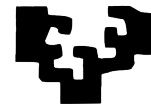


Xabier E. BARANDIARAN FERNANDEZ

eman ta zabal zazu

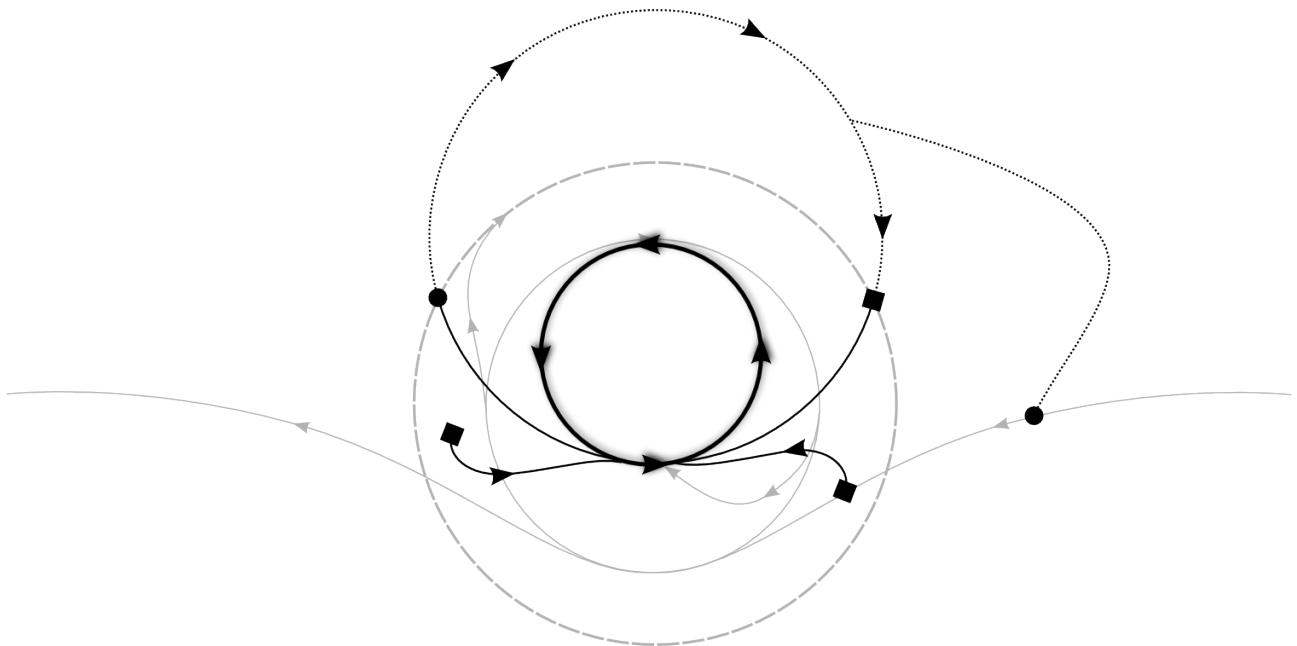


Supervised by:

Prof. Alvaro MORENO BERGARECHE

Mental Life

A naturalized
approach to the
autonomy of
cognitive agents



Submitted for the degree of Doctor of Philosophy,
April 2008, UPV-EHU, University of the Basque Country,

Para Laura

I hereby declare that this submission is my own original work and that, to the best of my knowledge, it contains no material previously published or written by another person, except where due acknowledgement has been made in the text. I also declare that this thesis has not been previously submitted, either in the same or different form, to this or any other university for a degree.

Xabier E. Barandiaran Fernandez.

