Proyecto de Tesis Doctoral:

Autonomía e Información en Sistemas Cognitivos Naturales y Artificiales

XABIER BARANDIARAN FERNANDEZ [1]

1 Introducción

Es este un proyecto de tesis filosófico, situado en la práctica interdisciplinaria que demarca a las ciencias cognitivas, y que busca construir desde ellas una elaboración teórica al tiempo que revierta sobre el dominio empírico como resultado de una reorganización intrateórica. El objetivo es articular conceptual y empíricamente las nociones de información y autonomía en el sistema nervioso, abarcando campos de investigación que van desde la biorrobótica hasta la teoría de la complejidad, pasando por las neurociencias, la cibernética y la biología teórica.

Metodológicamente la tesis busca integrar la investigación teórica (conceptual y empíricamente sostenida) con metodologías artificiales, valiéndose de simulaciones por ordenador de sistemas neuronales corporizados y situados como herramientas irrenunciables para estudiar los sistemas complejos que configuran los mecanismos cognitivos más básicos.

2 Antecedentes: autonomía e información en el contexto científico y filosófico actual

El estado actual de las ciencias cognitivas está marcado por la creciente crisis del funcionalismo computacional representacionalista (Putnam 1975, Fodor 1987, Block 1980) y el surgimiento de nuevos paradigmas experimentales y conceptuales. La crisis del funcionalismo tradicional viene propiciada por la robótica autónoma y situada (Maes 1991, Brooks 1991 y Pfeifer & Scheier 1999), el paradigma dinamicista en las ciencias cognitivas (Port & van Gelder 1995, Kelso 1995) y lo que Clark (1997) ha denominado la “revolución conexionista inacabada”. A este contexto hay

[1] Copyright © Copyleft 2003, 2004 Xabier Barandiaran:

Permiso para copiar, distribuir y/o modificar este documento bajo los términos de la Licencia de Documentación Libre GNU, Versión 1.2 o cualquier otra versión posterior publicada por la Free Software Foundation; sin secciones invariantes, sin cubierta frontal, sin cubierta posterior. Una copia de esta licencia puede encontrarse en: http://www.gnu.org/licenses/fdl.html; siempre y cuando se mantenga esta nota.

que añadir contribuciones importantes en diversos campos, como la neurociencia (Gazzaniga 1999, Dayan & Abbott 2001, Churchland 1989, Bechtel *et. al*. 2001), la simulación de sistemas complejos y adaptativos (Husbands *et. al.* 1997, Beer 1997, Webb 2001), y la creciente importancia de la biología para la fundamentación de los fenómenos cognitivos, tanto desde la vertiente evolutiva (Millikan 1984) como organísmica (Maturana & Varela 1980, Varela *et. al.* 1991).

En este contexto se abre un espacio en el que una reconceptualización de los conceptos de información y autonomía (y más concretamente su aplicación al sistema nervioso y a los procesos cognitivos) aparece como un tema de investigación de especial relevancia.

Por un lado, el concepto de *autonomía* ha permitido una fundamentación de los conceptos de funcionalidad y representación en la organización básica de la vida como fenómeno alejado del equilibrio (Collier 1999, Ruiz-Mirazo & Moreno 2000, Christensen & Bickhard 2002) al tiempo que articula todo un programa de ingeniería robótica (Maes 1991, Smithers 1997, Prem 2000) sobre el concepto de autoorganización interactiva de la conducta en tiempo real y situada en entornos físicos inestables (a diferencia de la robótica de corte funcionalista que operaba generalmente entornos virtuales o extremadamente simplificados). La autonomía así entendida hace referencia a la capacidad de un sistema para generar internamente su propia identidad definiendo el dominio de las posibles interacciones con su entorno. Por tanto, el sentido básico de la autonomía se manifiesta en aquellos sistemas cuyos comportamientos modifican las condiciones del entorno de manera tal que contribuyan al mantenimiento del *sujeto* de dichas acciones. Es en este sentido que el *hacer* del *ser* autónomo participa en la constitución de dicho *ser*. Por eso, un agente autónomo es un sistema que actúa no sólo por sí mismo, sino también “para sí mismo”, es decir, que es fuente y destinatario de sus acciones y por tanto capaz de generar una normatividad propia que, en su vertiente cognitiva devenga epistémica. De aquí el carácter intrínsecamente *funcional* de los procesos que constituyen y ejercen los sistemas autónomos. En relación a la neurociencia cognitiva, la idea de autonomía hace referencia a la dinámica cohesiva y operacionalmente cerrada (organizada en forma de red recurrente) que caracteriza al sistema nervioso (Maturana & Varela 1980, Varela 1979, Varela 1992), y que puede ser hoy día cuantificada a través de las medidas de complejidad desarrolladas para el análisis de procesos neuronales (Tononi *et.al.* 1998).

