# Artificial Mental Life.

## *A study of Brain-Body-World integration through neurocognitive robotics*

**Xabier E. Barandiaran**

http://barandiaran.net
xabier@barandiaran.net

**ABSTRACT:** The present project aims at investigating the theoretical foundations of cognitive science and philosophy of mind through the use of robotic simulations inspired on current theories of large-scale brain activity. The main concepts that will be investigated are those of normative functionality, intentionality and agential identity within the context of contemporary dynamicist, embodied and situated approaches in cognitive science. The connection with phenomenological approaches to mind and world will be explored together with a critical assement of the representational and computational theory of the mind and its evolutionary naturalization. Finally the epistemological status of simulation models of complex generative organizations and the way they transform our understanding of mechanistic explanations in philosophy of science.

**KEYWORDS:** Philosophy of Mind, Neurocognitive Robotics, Dynamicism, Embodiment, Situatedness, Simulation Models, Complex Systems.

## Table of Contents

# Research background: the mind as a complex generative mechanism and its study through neurocognitive robotics

## *From computational representationalism to embodied and situated neurodynamics*

The subject-object and mind-body dichotomies have permeated the history of philosophy from Plato to Descartes, from Kant to contemporary philosophy of mind and cognitive sciences. The problems posited by these dichotomies for a unified understanding of our world and the place of our experience on it are the source of some of the most important questions in philosophy. However, it is not until recently that the mechanisms underlying the generation of mindful properties have been amenable to scientific scrutiny and the possibility to create artificial devices with mind-like properties has become a reliable strategy to disclose the nature of the mind and cognition. Ongoing advances in the fields of neuroscience and robotics have transformed and informed renewed philosophical positions on the subject and have opened the way for more accurate and naturalizable foundations for a theory of the mind and cognitive science.

The representationalist and computationalist theory of the mind (Putnam 1960, Fodor 1975) has long dominated philosophy with profound influences on the fields of artificial intelligence, psychology and neuroscience. This theory states that cognition involves a subject-object dichotomy in which the cognizer (the subject) manipulates inner representational tokens of a computational nature. Representations (as units of cognitive organization) acquired their content or semantics (understood as a referential relationship with external states of affairs) by inferential procedures or through evolutionary history (Millikan 1984, Dennett 1995) and they are systematically transformed according to rational rules in order to deliver command actions into the world. Some recent trends in cognitive science (whose philosophical roots can be traced back to the phenomenological approach of Heidegger and Merleau-Ponty—see Dreyfus 1993, 2002, Wheeler 2005) have seriously challenged the computational-representationalist theory. In contrast with abstract and disembodied rational thinking, cognition is conceived as ongoing flexible and adaptive behaviour, in which the body's biological and biomechanical properties play an irreducible role, together with the situated nature of behaviour, capable to exploit contextual cues to achieve a complex behavioural repertoire (Hendriks-Jansen 1996, Clark 1997). This new paradigm has gained increasing acceptance through considerable advances on the fields of autonomous robotics (Brooks 1991), computational neuroethology (Cliff 1991), developmental psychology (Smith & Thelen 1994) and large-scale theories of brain functioning (Freeman 2000, Burzsáki 2006, Edelman & Tononi 2001), among others. The conceptual and methodological glue for these approaches is provided by dynamical system theory and its accompanying philosophical new foundations for cognitive science (Port & van Gelder 1995, Beer 1995, van Gelder 1998).