Copyright © 2008 Xabier E. Barandiaran Fernandez <xabier@barandiaran.net>
Mental Life. A naturalized approach to the autonomy of cognitive agents. v.1.1
<http://www.barandiaran.net/phdthesis/>

This work is available under the terms of the GNU Free Documentation License: *You are free to use, copy, modify and redistribute this work provided this notice is preserved.*
The full copyright license can be found at: <http://www.gnu.org/licenses/fdl.html>



AUTORIZACIÓN DEL DIRECTOR DE TESIS PARA SU PRESENTACION

Dr. Alvaro Moreno Bergareche como Director/a de la Tesis Doctoral:
“Mental Life. A naturalized approach to the autonomy of cognitive
agents” realizada en el Departamento de Lógica y Filosofía de la
Ciencia por el Doctorando Don Xabier E. Barandiaran Fernandez,
autorizo la presentación de la citada Tesis Doctoral, dado que reúne las
condiciones necesarias para su defensa.

En Donostia – San Sebastián a ____ de Abril de 2008

EL DIRECTOR DE LA TESIS

Fdo.: _____



CONFORMIDAD DEL DEPARTAMENTO

El Consejo del Departamento de Lógica y Filosofía de la Ciencia en reunión celebrada el día ____ de Abril de 2008 ha acordado dar la conformidad a la admisión a trámite de presentación de la Tesis Doctoral titulada: “Mental Life. A naturalized approach to the autonomy of cognitive agents” dirigida por el Dr. Alvaro Moreno Bergareche y presentada por Don Xabier E. Barandiaran Fernandez ante este Departamento.

En Donostia – San Sebastián a ____ de Abril de 2008.

Vº Bº DIRECTOR/A DEL
DEPARTAMENTO/

SECRETARIO/A DEL
DEPARTAMENTO/

Fdo.:

Fdo.:



ACTA DE GRADO DE DOCTOR ACTA DE DEFENSA DE TESIS DOCTORAL

DOCTORANDO DON. Xabier E. Barandiaran Fernandez

TITULO DE LA TESIS: *Mental Life. A naturalized approach to the autonomy of cognitive agents*

El Tribunal designado por la Subcomisión de Doctorado de la UPV/EHU para calificar la Tesis Doctoral arriba indicada y reunido en el día de la fecha, una vez efectuada la defensa por el doctorando y contestadas las objeciones y/o sugerencias que se le han formulado, ha otorgado por _____ la calificación de: *unanimidad ó mayoría*

En Donostia – San Sebastián a ____ de _____ de 2008

EL/LA PRESIDENTE/A,

EL/LA SECRETARIO/A,

Fdo.:

Dr/a: _____

Fdo.:

Dr/a: _____

VOCAL 1º

VOCAL 2º

VOCAL 3º

Fdo.:

Dr/a: _____

Fdo.:

Dr/a: _____

Fdo.:

Dr/a: _____

EL DOCTORANDO,

Fdo.

Summary

Recent advances in modelling complex adaptive systems through computer simulation have reconfigured the way in which mechanistic explanations can conceptualize the mind. The goal of the thesis is *to make explicit the construction of a model for Minds as a complex generative organization*. For doing so, and under the difficulties that current approaches find to specify cognitive systems as distinct from generic computational or dynamical ones, the construction of a model for minds departs from a minimalist, universal and naturalized conception of agency. Three conditions are state to provide a satisfactory account of agency: self-generated individuality, normativity and causal-asymmetry. A *morphophylogenetic* approach is proposed whereby increasingly complex models of agency are analysed. The morphophylogenetic reconstruction depart with the origins of life as the emergence of a self-encapsulated chemical network capable of actively producing itself. The *autonomous organization* of living organisms is shown to ground a naturalized conception of normative functionality, individuality and agency. Taking biological autonomy as a departure point the thesis covers the main organizational evolutionary transitions of agency that lead to Bilaterian organisms. A number of case studies (*E. coli*, *A. digitale* and *C. elegans*) are provided to illustrate different aspects of such transitions until *adaptive behaviour* (made possible by multicellular organisms endowed with a nervous systems and mechanically articulated bodies) is precisely defined and its modelling process made explicit. However, a number of problems are found for the project of grounding intentionality and mindfulness in biological (metabolic) normativity. An alternative research avenue is proposed in which it is the *autonomy of behaviour* (and not that of its underlying infrastructure) what serves to naturalize intentional agency. A conceptual model of *Mental Life* is proposed: the autonomous and normatively regulated organization of a web of sensorimotor neurodynamic structures, capable of generating a distinctive level of coherency and identity in continuous interaction with the world. An analysis of the evolution of brain, body and behaviour suggests that there is an evolutionary pathway towards more plastic, diverse and integrated agency leading to Mental Life. It is argued that current developmental cognitive neuroscience reveals the progressive organization of the brain through self-sustained behavioural interactions as opposed to the maturation of genetically pre-specified cognitive modules that would preclude the emergence of Mental Life. Large-scale models of neural dynamics are reviewed supporting the view of an integrated dissipative organization of neural activity and its self-regulatory capacity. Finally, it is shown how sensorimotor deprivation regimes may lead to a disintegration of behavioural organization revealing the deep entanglement between world and mind for the continued self-maintenance of the latter.