Por otro lado, la *información* se ha convertido en el concepto central que está permitiendo transformar un amplio abanico de problemas

científicos en biología, neurociencias, ciencias cognitivas y ciencias de la complejidad. Sin embargo el origen, naturaleza y estructura de la información en el sistema nervioso como origen de los fenómenos cognitivos se enfrenta a problemas y potencialidades aún por explorar.

Entre los problemas más acuciantes está el de cómo integrar la diversidad de usos del concepto de información en neurociencias (Dayan & Abbott 2001, Rieke *et. al.* 1997) y en filosofía de la mente (Dretske 1988) con la naturaleza dinámica del sistema nervioso y de la conducta situada y corporizada. Frente a las dificultades encontradas en esta tarea se ha propuesto el abandono del marco conceptual computacional clásico (y con él el del concepto de representación e información) y sustituirlo por el de la teoría de sistemas dinámicos (Beer 1997, van Gelder 1998, Maturana & Varela 1980, Varela 1992, Keijzer 2002, Chemero 2000) a través de simulaciones dinámicas del sistema nervioso, robótica situada y estudios empíricos neurocientíficos. Sin embargo el concepto de información puede resultar irrenunciable (como muchos filósofos han subrayado —Bechtel 1998, Clark 1997, Kirsh 1991) para reducir la complejidad de la explicación de los procesos neuronales y comprender cómo de éstos emergen las capacidades cognitivas animales y humanas.

3 Descripción del proyecto

3.1 Objetivos

En definitiva el objetivo de esta tesis es investigar qué papel juega el concepto de información para explicar el *surgimiento y mantenimiento de una complejidad creciente en la organización del sistema nervioso*. Se trata por tanto de elaborar un marco conceptual en el que integrar la dinámica neuronal e interactiva (conductual, situada y corporizada) con una organización funcional de los procesos cognitivos. Un marco que permita a su vez integrar estos procesos en el marco más amplio de los procesos biológicos (autoorganizativos y evolutivos), señalando continuidades y rupturas, y, finalmente, en el contexto más específico de las relaciones sociales (y los dominios de interacción lingüísticos). Para esta tarea trabajaremos con la hipótesis de que el concepto de *autonomía*, haciendo referencia al proceso recursivo, global y operacionalmente cerrado (Maturana & Varela 1980, Varela 1992) que caracteriza al sistema nervioso, así como a su relación con el automantenimiento del organismo (Ruiz-Mirazo & Moreno 2000), permite integrar la dinámica neuronal en el contexto global biológico (autoorganizativo, adaptativo y evolutivo). Esto permite a su vez la naturalización de la funcionalidad normativa (Christensen & Bickhard

2002, Barandiaran 2004) y con ello de la normatividad epistémica que ha sido considerado el máximo obstáculo (Kim 1988) para la naturalización de la epistemología misma, tal como propuso Quine (1969).

> Se trata, en definitiva, de dar una explicación *adecuada* de los conceptos de información y autonomía en el sistema nervioso que permita caracterizar la especificidad de lo cognitivo

Este objetivo requiere analizar una serie de problemas (1) el problema de la relación entre mecanismos neurodinámicos y conducta cognitiva, (2) el problema de la normatividad en procesos naturales, (3) el problema de la doble dimensión (instructiva y epistémica) de la información, y (4) el problema del incremento y sostenibilidad de la complejidad en los procesos neuronales y cognitivos. Durante la investigación se busca además (5) reflexionar sobre el valor epistemológico y metodológico de las tecnologías de simulación computacional de sistemas biológicos y (6) extraer consecuencias relevantes para la epistemología naturalizada y la filosofía de la mente.

