomy so far, in many aspects they still remain unclear and are often not so useful to advance in experimental or modelling research. **Chemero and Turvey's** contribution is precisely devoted to finding a general way to assess, clarify and compare these different theoretical schemes and their implications. They explore the value of hyperset theory as a tool to define the type of complexity that constitutes autonomy. Circularities involving sets being members of themselves were forbidden in the foundations of mathematics at the beginning of the 20th century, to avoid a number of paradoxes and inconsistencies. However, complex systems (and autonomous

systems in particular) belong to the class of systems that are best captured by circular models. Hyperset theory comes to solve this problem by providing the means to model circular processes. Chemero and Turvey apply the graphical notation of hyperset theory to distinguish the type of complexity found in Rosen's M-R systems, autopoietic systems and Kauffman's autocatalytic sets. They also draw the interesting conclusion that, given recent research on computability of impredicativities, models of autonomous systems would be, in fact, computable (against Rosen's claim, even if they consider most of his intuitions still relevant for understanding such systems).

A rather different characterization of autonomy comes out when the system is analysed externally, in terms of its relationship with the environment. The standard approaches to the problem (Varela's, Rosen's, Kauffman's, and so on) put the emphasis on the identity of the system as a cohesive, integrated, closed unity. At most (in the autopoietic case) claiming that it must actively distinguish itself from the surroundings. But the analysis of the relationship with the environment is always at a secondary level, introduced (if ever) after the defining criterion for autonomy is given. However, a materially grounded, less abstract formulation of the notion of autonomy requires that thermodynamic constraints be introduced right from the beginning (as pointed out by Moreno et al. this issue; for a more elaborate view, see Ruiz-Mirazo and Moreno, 2004; or also Bickhard, 2000; Christiansen and Hooker, 2000). This implies that autonomous systems must be open, intrinsically coupled with a flow of matter and energy from/to the environment, so the interactive dimension of autonomy is not something to be added or attached, once its basic way of organization is defined, but has to be included in its very definition.

Bertschinger and colleagues' contribution tries to formalize, in a rigorous, systematic way, these relational-interactive aspects of autonomy. They adopt an information theoretic approach and propose a measure that quantifies the interactive dimension of autonomy as a combination of two factors: (a) non-heteronomy (an autonomous system is not determined by environmental states) and (b) self-determination (the evolution of an autonomous system is determined by its past states, i.e. it is not a random system). The development of this measurement and its application to existing artificial life simulation models leads to interesting problems dealing with the reciprocal influences between system and environment (internal modelling, system's control on the environment, etc.). In order to solve some of the counterintuitive points that come out of a system with a high

capacity to modify its environment (thus sharing with it a high degree of mutual information), probabilistic-observational measurements are complemented with causal-interventional procedures that make explicit how much of the system–environment correlations are due to the pro-active nature of autonomous system. The paper acknowledges the difficulty of defining the appropriate observables and the system–environment distinction, which is pre-assumed within their information theoretic approach, even if, in reality, it should be generated by the autonomous system itself. It is the aspect of constitutive autonomy (in the sense of a self-maintaining/self-producing organization) that somehow escapes information theoretic approaches, which focus on the interactive dimension of autonomy.

Collier's way of conceiving autonomy combines the interactive and constitutive aspects of it. He claims that it is a mistake to think that operational closure must necessarily be complete. In fact, he makes a classification between different (increasingly stronger) forms of autonomy according to their way of achieving this closure, or relative closure, as well as to their anticipatory capacities. In his account, weak anticipation involves that future states of the environment are projected using a model of the environment and environmental data (inputs–outputs) alone; whereas strong anticipation requires that the system generate a model of itself (or part of itself), in order to project its future states from its current internal state. He then explores the consequences that this has for the design and accurate simulation of living systems. This contribution has the value of providing a conceptual bridge between the theoretical framework of autopoietic authors (Maturana and Varela, 1980; Varela, 1979) and that of Rosen's (1985; 1991). And it also goes somewhere beyond, proposing a naturalized version of functionality (based on the self-preservation or viability of an autonomous entity) that is connected to a stage of autonomy ('autonomy₃', as he calls it) in which closure is not complete, but just dominant. This would allow the system to be open "in its own logic" (in the sense that functionality can depend on future states involving already constructed closures), on top of being open to energy and matter flows.