Publications

As Editor

- Barandiaran, X. & Ruiz-Mirazo, P. (2008) Modelling autonomy. Special Issue *BioSystems Journal* **91**(2).

Articles in Journals

- Barandiaran, X. & Moreno, A. (2008) Adaptivity: from metabolism to behaviour. *Adaptive Behaviour*, accepted.
- Barandiaran, X. & Ruiz-Mirazo, P. (2008) Modelling autonomy. simulating the essence of life and cognition. *BioSystems* **92**(1): 295—304. <http://dx.doi.org/10.1016/j.biosystems.2007.07.001>
- Barandiaran, X. & Moreno, A. (2008) On the nature of neural information. A critique of the received view 50 years later. *Neurocomputing* **71**:681—692. <http://dx.doi.org/10.1016/j.neucom.2007.09.014>
- Barandiaran, X. & Guiu, L. (2006) Autonomía, Comunicación y Evolución en redes bacterianas y tecnológicas. *biTARTE* **38**:35—60.
- Barandiaran, X. and Moreno, A. (2006) On what makes certain dynamical systems cognitive. *Adaptive Behavior* **14**(2):171—185.
- Moreno, A. & Barandiaran, X. (2004) A naturalized account of the inside-outside dichotomy. *PHILOSOPHICA* **73**:11—26.

Book Chapters

- Moreno, A., Ruiz-Mirazo, K. & Barandiaran, X. (2008) The impact of the paradigm of complexity on the foundational frameworks of biology and cognitive science. In Hooker, C. (Ed.) *Philosophy of Complex Systems* (Vol. 10 of Handbook of the Philosophy of Science), New York: Elsevier, to appear.
- Barandiaran, X. (2007) Mental Life: conceptual models and synthetic methodologies for a post-cognitivist psychology. In Wallace, B., Ross, A., Davies, J. & Anderson, T. (Eds.) *The World, the Mind and the Body: Psychology after cognitivism*, Imprint Academic, pp. 49—90.

Conference Proceedings

- Barandiaran, X. & Moreno, A. (2008) Modelos Simulados, mediación virtual para el pensamiento complejo. *Ontology Studies / Cuadernos de Ontología* **7**: in press.

- Barandiaran, X. & Moreno, A. (2006) Una perspectiva naturalizada del concepto de información en el sistema nervioso. In Fernandez-Caballero, A. Gracia, M, Alonso, E. and Tomé, S.M. (Eds) *Actas del Congreso Internacional Campus Multidisciplinar en Percepción e Inteligencia*, Vol. I, Universidad de Castilla la Mancha, pp. 194—206.
- Barandiaran, X. and Moreno, A. (2006) Alife Models as Epistemic Artefacts. *Proceedings of the 10th International Conference on Artificial Life*. MIT Press, Boston, Massachussets, pp. 513—519.
- Barandiaran, X. (2004) La robótica evolutiva como epistemología experimental naturalizada. In *IV Congreso de la Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España*, pp. 124—127.
- Barandiaran, X. (2004) Behavioral Adaptive Autonomy. A milestone in the Alife route to AI?. In Pollack, J., Bedau, M. Husbands, P., Ikegami, T. & Watson, R. (Eds.) *Proceedings of the 9th International Conference on Artificial Life*, MIT Press, pp. 514—521.
- Barandiaran, X. and Feltrero, R. (2003). Conceptual and methodological blending in cognitive science. The role of simulated and robotic models in scientific explanation. In *Volume of abstracts of the 12th International Congress of Logic, Methodology and Philosophy of Science*, Oviedo (Spain), August 7-13, 2003. p. 171.