Para alcanzar ese objetivo, en esta tesis se pretende desarrollar un plan de trabajo en la que (i) la construcción de los conceptos se relacione con problemas empíricos actuales, de manera que (ii) se pueda dar razón de cómo surge la información y la autonomía en la historia natural; al tiempo que la investigación: (iii) revierta sobre el dominio empírico, mostrando su utilidad instrumental a la hora de reconceptualizar espacios empírico-experimentales y (iv) resulte sintéticamente operativa a la hora de *producir* sistemas artificiales

3.2 Descripción metodológica

La naturaleza de los sistemas integrados (como las redes neuronales altamente interconectadas) exige un tratamiento científico sintético, frente al analítico tradicional de descomposición y análisis de componentes. Es aquí donde el uso de simulaciones computacionales se impone como requisito fundamental para un tratamiento científico y filosófico del objeto de estudio. La biorrobótica (Webb 2001), la vida artificial (Langton 1996) y la simulación de conducta adaptativa (Beer 1997) se convierten así en disciplinas fundamentales para elaborar los conceptos de información y autonomía, integrando la complejidad interactiva y estructural de la que surge la cognición. En definitiva se trata de utilizar modelos artificiales como herramientas conceptuales (Dennett 1995, Di Paolo *et. al.* 2000), instrumentos de producción de hipótesis, pruebas de concepto, y reorganización conceptual. En concreto se

realizará una simulación de conducta cognitiva mínima. La simulación consistirá en la evolución artificial de una red neuronal con plasticidad sináptica como controlador de un robot simulado que deberá realizar una tarea cognitiva. Este trabajo se realizará siguiendo la metodología de la robótica evolutiva en entornos cognitivos mínimos (Husbands *et. al.* 1997, Beer 2003) y continuando con el trabajo ya desarrollado por el doctorando en la universidad de Sussex (Barandiaran 2002).

Los conceptos explicativos desarrollados en el análisis y la síntesis de estos modelos mínimos y en el trabajo conceptual que los acompaña serán a su vez volcado sobre el dominio empírico para ponerlos así a prueba. En concreto se preve analizar y conceptualizar el trabajo empírico neurobiológico desarrollado con animales modelo en neurociencias cognitivas (cuyos sistemas nerviosos han sido extensamente analizados en relación a interacciones cognitivas con su entorno).

3.3 Plan de trabajo

La planificación está comprendida en un proceso de tres etapas principales que se recorren repetidamente cada año académico:

- **Revisión bibliográfica**: Lectura de artículos científicos y filosóficos en filosofía de las ciencias cognitivas, neurociencia, robótica autónoma y teoría de sistemas.
- **Elaboración**: Elaboración y construcción teórica (conceptual) y de modelos simulados (programación).
- **Contrastación**: Contrastación de los modelos y marcos teóricos desarrollados mediante la puesta a prueba de los mismo con la literatura científica y la observación y análisis de comportamiento de simulaciones.

4 Bibliografía

Barandiaran, X. (2002) Adaptive Behaviour, Autonomy and Value Systems. Normative function in dynamical adaptive systems. Master's thesis, COGS, University of Sussex, Brighton, UK, 2002.

Barandiaran, X. (2004) Behavioural Adaptive Autonomy: A milestone in the Alife route to AI? *Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems, ALIFE'9* Boston, September 12th-15th, 2004. MIT Press, in press.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist challenge in cognitive science. *Cognitive Science,* **22:**295—318.

Bechtel, W., Mandik, P., Mundale, J. & Stufflebeam, R.S. (Eds.) (2001) *Philosophy and the Neurosciences*. Blackwell.

Beer, R. D. (1997) The Dynamics of Adaptive Behavior: A research program. *Robotics*

and Autonomous Systems, **20**:257—289.

Beer, R.D. (2000) Dynamical approaches to cognitive science. *Trends in Cognitive Sciences* **4**(3):91—99.

Beer, R. D. (2003) The dynamics of active categorical perception in an evolved model agent. Adaptive Behavior, **11**(4): 209—243.

Block, N. (1980). Introduction: what is functionalism? Block, N. (Ed.) *Readings in philosophy of psychology* **1**: 171—184. Harvard University Press, Cambridge, MA.

Brembs, B., Lorenzetti, F.D., Reyes, F.D., Baxter, D.A. & Byrne, J.H. (2002) Operant Reward Learning in *Aplysia*: Neuronal Correlates and Mechanisms. *Science* **296**:1706—1709.

Brooks, R.A. (1991). Intelligence without representation. *Artificial Intelligence Journal*, **47**:139—160.

Carew, T.J. (2002) Understanding the consequences. *Nature* **417**: 803—806.

Chemero, A. (2000) Anti-Representationalism and the Dynamical Stance. *Philosophy of Science*, **67**(4): 625—647.

Christensen, W.D. & Bickhard, M.H. (2002) The process dynamics of normative function. *Monist*, 85 (1):3-28, 2002.

Churchland, P.S. (1989) *Neurophilosopy: Towards a Unified Science of the Mind-Brain*. MIT Press.