3.2. Bottom-up Modelling: From Life to Cognition

The suitable characterization of living and cognitive autonomy requires, at this point (and increasingly so in the future), the use of specific models. Even the more theoretical contributions summarized above rely on models as target objects for their application, to ground the his-

torical and philosophical formulation of the problems addressed, or as the domain where theoretical advances need to be developed. In particular, *simulation* models, which are now thriving in all research areas, allow not only the empirical analysis of systems with multiple and non-linearly interacting components but, used as “tools for thinking”, also the exploration of a complex (yet explicit and systematic) *conceptual* space around those systems (Barandiaran and Moreno, 2006a). Given the complexity of our object of study, the naked human mind, the logical structure of natural language or the strict mathematical analysis, alone, are probably not enough to work out the intricate relationships that are established between certain variables and related concepts in our theories. This is particularly true for those concepts that involve or cut across different levels of organization, timescales and circular relationships: concepts whose understanding cannot be achieved without the mediation of computer simulation models. And autonomy is certainly one of such concepts.

It is not a coincidence that the early formulation of the concept of autopoiesis came together with a simulation model: “This model is significant in two respects: on the one hand, it permits the observation of the autopoietic organization at work in a system simpler than any known living system, as well as its spontaneous generation from components; on the other hand, it may permit the development of formal tools for the analysis and synthesis of autopoietic systems” (Varela et al., 1974: 189). Simulation models provide not only the possibility to reproduce and follow a complex organization “at work” (together with the development of specific tools for its analysis), but also to generate new hypothesis or reduce *ad absurdum* certain assumptions, whose consequences are not self-evident (due to the complexity involved). This is why some authors have labelled these models as “opaque thought experiments” (Bedau, 1998; Di Paolo et al., 2000). Thus, simulation modelling permits to establish a well-defined and dynamically unfolding arena, where theoretical disputes might be solved (or, at least, systematically made explicit) and also where a gradual approximation to the natural and empirically testable conditions of real autonomous systems can be attempted.

So the second half of this special issue is focused on specific simulation models that aim to throw some new light into the traditional conception of the origins and nature of autonomous systems. The first scenario addressed is the world of (proto-)biological chemical reactions, where minimal forms of autonomous agents can already be approached. Here comes **Fernando and Rowe**'s provocative approach to metabolic

autonomy as a result of an evolutionary artificial chemistry that includes a ‘natural selection’ algorithm. This model takes seriously into account (previously unconsidered—yet critical) aspects of side-reactions and thermodynamic constraints on the generation of autocatalytic networks in a prebiotic but fully Darwinian evolutionary scenario. The challenge of this novel and difficult-to-classify model for standard views or more traditional conceptions of the problem involves two main points: (a) first, their claim that natural selection mechanisms should start operating well before self-replicating polymers (like RNA molecules) appeared, being the real driving force to achieve autonomy and (b) second, that their proposed evolutionary dynamics of self-reproducing autocatalytic compartments would lead to full-fledged *agent* systems. Even if it remains controversial whether Fernando and Rowe's model actually achieves agency in any well-grounded, informative sense, it does show that a self-sustaining autocatalytic network is essential to create an open functional space on which selection can operate to produce increasingly complex regulatory forms that might eventually lead to the appearance of agential capacities.

