

## AUTONOMY: WHAT IS IT?

Margaret A. Boden  
University of Sussex

Autonomy is a buzz-word in current A-Life, and in some other areas of cognitive science too. And it's generally regarded as a Good Thing. However, neither the spreading use nor the growing approval have provided clarity.

This Special Issue of *BioSystems* is devoted to clarifying the concept, and to showing how it's being used in various examples of empirical research. Given the obscurity that still attends the concept, however, we shouldn't expect to find that all the 'clarifications' are equivalent--or even mutually consistent. Similarly, we shouldn't expect to find that the notion is applied identically in all the research that's reported here. So this preliminary sketch of the conceptual landscape may be helpful.

Very broadly speaking, autonomy is self-determination: the ability to do what one does independently, without being forced so to do by some outside power. The "doing" may be mental, behavioural, neurological, metabolic, or autopoietic: autonomy can be ascribed to a system on a number of different levels.

This doesn't rule out the possibility of the doing's being affected, even triggered, by environmental events. To the contrary: work in 'autonomous' (i.e. situated) robotics, and in most computational neuroethology (CNE), focusses specifically on a creature's reactive responses to environmental cues. Even research that's based on the theory of autopoiesis, which stresses the system's ability to form (and maintain) itself as a functioning unity, allows that a cell, or an organism, is closely coupled with its environmental surround--so much so, that from one point of view they can be regarded as a *single* system (Maturana and Varela 1980). So autonomy isn't isolation. But it does involve a significant degree of independence from outside influences.

One major "outside influence" which A-Life enthusiasts have in mind to deny is the alien hand of the programmer--and, for autopoietic theorists (though not for situated roboticists), even the engineer/designer. (The Designer in the sky is eschewed too, of course--in favour of biological evolution by natural selection.) The explanatory focus is on the specifics of the system's inherent structure and 'intrinsic' properties, not on any instructions that happen to be imposed on it by an outside agency. Only thus can the system's autonomy be respected, or even posited.

Moreover, the traditions of situated robotics and autopoiesis both deny the role of internal/cerebral representations--not only in computer models, but in living organisms too. Some pioneering, and influential, work in CNE posits such representations (e.g. Arbib 1981, 1987, 2003; Boden 2006: 14.vii). But most, perhaps because it deals with insect behaviour, does not.

As a result, most workers in A-Life and CNE reject GOFAI-based models of mind and behaviour--wherein programs are imposed on general-purpose machines, and internal representations are stressed. All too often, however, this rejection is expressed as scorn, more

ideology than science. Indeed, the editor of a professional newsletter in cognitive science has bemoaned the "frankly insulting" names commonly used by researchers for approaches different from their own, complaining that "The lack of tolerance [between different research programmes in AI/A-Life] is rarely positive, often absurd, and sometimes fanatical" (Whitby 2002). If any such insults have crept into the papers presented in this Special Issue, it's to be hoped that the reader will be more intellectually tolerant than the author. For, quite apart from professional etiquette, one important example of autonomy is best understood in largely GOFAI terms (see below).

Autonomy is a problematic concept partly because it can seem to be close to magic, or anyway to paradox. Self-determination is all very well--but how did the "self" (the system) get there in the first place? If the answer we're offered is that it spontaneously generated *itself*, this risks being seen as empty mystification. A key contribution of some current research on autonomy is that it disarms this paradox. But paradox isn't the only source of difficulty here. The concept of autonomy is problematic also because there are various types, and varying degrees, of independence.

Three aspects of a system's behaviour--or rather, of its control--are crucial here (Boden 1996). (The "system" in question may be a whole organism or a subsystem, such as a neural network or metabolic cycle, or a computer model of either of these.) However, the three aspects don't necessarily run alongside each other, nor keep pace with each other even when they do.

The first is the extent to which response to the environment is direct (determined only by the present state of the external world) or indirect (mediated by inner mechanisms partly dependent on the creature's previous history). The second is the extent to which the controlling mechanisms were self-generated rather than externally imposed. And the third is the extent to which any inner directing mechanisms can be reflected upon, and/or selectively modified in the light of general interests or the particularities of the current problem in the environmental context. In general, an individual's autonomy is the greater, the more its behaviour is directed by self-generated (and idiosyncratic) inner mechanisms, nicely responsive to the specific problem-situation yet reflexively modifiable by wider concerns.

