Autonomy: a review and a reappraisal

Tom Froese, Nathaniel Virgo and Eduardo Izquierdo

Centre for Computational Neuroscience and Robotics (CCNR)
Centre for Research in Cognitive Science (COGS)
University of Sussex, Brighton BN1 9QH, UK

{t.froese, n.d.virgo, e.j.izquierdo}@sussex.ac.uk

Abstract

In the field of artificial life there is no agreement on what defines 'autonomy'. This makes it difficult to measure progress made towards understanding as well as engineering autonomous systems. Here, we review the diversity of approaches and categorize them by introducing a conceptual distinction between *behavioral* and *constitutive* autonomy. Differences in the autonomy of artificial and biological agents tend to be marginalized for the former and treated as absolute for the latter. We argue that with this distinction the apparent opposition can be resolved.

1. Introduction

Two major research goals of artificial life are to 1) synt

GOFAI). Thus, this category includes all of those approaches which do not treat the autonomy of living beings as qualitatively (though, perhaps, quantitatively) different from the autonomy of most artificial agents. Three sub-categories can be distinguished:

1) The broadest use of the term ‘autonomy’ can be found in the context of engineering where the study of “autonomous systems” is basically equated with a concern for building robots (e.g. Smithers 1992). Thus, there is a sense in which even remotely controlled mobile robots (e.g. a Mars explorer) can be referred to as “autonomous agents” (e.g. Franklin 1995, p. 37). However, more commonly the notion is used to designate that the robot is engineered so as to be able to interact with its environment without requiring ongoing human intervention (e.g. Nolfi & Floreano 2000, p. 67). Brooks (1991), for example, uses the notion of autonomy to refer to tether-free robots, where all the energy and computational requirements are stored on board. Note that using the term ‘autonomy’ in this broad manner does not exclude agents whose behavior has been completely pre-specified. As such it can be criticized on the basis that the “agent can hardly be said to be autonomous because its behavior is largely dictated by the experimenter” (Nolfi & Floreano 2000, p. 148). A more restrictive notion is used by Pfeifer (1996) who proposes as the first design principle of autonomous agents that “they have to be able to function without human intervention, supervision, or instruction”. Nevertheless, it is clear that these requirements for autonomy are almost trivially fulfilled by many artificial agents and all organisms.

2) It is also often claimed that an autonomous system must be capable of satisfying some goal (or even of generating its own goals). For example, Beer (1995, p. 173) uses the term “autonomous agent” to meannty t be aborh1(nty t b bw5-1aw5-1 funaspec06 Tf)4(ic4(d)1s

natural ways (such as situatedness and robustness), others will not need to be solved since they are artifacts of the traditional approach (e.g. symbol grounding)".

2.2 Constitutive autonomy

This category includes all approaches to autonomy which can be traced to the autopoietic tradition, a movement which originated in theoretical biology in the 1970's (e.g. Varela, Maturana & Uribe 1974; Maturana & Varela 1980), and/or which are generally related to metabolism (e.g. Moreno & Ruiz-Mirazo 1999; Ruiz-Mirazo & Moreno 2000). It is generally claimed that autonomy in living systems is a feature of self-production or *autopoiesis*[2]. However, this restriction of autonomy to living systems is unsatisfactory because we also want to refer to some systems as autonomous even though they are not characterized by metabolic self-production, for example artificial and social systems (Luisi 2003).

Thus, the original account was followed by an attempt to conceptually separate the notion of autonomy from that of autopoiesis. In 1979 Varela published his *Principles of Biological Autonomy*, a book that continues to be an important reference for many researchers (e.g. Di Paolo 2005; Beer 2004; Bourgine & Stewart 2004; McMullin 2004; Ruiz-Mirazo & Moreno 2000), and in which he formulated the 'Closure Thesis' which states that "every autonomous system is organizationally closed" (Varela 1979,

Two main approaches can be distinguished according to whether their target is the 1) computational or 2) chemical domain.

1) The field of computational autopoiesis (McMullin 2004) attempts to explore the nature of living systems with the use of simulations. This research program originated over a decade in advance of the first Santa Fe Workshop on Artificial Life with the publication of a seminal paper by Varela, Maturana and Uribe (1974) in which the authors outline the first model of an autopoietic entity. It has subsequently given rise to a whole tradition of simulating autopoiesis (McMullin 2004). However, the question of whether such research can generate genuine autopoietic systems is still the subject of debate, with some researchers claiming for various reasons that computational entities can not be autopoietic in principle (e.g. Letelier, Marin & Mpodozis 2003; Thompson 2004; Rosen 1991; Varela 1997). Nevertheless it is clear that such modelling research has the potential to clarify some of the key ideas

autonomous or it is not. Nevertheless, there might be ways of treating the constitutive dimension as continuous. Bickhard (2000), for example, holds that an autonomous system is one which actively contributes to its own persistence and that "autonomy in this sense is a graded concept: there are differing kinds and degrees of such 'active contributions'". Barandiaran and Moreno (2006) outline another promising approach when they write that "while self-organization appears when the (microscopic) activity of a system generates at least a single (macroscopic) constraint, autonomy implies an open process of self-determination where an increasing number of constraints are self-generated".