Clark, A. (1997) *Being There: putting, body and world together again*. MIT Press, Cambridge, MA.

Collier, J. (1999). Autonomy and Process Closure as the Basis for Functionality. In Chandler, J.L.R. & van de Vijver, G., (Ed.), *Closure: Emergent Organizations and their Dynamics.* Volume 901 of the New York Academy of Sciences.

Dayan, P. & Abbott, L.F. (2001*) Theoretical Neuroscience. Computational and Mathematical Modeling of Neural Systems*. MIT Press.

Dennet, D. (1995) Artificial Life as Philosophy. In Langton, C. (Ed.), *Artificial Life. An overview*, pages 291—2. MIT Press, Cambridge, MA, 1995.

Di Paolo, E.A., Noble, J. & Bullock, S. (2000) Simulation Models as Opaque Thought Experiments. In M.A. Bedau, J.S. McCaskill, N.H. Packard, and S. Rasmussen, editors, *Artificial Life VII: The 7th International Conference on the Simulaiton and Synthesis of of Living Systems*.

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

Fodor. J.A. (1987) *Psychosemantics*. MIT Press, Cambridge, MA.

Gazzaniga, M.S. (Ed.) (1999). *The New Cognitive Neuroscience: Second Edition*. MIT Press.

Husbands, P., Harvey, I., Cliff, D., & Miller, G. (1997) Artificial Evolution: A New Path for Artificial Intelligence? *Brain and Cognition*, **34**:130—159.

Kandel, E.R. (2001) The Molecular Biology of Memory Storage: A Dialogue Between Genes and Synapses. *Science* **294**: 1030—1038.

Keijzer, F. (2002) Representation in dynamical and embodied cognition. *Cognitive Systems Research* **3**: 275—288.

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

Kim, J. (1988) What is naturalized epistemology? In Martín Alcoff, L. (Ed.) *Epistemology: The Big Questions*, pp.265—281. Blackwell, 1999 edition.

Kirsh, D. Today the earwig, tomorrow man. *Artificial Intelligence*, **47**:161—184, 1991.

Langton, C. (1996) Artificial Life. In M. Boden, editor, *The Philosophy of Artificial Life*, pages 39—94. Oxford University Press, Oxford, 1996.

Maes, P, (ed.) (1991) *Designing Autonomous Agents*. MIT Press.

Maturana, H.R. & Varela, F.J. (1980). Autopoiesis. The realization of the living. In H. Maturana and F. Varela, editors, *Autopoiesis and Cognition. The realization of the living*, pages 73-138. D. Reidel Publishing Company, Dordrecht, Holland.

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

Perrins, R. & Weiss, K.R. (1996) A Cerebral Central Pattern Generator in *Aplysia* and Its Connections with Buccal Feeding Circuitry. *Journal of Neurosciences* **16**(21): 7030—7045.

Pfeifer, R. & Scheier, C. (1999). *Understanding Intelligence*. MIT Press.

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

Prem, E. (2000). Changes of representational AI concepts induced by embobied autonomy. *Communication and Cognition Artificial Intelligence*, **17**(3—4):189—208.

Putnam, H. (1975). *Mind, language, and reality*. Cambridge, Cambridge University Press.

Quine, W.V.O. (1969) Epistemology Naturalized. In *Ontological Relativity and Other Essays*. Columbia University Press.

Rieke, F., Warland, D., van Steveninck, R.R. & Bialek, W. (1997) *Spikes. Exploring the neural code*. MIT Press, Cambridge, MA.

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

Smithers, T. (1997). Autonomy in Robots and Other Agents. *Brain and Cognition*, **34**: 88—106.

Tononi, G., Edelman, G., and Sporns, O. (1998). Complexity and coherency: integrating information in the brain. *Behavioural and Brain Sciences*, **2**(12): 474—484.

van Gelder, T. (1998) The dynamical hypothesis in cognitive science. *Behavioural and Brain Sciences*, **21**:615—665.

Varela, F. (1979). *Principles of Biological l Autonomy*. North-Holland, New York.

Varela, F. (1992) Autopoiesis and a biology of intentionality. In B. McMullin, editor, *Proceedings of a workshop on Autopoiesis and Percetion*, pp. 4—14.

Varela, F.J., Thompson, E. & Rosch, E. (1991) *The Embodied Mind. Cognitive science and human experience.* MIT Press, Cambridge MA.

Webb, W. (2001) Can robots make good models of biological behaviour? *Behavioural and Brain Sciences*, **24**:1033—1050.