Clearly, then, autonomy isn't an all-or-nothing property. And--even more confusing--the senses in which autopoietic systems or self-organizing networks are autonomous differ from each other, and from the sense in which situated robots are autonomous.

The confusion is compounded because, as the brief remarks above imply, autonomy is closely related to two other notoriously slippery notions: self-organization and freedom. No member of this problematic conceptual trio can be properly understood without also considering the other two.

Let's turn to self-organization, first (Boden 2006: 15.i.b). This is the central feature of life. Not only is it commonly listed as a defining characteristic of life, but all the other properties that are so listed are special cases of it. These vital properties are emergence, autonomy (sic), growth, development, reproduction, evolution, adaptation, responsiveness, and metabolism.

Self-organization may be defined as the spontaneous emergence (and maintenance) of order,

out of an origin that's ordered to a lesser degree. It concerns not mere superficial change but fundamental structural development, which can occur on successive levels of organization. And it is spontaneous, or autonomous, in that it results from the intrinsic character of the system (often in interaction with the environment) rather than being imposed by some external force or designer.

It's commonly, though not always, assumed that the generation *and the functioning* of a self-organized system is wholistic. In other words, they can't be explained as being due to interactions between independently definable sub-parts. Whereas a classical AI program, or a car engine, can be analyzed into separate pieces (procedures, mechanical parts), a self-organized system cannot. Each 'part' is to some degree dependent on other 'parts' for its very existence, and for its identity *as* a 'part' of the relevant type. Theoretical approaches based on autopoiesis are especially likely to stress this aspect of self-organized systems.

In the early days of A-Life, long before the field had received its name, the concept of self-organization was being viewed with mistrust by some pioneers even while they were studying the phenomenon in illuminating ways. William Ross Ashby is a case in point. His *Homeostat* was a major advance in the theory, and modelling, of self-organizing systems (Ashby 1947, 1948; Boden 4.viii.c-d, 15.xi.a). Nevertheless, he suggested that the term "self-organization" should be avoided. To be sure, he sometimes used it, and "self-coordinating" too (Ashby 1960: 10). But he also complained that such phrases were "fundamentally confused and inconsistent" and "probably better allowed to die out" (Ashby 1962: 269). He saw them as potentially mystifying because they imply that there's an organizer when in fact there isn't. Some modern researchers agree, carefully avoiding 'self-organization' and referring instead to *organisation* (or metaorganization): the ability to act as a unified whole (e.g. Pellionisz and Llinas 1985: Sectn. 4.1).

Yet the concept of self-organisation hasn't died out. It's employed today by many workers in A-Life and neuroscience--not to imply some mysterious 'inner organizer,' but to focus on *the spontaneous origin and development* of organisation at least as much as on its maintenance. As used in the papers collected below, the term normally carries that bias. The mystification, if not the marvelling, has lessened largely because computer models of various types of self-organization--from flocking (Reynolds 1987), through neural networks (von der Malsburg 1973; Linsker 1988, 1990), to the formation of cell-membranes (Zeleny 1977; Zeleny et al. 1989)--now exist. Clearly, none of these works by magic.

As for human freedom, commonly regarded as the epitome of autonomy, this too--like the vital properties listed above--is a special case of self-organization. A-Lifers, who concern themselves with organisms well below *Homo sapiens* in the phylogenetic scale, rarely mention it explicitly. Occasionally, they admit that their work doesn't cover it (e.g. Bird et al. 2006: 2.1). But sometimes, their words seem to imply that they confuse it with autonomy as such. That's a mistake. The examples of autonomy considered in A-Life show varying degrees of independence from direct outside control. But none has the cognitive/motivational complexity that's required for freedom (remember the third aspect of autonomy listed above).