At the epistemological side, functionalist modes of explanation have recently been reconsidered under a revival of mechanistic modes of explanation (Glennan 1996, Wright & Bechtel 2006). Far from the abstract computational paradigm that advocated for a multiple realizability of cognitive processes (disregarding the significance of the material and biological implementation of cognition) the mechanicist turn has centred the debate around the neurobiological basis of cognition (Bechtel 2008). In this sense, it is the rise of neuroimaging techniques what constitutes one of the most promising avenues of research into the nature of the mind. Large-scale brain imaging techniques (EEG, PET, fMRI, etc.) permit the study of the joint activity of distributed neural ensembles that altogether generate cognitive phenomena. New mathematical techniques, particularly the development of dynamical systems theory and functional network theory, provide the tools to disclose the self-organizing structure of neurocognitive processes. Computer simulations of complex neurodynamic processes permit to assemble neurobiological data and electromagnetic recordings into unified models that are capable to reproduce large-scale behaviour of neural processes. In addition, the need to integrate bodily and environmental features into cognitive models has pushed some neuroscientist to use neurobiologically informed robotic models, capable to display cognitive capacities in continuous interaction with the world (Friston et al. 1994—see also Clark 1997 for a review).

Nevertheless, a number of key questions still remain open to philosophical inquery. On the one hand, dynamical systems extend along all types of physical processes. On the other hand, complex cellular and molecular mechanisms extend along all life form. How, then, should cognition be characterized as a specific phenomenon distinct from the mere physical or biological? On the basis of what can we claim intentional and normative properties for mindful systems? If dynamical and situated approaches do not presuppose a causal priority between brain, body and environmental factors... how can the locus of agency be identified? What constitutes the identity of a cognitive agent?

## *Robotic neurodynamic models as tools for philosophical enquiry*

The new paradigm of neurodynamic cognitive robotics offers the opportunity to rethink some key aspects of the ontological and epistemological foundations of cognitive science. On the one hand, the mind might be conceived as a complex generative mechanism that integrates neurodynamic, bodily and environmental processes into a coherent flow. On the other hand, the understanding of this type of complex organizations appears intimately tied to the use of robotic and simulation models. These models are conceived as epistemic mediators, often analytically irreducible due to their characteristic complexity (their chaotic properties, holistic dependencies, sensitivity to small fluctuations and high number of non-linear interactions). As a result, the categories of agential causality, subject-object dichotomy and the very nature of cognitive processes requires a philosophical reformulation, together with traditional notions of representation, intentionality, normativity and agency. In particular, the concepts of representation, information, intentionality and semantics are currently under debate (Freeman 2001, Dreyfus 2002, Wheeler 2005, Calvo-Garzón in press).

Along this context the role of computer simulations of dynamical systems and robotic models for philosophical investigation appears analogous to the role played by logic and computational theory for the representational approach. The extreme complexity of natural systems often requires the exploration of philosophical issues through minimal *conceptual* models. These type of models do not try to fit any empirical target (animal behaviour, neural tissue or otherwise) but simulate artificial systems capable to illustrate or explore conceptual issues (Barandiaran & Moreno 2006a). As Daniel Dennet acknowledges, simulation models can be understood as thought experiments (Dennett 1991, see also Bedau 1998, Di Paolo et al. 2000), and a form of philosophy on its own right: they can be used to approach the essential organization of natural processes by means of its robotic or simulated implementation.

# Hypothesis, Goals and Methods

## *Hypothesis*

It is within the above scientific and philosophical context (and my research trajectory along it) that the present project is located. I have elsewhere addressed some of these philosophical issues, criticizing the inherited conception of information in cognitive neuroscience (Barandiaran & Moreno 2008b), analysing the subject-object dichotomy (Moreno & Barandiaran 2004), addressing the epistemological novelty of complex simulation techniques (Barandiaran & Moreno 2006a, 2007), making explicit the emergence of adaptive and behavioural capacities from living organization (Barandiaran & Moreno 2008a) and addressing the origin of cognitive normativity and identity (Barandiaran & Moreno 2006b). In particular, my PhD thesis was devoted to a bottom-up, naturalist and biologically grounded notion of *Mental Life*.