Ruiz-Mirazo and Mavelli's model, coming next, can be classified, instead, as a clear ‘metabolism-first’ approach, in which the protocellular organization of chemical networks (i.e., the – originally autopoietic idea of – complementarity between boundary and metabolism) is revisited, by means of a recently developed stochastic simulation platform. Apart from this computational platform, the novelty of their contribution rests on the realistic way in which membrane processes and physico-chemical constraints (osmotic pressure, surface to volume relationship, and so on) are modelled. This also allows the authors to make a well-defined distinction between protocells (lipid-peptide membranes enclosing an aqueous core where reactions take place) and their environment, studying their continuous interactions (in particular, the coupling of transport and reaction processes), with the final purpose of projecting a theory for the appearance of minimal agency in that context. Although their present simulation results, as acknowledged, do not get that far, they show the way to tackle, in the future, in explicit terms, the problem of the origins of agency, by looking into the *active* role that the membrane must play to control the flow of matter and energy through the system, so as to achieve robust self-construction and avoid, e.g., osmotic bursting.

But can these systems really *act on their own behalf*? The question remains open, and perhaps requires moving beyond ‘bare chemistry’. This is the focus of the following articles. **Ikegami and Suzuki**'s work is meant to

link metabolic self-producing processes with adaptive self-movement, paying special attention to the constitution of a 'self' and its behavioural regulation within viability constraints and merging, in this way, Maturana and Varela's autopoietic theory with Ashby's framework for the modelling of adaptive systems. However, Ikegami and Suzuki manage to push this fusion a step forward: according to them, the critical transition for the origin of autonomous agents requires a framework that allows to investigate 'homeodynamic' systems, as opposed to systems that self-maintain or self-regulate by means of bare homeostatic mechanisms. They introduce two models (one is an expansion of Varela et al.'s tessellation automata; the other, an imaginative elaboration of Daisy world dynamics) with which they illustrate and develop their main point. In particular, the first permits to explicitly establish a link between the self-constructive and interactive dimensions of autonomy, by allowing the tessellation automata to grow differentially in its environment, opening the possibility for self-movement within a gradient of metabolic substrates. The autonomy of the system is, thus, not limited to a circular process of self-production but includes an open interaction with the environment towards preferred environmental conditions for self-reproduction. So we can take this as an example of a model of autopoietically grounded motile agency.

It is precisely this type of relationship between constitutive and interactive autonomy that motivates some criticisms of the embodied and situated approaches to autonomous robotics. **Ziemke's** contribution reviews some of these criticisms, highlighting the need for a more profound sense of autonomy that goes beyond mere physical or sensorimotor embodiment. Robotics needs to be grounded on internal homeostatic regulation. And this is what current theories of emotion in neuroscience just point out: that cognition and consciousness heavily rely on the organismic embodiment of neural and behavioural processes; that is to say, on the interplay between internal bioregulatory processes and neurally guided sensorimotor interactions. Thus, the new challenge for autonomous robotics is to build behaviour upon a hierarchy of levels of automated homeostatic regulation. Ziemke briefly sketches how the mammalian brain anatomy serves as inspiration for the ICEA project: an architecture that integrates emotional and cognitive aspects into an autonomous robot. Actually, it is by making detailed and integrated models of how real animals regulate their behaviour (according to internal homeostatic needs) that future advances on autonomy can be achieved. In this sense the situated and embodied robotic "revolution", however incomplete, allows today

the integration of neurophysiological models with situated behaviour, in order to explore the holistic principles that underlie autonomous behaviour.