That's why the often-scorned GOFAI has got closer to an understanding of freedom than A-Life has done. Freedom is best understood in terms of a particular form of complex

computational architecture (Dennett 1984; Boden 2006: 7.i.g-i). It requires psychological resources wherein reasoning, means-end planning, motivation, various sorts of prioritizing (including individual preferences and moral principles), analogy-recognition, the anticipation of unwanted side-effects, and deliberate self-monitoring can all combine to generate decisions/actions selected from a rich space of possibilities. (In the paradigm case, the choice is largely conscious. But an action may be termed "free" because, given the computational resources possessed by the person in question, it *could* have been consciously considered by them, and the decision could have differed accordingly.)

Compromises of freedom occur (for instance) in the clinical apraxias, in hypnosis, and when someone obeys hallucinated instructions from 'saints' or 'aliens'. All these phenomena, wherein a person's autonomy is significantly lessened, have been helpfully theorized and/or modelled in partly GOFAI terms (Boden 2006: 7.i.h-i; 12.ix.b).

In hypnosis and hallucination, for example, a certain type of high-level self-monitoring is inhibited, leaving the person at the mercy of directives imported from outside or internally generated in an unconsidered way (Dienes and Perner in press). As for apraxia, a brain-damaged patient may be unable to plan a simple task, or to perform the relevant sub-tasks in the correct order; or they may be constantly diverted onto a different task while trying to carry out the first one. These debilitating syndromes involve the inappropriate activation and/or execution of hierarchical action-schemas (Norman and Shallice 1980/86; Cooper et al. 1995, 1996). Such schemas may malfunction in various ways, and/or they may be triggered irrelevantly by pattern-recognition mechanisms that divert control of the action onto unwanted paths. In short, apraxias are being modelled by hybrid systems, implementing both GOFAI and connectionist computations.

These theories/models of human freedom, and of its impairments, are relatively broad-brush. They are joined by a wide variety of A-Life models that seek to show even more precisely *how* autonomy, of various kinds, can occur.

CNE has provided some highly detailed explanations of certain aspects of insect behaviour, for example. The computer models concerned include 'virtual' simulations (e.g. Cliff 1991a,b) and robots (e.g. Webb 1996; Webb and Scutt 2000; Beer 1990). Research scattered across A-Life, connectionism, and neuroscience has offered many intriguing suggestions about spontaneous self-organization from a random base, and has provided demonstrations of this phenomenon too (see Boden 2006: 12.ii, 12.v, 14.vi, 14.viii.c, and 15.vii-viii). Some of these results are highly counterintuitive (e.g. von der Malsburg 1973; Linsker 1988, 1990). Further examples are described/cited in the new papers that follow.

One thing's for sure: autonomy is marked on our intellectual map. And the ambiguity and unclarity that attend the concept haven't swamped the excitement. There's plenty of that, here.

## References:

Arbib. M. A. (1981), 'Visuomotor Coordination: From Neural Nets to Schema Theory', *Cognition and Brain Theory*, 4: 23-39.