The central hypothesis of this project is that cognition can be understood through a generative mechanistic model of Mental Life: the far-from-equilibrium and adaptively self-maintaining neurodynamic organization of behaviour. Behaviour, in turn, is understood as a dynamical system that is decoupled but embedded on a living and mechanically articulated sensorimotor body, situated in the world it co-defines by actively and selectively integrating different aspects of its environment. The key into a mechanistic characterization of Mental Life is the notion of an autonomous organization: i.e. the orchestrated functioning of the parts of a mechanistic system so as to preserve and activelly maintain its own structure. The concept of Mental Life is drawn in analogy with the minimal living organization as that of an autonomous self-organized chemical network capable to maintain itself through a flow of matter and energy and adaptively negotiate its viability or survival conditions through the environment (Maturana & Varela 1973, Ruiz-Mirazo & Moreno 2001).

The specificity of cognitive organization relies on its occurrence in the open domain of sensorimotor interactions made possible by the nervous system (rather than the domain of chemical or metabolic reactions characteristic of biological processes). As a result, cognitive or mental organization displays a coherency and identity of its own, achieved through an open interactive developmental history and not reducible to mere biological or metabolic processes or constraints (yet grounded and made possible by them). Neurodynamic dissipative structures (also named dynamic cell assemblies, global neurodynamic attractors, etc.) can be taken as emergent organizational components of cognitive organization as proposed by recent neuroimaging techniques and large-scale models of brain activity (Varela et al. 2001, Fuji et al. 1996, Edelman & Tononi 2000, Freeman 2000, Llinás 2001, Tsuda 2001, Fries 2005). Neurodynamic structures have been shown to appear as the temporal grouping of local clusters of neurons on a large-scale assembly formed in a task-dependent manner, in which the anatomical disposition, the endogenous activity of the brain and its sensorimotor context, spontaneously forms a self-organized pattern of synchronous integration of distributed clusters of neurons. These neurodynamic structures can interactivelly regulate their stability and collectivelly generate an autonomous organization, whose development and historical transformation constitutes the Mental Life of a cognitive agent.

## *Goals*

*The goal of the project is to develop a simulation model of a robotic agent that instantiates the minimal organization of Mental Life in order to address some key aspects of contemporary philosophy of mind and philosophy of cognitive science*. More specifically, the robotic model would be used to achieve more specific goals that constitute the theoretical core of this project:

- **Neurocognitive modelling:** To integrate essential aspects of empirical models in large-scale dynamical neuroscience into a unified conceptual robotic model that captures the essential organization capable of displaying minimally cognitive properties (perceptual categorization and operand learning).
- **Conceptual reconstruction of central notions for the philosophy of mind:** To illustrate and re-conceptualize through the simulation model three central and foundational concepts of philosophy of mind: a) intentionality, b) normative functionality and c) agential identity.
- **Discussion with other philosophical approaches:** a) To critically asses traditional representationalist and computationalist theories of the mind with particular focus on the concepts of intentionality and representation/information (Dretske 1981, 1988) and the evolutionary approach to their normative functionality and meaning (Millikan 1984, Dennett 1995); and b) to connect the proposed model with the Heideggerian notion of *Dasein*, and Merleau-Ponty's conception of *dynamic structures of comportment* and the intentional-arc.
- **Discussion of epistemological issues:** To address the epistemological status of simulation models of complex generative organizations and the way they transform our understanding of mechanistic explanations in philosophy of science.

### *Experimental and modelling design: Evolutionary robotics, dynamical explanation and philosophical interpretation*

The **robotic model** will be developed within the paradigm of Evolutionary Robotics (Harvey et al. 2005) with which I am already familiar[1]. The minimally cognitive task to be modelled consists on a perceptual categorization between two alternative food sources (distinguishable by their colour and energy value) and the operand conditioning of colour association to energy profitability: the agent needs to discriminate profitable from non profitable food sources in a changing environment composed of different food items. This minimalist framework permits to integrate perception, action, categorization and learning into a single task.