Another dimension of the discussion on cognitive, behavioural autonomy is whether it is something that happens exclusively in the (bio-)chemical self-producing/autopoietic domain (expanded through motility) or if it can re-appear again at a purely cognitive level, as some kind of closure on neural dynamics coupled to the behaviour they generate. The consequences of each alternative for behavioural and cognitive sciences are critical. If the first possibility were the case, i.e. if autonomy should only be relevant in so far as metabolism is involved, roboticists would do well to ignore the concept of autonomy and merely focus on the satisfaction of viability constraints that adaptive behaviour should encompass (probably enriched with internal bioregulatory functions). **Di Paolo and Iizuka's** work on evolutionary robotics and its theoretical implications, however, embraces the second alternative. They criticise that the importance of autonomy at the behavioural level is often, either disregarded, as non-externally controlled behaviour, or neglected as belonging to the metabolic level. As a consequence, behaviour tends to be modelled as satisfaction of certain performance criteria, independently of the effect of this performance on the agents behavioural organization. They provide an evolutionary robotic modelling example in which behavioural preference (for two distinct sources of light on a phototactic task) emerges as a result of internal and contextual factors through switching between two different stable attractors of the robot controller (a fully connected dynamical neural network). Yet, there is no internal module acting as an internal homuncular controller, which drives that switching, nor a specific environmental stimulus that reactively triggers the response. According to Di Paolo and Iizuka, preference, value, semantics and other autonomy-related cognitive properties cannot be studied as isolated functional modules controlling behaviour (as it is typically done, even among biologically inspired cognitive scientists). So their work provides a first step into the study of how the homeostatic organization of behaviour is capable of generating, autonomously, its own values and preferences. By working with minimalist dynamical models of embodied simulated agents (unlike detailed and often excessively complex anatomical models of natural cognitive systems), this methodology stands as one of the most valuable tools to study the holistic nature of brain–body–environment interactions that lead to autonomous behaviour (see also **Di Paolo, 2003**; or **Barandiaran and Moreno, 2006b**).

4. Bringing forth a World with Autonomous Systems

As a way of concluding the special issue, we considered suitable to include a reflection on our own nature as autonomous cognitive agents and, particularly, on the implications that this has in our scientific ways of understanding the very concept of autonomy. In fact, autonomous systems research has often overlapped with epistemological issues: it is the inevitable consequence of a reflexive folding of the model into the modeller and the modelling process. **Rohde and Stewart's** contribution, the final of the issue, deals with this problem: they draw some of the consequences of taking a constructivist standpoint (implicitly or explicitly adopted by many researchers in the field: Maturana, 1978; Stewart, 1996; von Foerster, 2003). How is it that we attribute autonomy to other systems? Can we really ground and justify such an ascription? Or are we condemned to play an endless game of imitation? The authors argue that there is no 'Turing-test' for autonomous systems, nothing on their surface behaviour will permit to avoid 'potential cheating'. Thus, the intuitive attribution of autonomy to certain natural systems needs to be a *mechanistically informed* ascription; i.e., behaviour (if it is to be judged autonomous) must be produced by a particular type of generative mechanism. However, a mechanistic analysis may not always provide a satisfactory clue, especially if the context dependence and the situatedness of behaviour are intrinsically involved in the generative mechanism. Rohde and Stewart examine the different implications and assumptions that some of the models presented in this special issue bear for the attribution of autonomy, with special attention to the mechanisms that generate behaviour in each model.

So, after going through the collection of articles that follow, the reader may have the feeling that they are not so conclusive, after all; that everything remains open, up in the air, within this field. To a certain extent, this is true: the mathematical formalizations of autonomy have trouble to find an adequate balance between constitutive and interactive aspects of it; simulations of chemical system candidates for minimal biological autonomy do not agree on whether the approach should be strictly evolutionary (natural selection involved) or, so to speak, developmental (just self-organization principles involved); agency seems to be achievable at the biochemical-metabolic level, for some authors, while others claim that the coupling of self-production dynamics with 'motion' (in the widest sense of the term) is crucial for it; cognitive autonomy, in turn, could be based on the 'homeodynamic' regimes that come out of that coupling or, alternatively,

stem from behavioural, system–environment interactive loops that are decoupled (in a strong sense) from the metabolic domain; and so on. However, at the same time, it is quite clear – although the reader should be the final judge on that, of course – that progress is being made, by the way of addressing the relevant questions and the way of discussing and criticising the tentative answers of active researchers in the field. It is precisely the advances in reframing old questions and the subsequent approaching to empirically testable hypotheses (both possible through the construction of computer simulation models and new formalisms) that this special issue is meant to capture.

