Artificial mental life

Xabier E. Barandiaran[1] and Ezequiel Di Paolo[2]

[1]Dept. of Logic and Philosophy of Science, University of the Basque Country, Spain
[2]Centre for Computational Neuroscience and Robotics, University of Sussex, UK
xabier@sindominio.net

Work in Artificial Life aimed at informing Artificial Intelligence (Steels & Brooks, 1994, *Artificial Life route to Artificial Intelligence*, Lawrence Erlbaum) has drawn inspiration from biology mainly at two levels: i) a bottom-up modelling approach conceiving cognition as the evolutionary complexification of adaptive behaviour, and ii) appeals to self-organization in the domain of behaviour and neural dynamics in analogy with self-organized chemical and biological processes. But little attention has been paid to the possibility of conceiving (and modelling) behaviour in terms of a self-maintaining organized unity in analogy with minimal forms of proto-cellular (or autopoietic) life. We propose that the behavioural counterpart of a network of self-sustaining chemical reactions should be a network of interactively maintained sensorimotor dissipative structures (*habits*) that emerge from the continuous reciprocal interaction between brain, body and world (and not, as in previous attempts, between molecular processes and neural processes, conceiving the nervous system as *operationally closed*Varela, 1979, *Principles of Biological Autonomy*, Elsevier).

Despite its popularity among pre-Darwinian biologists (such as Aristotle, Lamarck or Bichat), pragmatists and phenomenologists alike (Dewey, Merleau-Ponty) and among pre-computationalist psychologists (like James, Goldstein, Ivo Kohler or Piaget) the notion of habit has received little attention within Artificial Life. Habits posses key properties that make them extremely attractive for modelling the organization of behaviour: a) the structure of habits can be traced back to a fully operational-dynamicist framework, b) they do not presuppose a distinction or a causal priority between perception and action, c) habits are inherently situated or enactive structures cutting across brain, body and environment, d) habits are plastic and malleable, e) habits provide a concrete sense of self-maintenance (they are both cause and effect of their occurrence) potentially implying an intrinsic and a interactive teleology and f) habits can be nested or composed at different scales. This opens up the possibility for an operational notion of what might be called *Mental Life* (Barandiaran, 2007, *The World, the Mind and the Body*, p. 49, Imprint Academic) as the continued formation of a web of habits through sensorimotor interactions whose cohesive self-maintenance constitutes the identity of a *cognitive* (as opposed to barely biological) agent and the world it thereby co-defines.

We use some recent evolutionary robotic models on preference and habit formation (Di Paolo & Iizuka, 2008, *Biosystems*, 91, p. 409) to illustrate and explore the theoretical and philosophical implications of taking sensorimotor habits as the building blocks of behavioural organization. This organization takes the form of an attractor landscape whose stability is homeodynamically maintained through sensorimotor coupling. Mental Life opens up a new object of modelling in its own right, closer to the Aristotelian notion of psyche (or even the Heideggerian notion of Dasein) than to the notion of information processing, adaptive problem solving or weak conceptions of autonomy in robotics. Artificial Mental Life involves a shift from building artificial systems that satisfy externally imposed norms (engineering or evolutionary) to systems capable of generating their own norms: those required to sustain their own behavioural organization. In turn, it can become a source of new research questions to investigate the dynamics of assimilation and accommodation into an existing organization, its shaping by social interactions and institutions, or mental disorders dealing with stability, stress, identity, etc.