Another possibility would be to measure the dimensions of autonomy along an increase in organizational requirements. For example, one could go from negative feedback, to homeostasis, and finally to autopoiesis[5]. This might make it possible to trace behavioral and constitutive autonomy from what might be called a 'weaker' sense to a 'stronger' sense, a continuum which roughly coincides with a transition from a more technological to a more biological usage of the term, and which finally culminates in a complete restriction of the term's applicability to actual living organisms. However, if this hierarchy of organizational requirements is to be actually useful in measuring autonomy, further work needs to be done to define the terms and their relationships more precisely.

3.2 Life as constitutive and behavioral autonomy

After conceptually teasing the constitutive and behavioral domain of autonomy apart, it is nevertheless quite clear that they do somehow relate in living systems. Varela (1997), for example, relates constitutive autonomy to the behavioral domain: "To highlight autonomy means essentially to put at center stage two interlinked propositions: Proposition 1: Organisms are fundamentally the process of constitution of an identity. [...] Proposition 2: The organism's emergent identity gives, logically and mechanistically, the point of reference for a domain of interactions"[6]. However, it is a non-trivial question as to exactly how the organism distinguished in the constitutive domain relates to its behavior distinguished in the behavioral domain. Moreover, this connection only works for some conceptions of behavioral autonomy, and a more precise definition of how such autonomy relates to living systems is needed before the relationship can be stated more formally.

While such further conceptual clarification is important for the development of a coherent theory of autonomy, it is also of practical interest for current artificial life research. Bourgine and Stewart (2004), for example, conceptualize autopoiesis and cognition as distinct aspects of living systems in such a way that it allows them to refer to artificial agents as 'cognitive' without them having to be autopoietic. This view is clearly a useful theoretical justification for using evolutionary robotics as a methodology for studying behavioral autonomy in the form of cognition (e.g. Harvey *et al.* 2005) without having to address the problem of constitutive autonomy.

[5] Thanks to Barry McMullin for pointing this out. This hierarchy is enhanced when we consider that "an autopoietic machine is an homeostatic (or rather a relations-static) system which has its own organization (defining network of relations) as the fundamental variable which it maintains constant" (Maturana & Varela 1980, p. 79). See also Varela (1979, p. 13).

[6] This was clearly also a part of his vision for ECAL, as is evident in Bourgine and Varela (1992).

Similarly, Beer's (2004) approach to cognition follows directly from an autopoietic perspective on life when two key abstractions are made:

> 1) Focus on an agent's behavioral dynamics. An agent's behavior takes place within its cognitive domain, which is a highly structured subset of its total domain of interaction.
>
> 2) Abstract the sets of destructive perturbations that an agent can undergo as a viability constraint on its behavioral dynamics.

Thus, we assume the existence of a constitutively autonomous agent, but model only its behavior and not the constitutive aspects of its autonomy. In other words, the agent is constitutively autonomous by definition only.

However, there are reasons for holding that in living systems autopoiesis and cognition are more tightly interlinked than the possibility of strict conceptual separation seems to indicate (Bitbol & Luisi 2004). Thus, as Beer (1997) himself makes clear, some of the abstractions made in artificial life research are not completely satisfactory:

> "[T]his explicit separation between an animal's behavioral dynamics and its viability constraint is fundamentally somewhat artificial. An animal's behavioral dynamics is deeply intertwined with the particular way in which its autopoiesis is realized. Unfortunately, a complete account of this situation would require a theory of biological organization, and the theoretical situation here is even less well developed than it is for adaptive behavior. [...] However, if we are willing to take the existence of an animal for granted, at least provisionally, then we can assume that its viability constraint is given a priori, and focus instead on the behavioral dynamics necessary to maintain that existence" (Beer 1997, p. 265).

It is clear from these considerations that, while the general aim of evolutionary robotics is not to study the mechanisms underlying constitutive autonomy, more thou000anecessary tedcate bepriori, cate how natu(of are m)8(ostitu)s ft echan05 Tc 0.2578 T

behaviorally autonomous (than at the start of ECAL, for example). Most of the work that is done in the artificial sciences under the banner of autonomous systems research is providing a wealth of tools of analysis and ways of understanding of how externally defined constraints can be successfully satisfied by increasingly complex artificial agents. However, the vast majority of this kind of research is not tackling the question of how such viability constraints (and, more importantly, an agent's identity) can emerge from the internal operations of those autonomous systems while coupled to their environments, though more work is starting to be done in this area.