We have long lived in a scientific world-view where molecules, genes, individual neural structures, or disembodied algorithms were considered the only and ultimate causes of our experience. Yet that atomistic world-view is starting to change considerably and the complex systems that we are and we live with can start to be understood as holistic, dynamically integrated systems, whose most characteristic properties are not reducible to isolated components. Autonomy is certainly one of such properties, if not the source of many of them: it can be condensed as the capacity to generate and regenerate the boundaries/limits and rules/norms that define those complex dynamic systems. It is in our hands to bring forth a scientific world-view where systems' autonomy is acknowledged. But only further research and time will allow for a more conclusive evaluation of how crucial or helpful the concept of autonomy is in our attempts to capture the essence of life and mind. It is the uncertainty of the final destination that pushes us to walk this road.

Acknowledgements

We would like to acknowledge financial support from the EuCognition network (action 126.01), plus the additional help of the University of the Basque Country, the Department of Logic and Philosophy of Science and Kutxa bank, which made possible, financially and institutionally, the organization of the workshop that was the main driving force to put together this special issue. We also wish to thank all the participants in this and the previous workshop (in ALifeX) for their contributions, their active participation and their joint motivation to arrive at this point. In addition, X.B. has the support of the doctoral fellowship BFI03371-AE from the Basque Government, while K. R. M. holds a *Ramón y Cajal* research position. Thanks also to research grants 9/UPV00003.230-15840/2004 (UPV-EHU), HUM2005-02449 and BFU2006-01951/BMC

(Ministerio de Educación y Ciencia). Finally, many thanks to the members of our IAS research group (Alvaro, Arantza and Jon) for their continuous support during the organization of the workshop in San Sebastian, the editing work, and for helpful comments and corrections to this introductory manuscript.