- Arbib, M. A. (1987), 'Levels of Modelling of Visually Guided Behavior', *Behavioral and Brain Sciences*, 10: 407-465.
- Arbib, M. A. (2003), 'Rana computatrix to Human Language: Towards a Computational Neuroethology of Language', *Philosophical Transactions of the Royal Society of London A*, 361: 2345-2379. (Special issue on 'Biologically Inspired Robotics'.)
- Ashby, W. R. (1947), 'The Nervous System as a Physical Machine: with Special Reference to the Origin of Adaptive Behaviour', *Mind*, 56: 44-59.
- Ashby, W. R. (1948), 'Design for a Brain', *Electronic Engineering*, 20: 379-83.
- Ashby, W. R. (1960), *Design for a Brain: The Origin of Adaptive Behaviour* (2nd edn., revd.) (London: Chapman & Hall).
- Ashby, W. R. (1962), 'Principles of the Self-Organizing System', in H. von Foerster and G. W. Zopf (eds.), *Principles of Self-Organization* (New York: Pergamon Press), 1962, 255-278.
- Beer, R. D. (1990), *Intelligence as Adaptive Behavior: An Experiment in Computational Neuroethology* (Boston: Academic Press).
- Bird, J., Stokes, D., Husbands, P., Brown, P., and Bigge, B. (2006), 'Towards Autonomous Artworks', unpublished working paper: COGS/CCNR, University of Sussex. (Available from jonba@sussex.ac.uk.)
- Boden, M. A. (1996) 'Autonomy and Artificiality', in M. A. Boden (ed.), *The Philosophy of Artificial Life* (Oxford: Oxford University Press), 95-108.
- Boden, M. A. (2006), *Mind as Machine: A History of Cognitive Science*, 2 vols. (Oxford: Oxford University Press).
- Cliff, D. (1991a), 'The Computational Hoverfly: A Study in Computational Neuroethology', in J.-A. Meyer and S. W. Wilson (eds.), *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior* (Cambridge, Mass.: MIT Press), 87-96.
- Cliff, D. (1991b), 'Computational Neuroethology: A Provisional Manifesto', in J.-A. Meyer and S. W. Wilson (eds.), *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior* (Cambridge, Mass.: MIT Press), 29-39.
- Cooper, R., Fox, J., Farringdon, J., and Shallice, T. (1996), 'Towards a Systematic Methodology for Cognitive Modelling', *Artificial Intelligence*, 85: 3-44.
- Cooper, R., Shallice, T., and Farringdon, J. (1995), 'Symbolic and Continuous Processes in the Automatic Selection of Actions', in J. Hallam (ed.), *Hybrid Problems, Hybrid Solutions* (Oxford: IOS Press), 27-37.

Dennett, D. C. (1984), *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, Mass.: MIT Press).

Dienes, Z., and Perner, J. (in press), 'The Cold Control Theory of Hypnosis', in G. Jamieson (ed.), *Hypnosis and Conscious States: The Cognitive Neuroscience Perspective* (Oxford: Oxford University Press), ch. 16..

Linsker, R. (1988), 'Self-Organization in a Perceptual Network', *Computer*, 21: 105-117.

Linsker, R. (1990), 'Perceptual Neural Organization: Some Approaches Based on Network Models and Information Theory', *Annual Review of Neuroscience*, 13: 257-281.

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

Norman, D. A., and Shallice, T. (1980/86), *Attention to Action: Willed and Automatic Control of Behavior*. CHIP Report 99, University of California San Diego, 1980. (Officially published in R. Davidson, G. Schwartz and D. Shapiro (eds.), *Consciousness and Self Regulation: Advances in Research and Theory, Vol. 4* (New York: Plenum), 1986, 1-18.)

Pellionisz, A., and Llinas, R. (1985), 'Tensor Network Theory of the Metaorganization of Functional Geometries in the Central Nervous System', in A. Berthoz and G. Melvill Jones (eds.), *Adaptive Mechanisms in Gaze Control* (Amsterdam: Elsevier), 223-232.

Reynolds, C. W. (1987), 'Flocks, Herds, and Schools: A Distributed Behavioral Model', *Computer Graphics*, 21: 25-34.

Von der Malsburg, C. (1973), 'Self-Organization of Orientation Sensitive Cells in the Striate Cortex', *Kybernetik*, 14: 85-100.

Webb, B. (1996), 'A Cricket Robot', *Scientific American*, 275(6): 94-99.

Webb, B., and Scutt, T. (2000), 'A Simple Latency-Dependent Spiking-Neuron Model of Cricket Phonotaxis', *Biological Cybernetics*, 82: 247-269.

Whitby, B. (2002), 'Let's Stop Throwing Stones', *AISB Quarterly*, no. 109, 1 (one page only).

Zeleny, M. (1977), 'Self-Organization of Living Systems: A Formal Model of Autopoiesis', *International Journal of General Systems*, 4: 13-28.

Zeleny, M., Klir, G. J., and Hufford, K. D. (1989), 'Precipitation Membranes, Osmotic Growths, and Synthetic Biology', in C. G. Langton (ed.), *Artificial Life: The Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems* (held September 1987), (Redwood City, CA: Addison-Wesley), 125-139.