The agent's body is a two wheeled circular robot with an array of visual sensors and an internal sensor capable of measuring the profitability of the food sources when it approaches and "eats" them. Basic physical properties of the agent and the environment are modelled (including velocities, momentum, light intensity, etc.). The behaviour

1 I already developed a similar evolutionary robotics model for my MSc dissertation (Barandiaran 2002) and I am currently working with Ezequiel Di Paolo into a similar model.

generating mechanisms (or simulated brain) of the robot will be modelled as a continuous-time recurrent neural network (CTRNN). This type of networks constitute, for a dynamicist approach, a loose equivalent of what Turing machines constitute for computational approaches. This is due to the fact that CTRNNs has been mathematically proved to be capable of approximating any dynamical system (Funahashi & Nakamura 1993) and become, thus, universal dynamic approximators. In order to introduce additional plasticity and neurobiological accuracy the connections between nodes will be subject to homeostatic Hebbian learning mechanisms. This form of synaptic plasticity has been discovered to operate in real brains (Turrigiano 1999) and has already been successfully implemented in robotic architectures (Di Paolo 2000, Di Paolo and Iizuka 2008). Connection weights between the nodes of the network can be initialized at value zero (or at a random value) so that the early sensorimotor activity of the agent is capable to trigger a developmental process that stabilizes the network connectivity so as to solve the minimally cognitive task.

The complexity of the control architecture is impossible to design by hand. In order to achieve an appropriate set of parameters, artificial evolution is used as an optimization technique (Mitchell 1996). This method starts by generating a set of agents whose neural parameters are randomly chosen. A fitness function selects for the best performing agents and the neural parameters of these agents are combined and randomly mutated to create a new generation of candidate agents. The process is repeated until an agent is capable to solve the task. The fitness function (on the basis of which agents are selected every generation) is designed to maximize the energy of the agent (according to the choice of food sources that the agent has made during its simulated lifetime).

Once the control parameters of an agent have been optimized to solve the task, a number of experimental methods are applied to deliver a satisfactory explanation of its behaviour, including behavioural analysis, "lesion" studies and a dynamical system analysis that integrates neural, bodily and environmental features in the production of behaviour. The dynamical explanation of behaviour can then be used to explore conceptual and epistemological questions into the nature of cognitive explanations.

After the modelling and analysis phase, the project incorporates a phase of **philosophical discussion and interpretation**. In particular the resulting model will be examined under Dretske's theory of cognition as information processing (Dretske 1981, 1988). Dretske is here chosen as a paradigmatic example of the computational and representational theory of mind because of his significant influence in cognitive neuroscience and its precise opertional framework. An evolutionary interpretation of semantics will also be assessed, choosing Millikan's influential etyological approach as the target theory (Millikan 1984). The model is particularly well suited to explore this topic since the use of artificial evolutionary techniques permits to explore the relationship between evolutionary theory and cognitive behaviour. Another line of philosophical investigation will involve the connection of the proposed model with the philosophy of mind of Heidegger and Merleau-Ponty. In particular the notion of *Dasein* will be re-interpreted as the neurodynamic organization of continuous agent-environment coupling (following the philosophical development of Heidegger by Dreyfus—1993) and Merleau-Ponty's notion of *the structure of comportment* and the *intentional-arc* will be revisited through the notions of neurodynamic structures and embodied behaviour. The above investigation will finally permit to explore, within the wider context of philosophy of science, the role of dynamical simulation models in cognitive science.

### *Expected Results*

The results of the proposed research project should give rise to at least 6 publications in international journals and 4 presentations to international conferences. The simulation model is planned to give rise to one or two publications on the following journals: *Adaptive Behavior, Artificial Life or Biosystems.* The critical analysis of representational theories of mind will be used to submit a paper to the following journals: *Philosophical Psychology, Mind & Machines or Cognition.* The discussion of evolutionary accounts of normativity and meaning could give rise to a paper for publication at: *Biological Theory, Biology & Philosophy or Bioessays*. Results of the philosophical interpretation of the model in connection with phenomenology will submitted to the journals *Phenomenology and Cognitive Sciences, Metaphilosophy, Journal of Consciousness Studies or International Philosophical Quarterly.* The epistemological analysis of the model and its relationship with neurodynamic theories of brain functioning will giver rise to a paper to be potentially published by the journals *Philosophy of Science, British Journal of Philosophy of Science* or *Synthese.*

Results should also be presented at some of the following conferences: European Conference on Artificial Life, International Conference on Artificial Life, Simulation of Adaptive Behaviour, Conference of the European Philsophy of Science Association, Meeting of the International Society for the History, Philosophy and Social Studies of Biology, Conference of the American Association of Artificial Intelligence, Meeting of the Canadian Society for the History and Philosophy of Science and Meeting of the Cognitive Neuroscience Society.