References

- Bacigalupo, J., Palacios, A. (Eds.), 2003. A tribute to Francisco Varela, Special issue of *Biological Research* 36 (1).
- Barandiaran, X., Moreno, A., 2006a. ALife models as epistemic artefacts. In: *Proceedings of the 10th International Conference on Artificial Life*. MIT Press, Cambridge, MA, pp. 513–519.
- Barandiaran, X., Moreno, A., 2006b. On what makes certain dynamical systems cognitive: a minimally cognitive organization program. *Adaptive Behav.* 14 (2), 171–185.
- Bechtel, W., 2007. Biological mechanisms: organized to maintain autonomy. In: Boogerd, F., et al. (Eds.), *Systems Biology; Philosophical Foundations*. Elsevier, Amsterdam.
- Bedau, M.A., 1998. Philosophical content and method of artificial life. In: Bynum, T.W., Moor, J.H. (Eds.), *The Digital Phoenix: How Computers are Changing Philosophy*. Basil Blackwell, Oxford, UK, pp. 135–152.
- Beer, R., 1997. The dynamics of adaptive behavior: a research program. *Robotics Autonomous Syst.* 20 (2–4), 257–289.
- Benner, S.A., Sismour, A.M., 2005. Synthetic biology. *Nat. Rev. Genet.* 6, 533–543.
- Bickhard, M.H., 2000. Autonomy, function, and representation. *Commun. Cogn. Artif. Intell.* 17 (3–4), 111–131.
- Boden, M., 1996. Autonomy and artificiality. In: Boden, M. (Ed.), *The Philosophy of Artificial Life*. Oxford University Press, Oxford, pp. 95–108.
- Boogerd, F., Bruggeman, F., Hofmeyr, J., Westerhof, H. (Eds.), 2007. *Systems Biology: Philosophical Foundations*. Elsevier, Amsterdam.
- Buszaki, G., 2006. *Rhythms of the Brain*. Oxford University Press, New York.
- Christiansen, W.D., Hooker, C.A., 2000. Autonomy and the emergence of intelligence: organised interactive construction. *Commun. Cogn. Artif. Intell.* 17 (3–4), 133–157.
- Damasio, A.R., 1999. *The Feeling of What Happens: Body Emotion and the Making of Consciousness*. Vintage, London.
- Damasio, A.R., 2003. *Looking for Spinoza: Joy Sorrow and the Feeling Brain*. Orlando, FL, Harcourt.
- Di Paolo, E.A., 2003. Organismically inspired robotics: homeostatic adaptation and natural teleology beyond the closed sensorimotor loop. In: Murase, K., Asakura, T. (Eds.), *Dynamical Systems Approach to Embodiment and Sociality*. Advanced Knowledge International, Adelaide, Australia, pp. 19–42.
- Di Paolo, E.A. (Ed.), 2004. *Francisco Varela and Artificial Life*, Special Issue of *Artificial Life* 10 (3).
- Di Paolo, E., Noble, J., Bullock, S., 2000. Simulation models as opaque thought experiments. In: Bedau (Ed.), *Proc. Artificial Life VII*. MIT Press, Cambridge MA, pp. 497–506.
- Etxeberria, A., Moreno, A., Umerez, J., 2000. The contribution of artificial life and the sciences of complexity to the understanding of autonomous systems (special issue). *CC AI: Commun. Cogn. Artif. Intell.* 17 (3–4).
- Freeman, W., 2001. *How Brains Make up their Minds*. Columbia University Press, New York.
- Forster, A.C., Church, G.M., 2006. Towards synthesis of a minimal cell. *Mol. Syst. Biol.* 2, 45.
- Gánti, T., 1975. Organization of chemical reactions into dividing and metabolizing units: the chemotons. *BioSystems* 7, 15–21.
- Gánti, T., 2003. *The principles of life*. With a commentary by J. Grieseimer and E. Szathmáry. Oxford University Press, Oxford.
- Gánti, T., 2004. *Chemoton Theory*, Vols. I and II. Kluwer Academic/Plenum Publishers.
- Kauffman, S., 1986. Autocatalytic sets of proteins. *J. Theor. Biol.* 119, 1–24.
- Kauffman, S., 2000. *Investigations*. Oxford University Press, Oxford.
- Keller, E.F., 2000. *The Century of the Gene*. Harvard University Press, Cambridge, MA.
- Kitano, H. (Ed.), 2001. *Foundations of Systems Biology*. MIT Press, Cambridge MA.
- Letelier, J.C., Soto-Andrade, J., Guñez-Abarzúa, F., Cornish-Bowden, A., Cárdenas, M., 2005. Organizational invariance and metabolic closure: analysis in terms of (M R) systems. *J. Theor. Biol.* 238 (4), 949–961.
- Luisi, P.L., 2006. *The Emergence of Life*. Cambridge University Press, Cambridge, MA.
- Maturana, H., 1978. Biology of language the epistemology of reality. In: Miller, G.A., Lenneberg, E. (Eds.), *Psychology and Biology of Language and Thought: Essays in Honor of Eric Lenneberg*. Academic Press, New York, pp. 27–63, Chapter 2.
- Maturana, H., Varela, F.J., 1980. *Autopoiesis and Cognition the Realization of the Living*. D. Riedel Publishing Company, Dordrecht.
- McMullin, B., Varela, F., 1997. Rediscovering computational autopoiesis. In: Husbands, P., Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life*. MIT Press, Cambridge, MA, pp. 38–47.
- Moreno, A., Lasa, A., 2003. From basic adaptivity to early mind. *Evol. Cogn.* 9 (1), 12–30.
- Pattee, H.H., 1973. The physical basis and origin of hierarchical control. In: Pattee, H.H. (Ed.), *Hierarchy Theory*. Braziller, New York, pp. 73–108.
- Pattee, H.H., 1982. Cell psychology: an evolutionary approach to the symbol-matter problem. *Cogn. Brain Theory* 4, 325–341.
- Rosen, R., 1971. Some realizations of (M R)-systems and their interpretation. *Bull. Math. Biophys.* 33, 303–319.
- Rosen, R., 1985. *Anticipatory Systems: Philosophical Mathematical and Methodological Foundations*. Pergamon Press.
- Rosen, R., 1991. *Life itself: A Comprehensive Inquiry into the Nature, Origin and Fabrication of Life*. Columbia Univ. Press, New York.
- Ruiz-Mirazo, K., Moreno, A., 2004. Basic autonomy as a fundamental step in the synthesis of life. *Artif. Life* 10 (3), 235–259.
- Rupin, E., 2002. Evolutionary autonomous agents a neuroscience perspective. *Nat. Rev. Neurosci.* 3, 132–141.
- Smithers, T., 1997. Autonomy in robots and other agents. *Brain Cogn.* 34, 88–106.
- Solé, R.V., Munteanu, A., Rodriguez-Caso, C., Macía, J., 2007. Synthetic Protocell Biology: from reproduction to computation. *Phil. Trans. Royal Soc. London B*, 362, 1727–1739.
- Steels, L., Brooks, R.A. (Eds.), 1995. *The Artificial Life Route to Artificial Intelligence: Building Embodied Situated Agents*. Lawrence Erlbaum, Hillsdale, NJ.
- Stewart, J., 1996. Cognition = life: implications for higher-level cognition. *Behav. Process.* 35, 311–326.
- Stewart, J., Gapenne, O., Di Paolo, E.A. (Eds.), 2007. *Enaction: Towards a New Paradigm for Cognitive Science*. MIT Press, Cambridge, MA.