Acknowledgements

I would like to express my intellectual debt with the San Sebastian group of Philosophy of Biology (IAS-research). All the ideas presented here are, one way or another, a continuation of the work that has inspired me within that group for several years now. Thanks to Arantza, Jon, Kepa and specially to Alvaro Moreno for creating such an inspiring intellectual environment. The Centre for Computational Neuroscience and Robotics (at University of Sussex) has been my second intellectual home. The present work could not had been possible without the warm and inspiring stances at the CCNR where so many people (thanks Ezequiel, Eduardo and Marieke!) have made me feel part of a collective endeavour for discovering the secrets of life and mind.

A number of colleagues and researchers have contributed to this thesis through valuable discussions of some central ideas and by suggesting key reference readings. I am grateful to Anthony Chemero, Arantxa Etxeberria, Chrisantha Fernando, Cliff Hooker, Eduardo Torres-Izquierdo, Ezequiel Di Paolo, Gorka Zamora, Inman Harvey, James Bernat, Jamie Twycross, Jesús Siqueiros, Jon Umerez, Juan Vidal, Juanba Bengoetxea, Kepa Ruiz-Mirazo, Margaret Boden, Marieke Rohde, Rachel Wood, Thomas Hughes Castañeda, Simon McGregor, Steve Torrance and Tomás García. In addition, some of them have gently revised early (and often terribly written) chapters of the thesis. The final result owes much to their careful and dedicated corrections: Aita, Arantza Etxeberria, Borja Esteve, Ezequiel Di Paolo, Jon Umerez, Juan Vidal, Juanba Bengoetxea and Kepa Ruiz-Mirazo. Edorta Txopitea and Paul Feith have gently responded to my last-minute requests for solving problems with vocabulary and grammar. Without Maite's help I would be now helplessly lost on a bureaucratic nightmare.

Many others have provided emotional support and psychological care for this intense (at time neurotic) enterprise: my office colleagues for such a long time (Leire and Jesús but also intermittently Marila, Kepa and Juanba), friends and flatmates (Aurre, Ibai, Jesús, Egaña, Azalai, Santi, Cañete, Patxangas, ...) and many others (the Txoko crew!) but particularly my family (Ama, Aita and Jone) were always supportive and comprehensive throughout the many breakdowns and stressful situations that the writing of this work has made me through. Absorbed working on my thesis has made me disengage, for too long, from close friends, communities (Txoko, Metabolik, SinDominio, Hacklabs, Hackmeeting, CEB, 23 ...) and my family: demasiado tiempo sacrificado por esta tesis, os he echado muchísimo de menos, espero que haya merecido la pena!

Two persons deserve a special mention. One is Alvaro Moreno. He embarked on this trip long ago. His philosophical conversations at occasional dinners at home captured my attention when I was still a child. I could not stop critically inquiring the world since then. Alvaro has not only supervised this work with exclusive dedication and care but has also channelled my almost infinite drift and has always offered opportunities to deepen and refine every single line of thought. Last, but not least, he has taken care of the author not only of his thesis (a rare exception in the academic jungle) and sew the threads of destiny for me to meet Laura. She is the second person that deserves a special mention. Laura has read, discussed, corrected and contributed greatly to almost every piece of this work. She has been the most supportive, intelligent and comprehensive mate I could ever had imagined; and this work has stolen too much of myself and our time. Así que esto va por tí. Maite zaitut.