## References

Barandiaran, X. (2002) *Adaptive Behaviour, Autonomy and Value systems. Normative function in* *dynamical adaptive systems.* MSc dissertation, University of Sussex, Brighton, UK: http://www.ehu.es/ias-research/doc/2002_ba_msth_fin.pdf

Barandiaran, X. & Moreno, A. (2006a) Alife Models as epistemic artefacts. En Rocha, L.M., Yaeger, L.S., Bedau, M.A., Floreano, D., Goldstone, R.L. and Vespignani, A. (eds) *Artificial Life X: 10th International Conference on the Simulation and Synthesis of Living Systems.* MIT Press. pp. 513—519.

Barandiaran, X. and Moreno, A. (2006b) On what makes certain dynamical systems cognitive. *Adaptive Behavior* **14** (2):171—185.

Barandiaran, X. & Moreno, A. (2008) Adaptivity: from metabolism to behaviour. *Adaptive Behaviour,* in press.

Bedau, M.A. (1998) Philosophical content and method of artificial life. In Bynum, T.W. and Moor, J.H. editors. *The digital phoenix: How computers are changing philosophy*. Oxford, UK: Basil Blackwell, pp. 135—152.

Bechtel, W. (2008) *Mental mechanisms*. MIT Press. Cambridge.

Beer, R. D. (1995) A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence* **72**:173—215.

Brooks, R.A. (1991) Intelligence without representation. *Artificial Intelligence* **47**:139—159.

Burzsáki, G. (2006) *Brain rhythms.* Oxford University Press.

Calvo Garzón, F. (in press) Towards a general theory of antirepresentationalism, *The British Journal for the Philosophy of Science*.

Clark, A. (1997) *Being there*. *Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.

Di Paolo, E. (2000) Homeostatic adaptation to inversion of the visual field and other sensorimotor disruptions. In Meyer, J.-A., Berthoz, A., Floreano, D., Roitblat, H., and Wilson, S., editors, *From Animals to Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*, pages 440—449. Harvard, MA: MIT Press.

Di Paolo, E.A. & Iizuka, H. (2008) How (not) to model autonomous behaviour. *Biosystems* **91**(2): 409—423.

Dretske, F. (1981) Knowledge and the Flow of Information. Cambridge: MA, MIT Press.

Dretske, F. (1988) *Explaining Behavior: Reasons in a World of Causes.* Cambridge: MA, MIT Press.

Dreyfus, H. L. (1993) Heidegger's critique of the Husserl/Searle account of intentionality. *Social Research* **63**(1):17-38.

Dreyfus, H. L. (2002) Intelligence without representation – Merleau-Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences* **1**:367—383.

Edelman, G.M. & Tononi, G. (2000) *A universe of consciousness. How matter becomes imagination*.

Fodor, J.A. (1975) *The language of thought.*New York: Thomas Y. Crowell.

Freeman, W. (2000) *How Brains Make Up Their Minds*. Columbia University Press.

Friston, K.J., Tononi, G., Reeke, G.N., Sporns, O., Edelman, G.M. (1994) Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience* **59**(2):229—43.

Fuji, H., Ito, H., Aihara, K., Ichinose, N. and Tsukada, M. (1996) Dynamical cell assembly hypothesis — Theoretical possibility of spatio-temporal coding in the cortex. *Neural Networks* **9**(8): 1303—1350.

Funahashi, K., & Nakamura, Y. (1993) Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks* **6**:801—806.