- Thompson, E., 2007. *Mind in Life*. Harvard University Press, Cambridge, MA.
- Tsuda, I., 2001. Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behav. Brain Sci.* 24 (5), 793–847.
- Van der Vijver, G., Chandler, R. (Eds.), 2000. *Closure: Emergent Organizations and Their Dynamics*, volume 901 of the *Annals of the New York Academy of Sciences*.
- Varela, F.J., 1979. *Principles of Biological Autonomy*. Elsevier, New York.
- Varela, F.J., 1988. *Connaître: Les Sciences Cognitives, Tendances et Perspectives*. Editions du Seuil, Paris.
- Varela, F.J., Bourgine, P. (Eds.), 1991. *Toward a Practice of Autonomous Systems*. MIT Press, London.
- Varela, F.J., Maturana, H., Uribe, R., 1974. Autopoiesis: the organization of living systems, its characterization and a model. *BioSystems* 5, 187–196.
- Varela, F.J., Thompson, E., Rosch, E., 1991. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, Cambridge, MA.
- Varela, F.J., Lachaux, J.P., Rodriguez, E., Martinerie, J., 2001. The brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* 3, 229–239.
- von Foerster, H., 2003. *Understanding Understanding: Essays on Cybernetics and Cognition*. Springer-Verlag.
- Walde, P., Wick, R., Fresta, M., Mangone, A., Luisi, P.L., 1994. Autopoietic self-reproduction of fatty acid vesicles. *JACS* 116, 11649–11654.
- Weber, A., Varela, F.J., 2002. Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomen. Cognitive Sci.* 1 (2), 97–125, 29.
- Westerhoff, H., Palsson, B., 2004. The evolution of molecular biology into systems biology. *Nat. Biotechnol.* 22 (10), 1249–1252.
- Ziemke, T., 1998. Adaptive behavior in autonomous agents. *Presence* 7 (6), 564–587.

Xabier Barandiaran*

Kepa Ruiz-Mirazo

*IAS-Research Group, Department of Logic and
Philosophy of Science, University of the
Basque Country, Spain*

* Corresponding author.

E-mail addresses: xabier@barandiaran.net

(X. Barandiaran),

kepa.ruiz-mirazo@ehu.es

(K. Ruiz-Mirazo)