Finally, it is important to note that the widespread disregard of the dimension of constitutive autonomy is a serious shortcoming not only for scientific research, but

References

Barandiaran, X. & Moreno, A. (2006), “On what makes certain dynamical systems cognitive: A minimally cogniti

Harvey, I., Di Paolo, E.A., Wood, R., Quinn, M. & Tuci, E. A. (2005), "Evolutionary Robotics: A new scientific tool for studying cognition", *Artificial Life*, **11**(1-2), pp. 79-98

Iizuka, H. & Di Paolo, E.A. (submitted), "Toward Spinozist robotics: Exploring the minimal dynamics of behavioural preference", *Adaptive Behavior*

Kauffman, S. (2000), *Investigations*, New York, NY: Oxford University Press

Kelso, J.A.S. (1995), *Dynamic Patterns: The Self-Organization of Brain and Behavior*, Cambridge, MA: The MIT Press

Letelier, J.C., Marin, G. & Mpodozis, J. (2003), "Autopoietic and (M, R) systems", *Journal of Theoretical Biology*, **222**(2), pp. 261-272

Luisi, P.L. (2003), "Autopoiesis: a review and reappraisal", *Naturwissenschaften*, **90**, pp. 49-59

Maturana, H.R. & Varela, F.J. (1980), *Autopoiesis and Cognition: The Realization of the Living*, Dordrecht, Holland: Kluwer Academic Publishers

Mavelli, F. & Ruiz-Mirazo, K. (2007), "Stochastic simulations of minimal self-reproducing cellular systems", *Phil. Trans. R. Soc. B*, in press

McMullin, B. (2004), "Thirty Years of Computational Autopoiesis: A Review", *Artificial Life*, **10**(3), pp. 277-295

Moreno, A. & Ruiz-Mirazo, K. (1999), "Metabolism and the problem of its universalization", *BioSystems*, **49**(1), pp. 45-61

Nolfi, S. & Floreano, D. (2000), *Evolutionary Robotics: The biology, intelligence, and technology of self-organizing machines*, Cambridge, MA: The MIT Press

Pfeifer, R. (1996), "Building 'Fungus Eaters': Design Principles of Autonomous Agents", in: P. Maes *et al.* (eds.), *Proc. of the 4thInt. Conf. on the Simulation of Adaptive Behavior*, Cambridge, MA: The MIT Press, p. 3-12

Pfeifer, R. & Verschure, P. (1992), "Distributed Adaptive Control: A Paradigm for Designing Autonomous Agents", in: F.J. Varela & P. Bourgine (eds.), *Proc. of the 1st Euro. Conf. on Artificial Life*, Cambridge, MA: The MIT Press, pp. 21-30

Rosen, R. (1991), *Life Itself: A Comprehensive Inquiry into the Nature, Origin and Fabrication of Life*, New York, NY: Columbia University Press

Ruiz-Mirazo, K. & Moreno, A. (2000), "Searching for the Roots of Autonomy: The natural and artificial paradigms revisited", *Communication and Cognition – Artificial Intelligence*, **17**(3-4), pp. 209-228

Smithers, T. (1992), "Taking Eliminative Materialism Seriously: A Methodology for Autonomous Systems Research", in: F.J. Varela & P. Bourgine (eds.), *Proc. of the 1st Euro. Conf. on Artificial Life*, Cambridge, MA: The MIT Press, pp. 31-40

Thompson, E. (2004), "Life and mind: From autopoiesis to neurophenomenology. A tribute to Francisco Varela", *Phenomenology and the Cognitive Sciences*, **3**(4), pp. 381-398

van Gelder, T. & Port, R.F. (1995), "It's About Time: An Overview of the Dynamical Approach to Cognition", in: R.F. Port. & T. van Gelder (eds.), *Mind as Motion: Explorations in the Dynamics of Cognition*, Cambridge, MA: The MIT Press, pp. 1-43

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

Varela, F.J. (1997), "Patterns of Life: Intertwining Identity and Cognition", *Brain and Cognition*, **34**(1), pp. 72-87

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

Weber, A. & Varela, F.J. (2002), "Life after Kant: Natural purposes and the autopoietic foundations of biological individuality", *Phenomenology and the Cognitive Sciences*, **1**, pp. 97-125

Wheeler, M. (1997), "Cognition's Coming Home: the Reunion of Life and Mind", in: P. Husbands & I. Harvey (eds.), *Proc. of the 4th Euro. Conf. on Artificial Life*, Cambridge, MA: MIT Press, pp. 10-19