Glennan, S. (1996) Mechanisms and the nature of causation. *Erkenntnis* **44**: 49–71.

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

Llinás, R.R. (2001) *The I of the vortex. From neuron to self.* Cambridge, MA: MIT Press.

Maturana, H.R. & Varela, F. (1973/1980) *De Máquinas y Seres Vivos: Una teoría sobre la organización biológica*, Santiago de Chile: Editorial Universitaria. English version: Maturana, H. & Varela, F. (1980) Autopoiesis and Cognition: The realization of the living. D. Reidel.

Millikan, R.G. (1984*) Language, Thought and Other Biological Categories*. Cambridge, MA: MIT press.

Mitchell, M. (1996) *An Introduction to Genetic Algorithms.*Cambridge MA: MIT Press.

Moreno, A. & Barandiaran, X. (2004) A Naturalized Account of the Inside-Outside Dichotomy. *Philosophica* **73**: 11—26.

Port, R. & van Gelder, T. (1995) *Mind as motion: Explorations in the dynamics of cognition.* Cambridge, MA: MIT Press.

Putnam, H. (1960) Minds and Machines. In Hook, S. (Ed.) *Dimensions of Mind*, pp.148—179. New York: New York University Press.

Ruiz-Mirazo, K and Moreno, A. (2000) Searching for the roots of autonomy: The natural and artificial paradigms revisited. *Comunication and Cognition-Artificial Intelligence*, **17**(3–4): 209—228.

Smith, L. B. & Thelen, E. (2003) Development as a dynamical system. *Trends in Cognitive Sciences*, **7**(8): 343—348.

Tsuda, I. (2001) Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems, *Behavioral and Brain Sciences***24**(5):793—847.

Turrigiano, G.G. (1999) Homeostatic plasticity in neuronal networks: the more things change, the more they stay the same. Trends in Neuroscience 22:221-228.

van Gelder, T. (1998) The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences***21**(5):615—628.

Varela, F., Lachaux, J.P., Rodriguez, E. & Martinnerie, J. (2001) The Brain Web: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience***2**: 229—239.

Wright, C. & Bechtel, W. (2006) Mechanisms and Psychological Explanation. In Thagard, P. (Ed.) *Philosophy of Psychology and Cognitive Science* (Volume 4 of the Handbook of the Philosophy of Science). New York: Elsevier.

## Resumen

El problema mente-cuerpo juega un papel fundamental en la historia de la filosofía. Durante los últimos años, la teoría representacionalista y computacionalista de la mente ha pretendido haber resuelto dicho problema y ha ejercido una considerable influencia en la fundamentación de las ciencias cognitivas. Sin embargo, durante las dos últimas décadas nuevas corrientes en ciencias cognitivas han hecho peligrar el programa representacionalista. Desde esta perspectiva, la mente no se concibe ya como una manipulación computacional de símbolos representacionales de acuerdo con reglas lógicas o sintácticas, sino como la conducta flexible y adaptativa que emerge de la interacción recíproca entre procesos neuronales, corporales y ambientales. El presente proyecto propone investigar la fundamentación de las ciencias cognitivas y la filosofía de la mente a través de una simulación robótica inspirada en la recientes teorías del funcionamiento dinámico del sistema nervioso. Desde una perspectiva fenomenológica (heredera del pensamiento de Heidegger y Merleau-Ponty), y en oposición a los planteamientos representacionalistas clásicos, se propone explorar una serie de conceptos fundamentales: *funcionalidad normativa*, *intencionalidad* e *identidad agencial*. Asimismo, se pretende analizar los aspectos epistemológicos derivados del uso de modelos dinámicos simulados para el estudio de la complejidad de los sistemas cognitivos. El proyecto entronca con el trabajo previo del solicitante (avalado por más de una docena de publicaciones), así como con las líneas de investigación del centro receptor, de reconocido prestigio internacional en la construcción y análisis de modelos robóticos neurobiológicamente inspirados.