I am finally grateful to the free-software community for developing, supporting and documenting all the computer tools necessary and sufficient for doing this work. I have also tremendously benefited from anonymous feeders who have created a high quality on-line peer-to-peer scientific library. I will always be grateful to their everyday honourable effort to scan and compile scientific textbooks in order to make the power of knowledge accessible to anyone. Open Access initiatives, like the Public Library of Science or the WormBook research community, are also drawing the contours of a scientific era where scientists and society can liberate their knowledge from the tyranny and monopoly of big publishing companies. Thank you all.

Funding for this work was provided by grants: 9/UPV 00003. 230–3707/2001 from the University of the Basque Country and BMC2000–0764 and HUM2005- 2449/FISO from the Ministry of Science and Technology. In addition the author had benefited from the economic support of the doctoral fellowship BFI03371-AE from the Basque Government.

Table of Contents

PART I: A QUESTION OF METHOD

Chapter 1: The ghost and the machine	23
Chapter 2: Complex mechanism, some methodological considerations	31
1. Behaviourist, functionalist and mechanistic explanatory paradigms	31
2. From atomistic reduction to emergent organizations	36
2.1. Internal emergent functionality	41
2.2. Interactive emergence	44
2.3. Chaos and fluctuations	45
2.4. Hierarchies, levels, mechanisms and organization	47
3. Modelling complexity: a synthetic bottom-up approach	50
3.1. Mechanistic-empirical models	54
3.2. Functional models	56
3.3. Generic Models	58
3.4. Conceptual models	59
3.5. Animal models and in vivo synthetic approaches	62
4. Complex thinking and its foundational impact on biological and cognitive sciences	63
Chapter 3: What a Concept for Minds Requires	67
1. Models, Theories and Concepts	67
2. Conceptual models as explanatory and generative definitions	71
3. Minimalism, Universality and Naturalism: A MUNdane declaration of principles.	74
3.1. Naturalism	74
3.2. Universalism	75
3.3. Minimalism	76
3.4. MUN-modelling: the epistemic engine	77
4. Enclosing cognition: the problem of delimiting boundaries	78
5. What agency requires	82
5.1. Individuality	84
5.2. Causal asymmetry	85
5.3. Normativity	86
6. A Morphophylogenetic Approach: the route for reconstruction	87

PART II: THE MORPHOPHYLOGENY OF AGENCY

Chapter 4: Life: minimal agency in basic autonomous systems	91
1. Types of cohesion	92
2. Dissipative order as self-organization	94
3. Towards organized complexity	97
3.1. Metabolism: chemical component production networks	97
3.2. Replication: template molecules	99
3.3. The early exploration of the complexity space	103
4. Basic autonomous systems: constructive and interactive cycles	105

5. CASE STUDY: Characterization of minimal autonomy through a simulation model	109
5.1. The constructive cycle	111
5.2. Interactive processes	112
6. Individuality and environment in minimal autonomous agents	116
7. Normativity and functionality in minimal autonomous agents	119
8. Causal asymmetry in minimal autonomous agents	124
9. Life and its intrinsic agential nature: minimal agency naturalized	127
Chapter 5: Adaptivity and sensorimotor coupling in chemodynamic autonomous agents	131
1. The limits of proto-metabolic robustness without explicit regulatory control	131
2. Adaptivity and decoupled control mechanisms	135
3. Recurrent adaptive interactions: minimal agency revisited	141
4. CASE STUDY: Bacterial Chemotaxis	143
5. Motility: the enaction of an environment	148
6. Organizational bottlenecks of motile agency in uni and multicellular organisms	151
Chapter 6: The Nervous System: origins, evolution and organization of Behavioural Agency	157
1. The Origins of the NS	158
2. CASE STUDY: A science fusion journey with Jelly Fish <i>Aequorea victoria</i>	164
3. Cnidarian agency	168
4. The evolutionary development of the early NS and Bilaterian agency	173
5. CASE STUDY: a science fusion journey with <i>Caenorhabditis elegans</i>	178
6. The Hierarchical Decoupling of the Nervous System	184
7. Dynamical modelling of the neurodynamic domain	189
8. The embodiment and situatedness of neural dynamics	196
9. The adaptive organization of embodied and situated neurodynamics.	205
10. RECAPITULATION: Agency revisited through adaptive behaviour.	215
 PART III: THE MIND HAS A LIFE OF ITS OWN	
Chapter 7: Failed intentions. Adaptive behaviour as intentional agency?	221
1. The promised land?	222
2. Cognition as closed sensorimotor loop	224
2.1. The sensorimotor “world”	225
2.2. Meaningful problems: behaviour without frustration	229
3. Biological grounding of the sensorimotor loop	234
3.1. Living as a process is a process of cognition	235
3.2. Making it explicit: adaptive mechanisms need find their place	237
4. Phenomenology meets the mechanisms of adaptive behaviour	242
4.1. Towards a minimal structure of intentional agency	243
4.2. Phenomenological refutations	248
5. Interrogating mechanisms: adaptive behaviour and intentionality	249
5.1. The case of metabolism-independent chemotaxis	250
5.2. The case of adaptive evaluation of behaviour	252
5.3. The case of adaptive initiation of behaviour	256

6. Some in-principle problems for the biological grounding of intentional agency and a disjoint continuity	258
6.1. Intentional guts?	259
6.2. The problem of dissociation between behavioural mechanisms and metabolic norms	260
6.3. The mind has a life of its own	265
Chapter 8: Mental Life: the autonomy of behaviour	269
1. Scape from mind preclusion: two fables and a jail	269
1.1. C. elegans, locked inside the jail of adaptive constraints	269
1.2. The dialectics of subordination and under-determination. A fable on academic research	271
1.3. A bite on cognitive or behavioural autonomy, a fable on canine rage and the freedom-reflex	273
2. The Autonomous Organization of Behaviour: unpacking a conceptual model for Mental Life	275
2.1. Searching for autonomy in behaviour	275
2.2. A bundle of habits	279
2.3. Neurodynamic structures and hyperdescriptions	284
2.4. Interactively dependent stability: a case study in homeostatic plasticity and a single habit	286
2.5. Extending the case of single habit	290
2.6. The autonomy of behaviour revisited	293
3. The evolutionary origins of Mental Life: environment, behaviour, brain and body	295
3.1. How evolution might open up a complex behavioural and environmental space for Mental Life	296
3.2. Trends in neocorticalization: evidence for the evolution of complex behaviour generating mechanism	303
3.3. Encephalized bodyplans: co-evolution of embodied brains and embrained bodies	306
4. Neurocognitive development: becoming Mental Life	309
4.1. Some general genetic and developmental considerations	310
4.2. Embryonic (prenatal) brain-body development	311
4.3. Embodied and situated behavioural development	313
4.4. Neural constructivism	315
5. Mental Life in Action (or 2001 a brain odyssey into the cosmic dance)	317
5.1. Micro-Meso-Macro: the emergence of integrative scales	320
5.2. Neurodynamic cortical structures in the gamma-band	323
5.3. Embodiment, emotions and appraisals: the integration of the sensorimotor and the interior nervous system	329
5.4. Autonomy and normative regulation in neurodynamic organization	333
6. Mental Death: the effects of sensorimotor deprivation and solitary confinement	336
Chapter 9: Recapitulation and conclusion	343
1. Extended Schematic Summary: from concepts to complex generative models, from life to mind.	343
2. A Concept for Minds	353
References	357

PART I

A Question of Method

The third was to carry on my reflexions in due order, commencing with objects that were the most simple and easy to understand, in order to rise little by little, or by degrees, to knowledge of the most complex, *assuming an order, even if a fictitious one, among those which do not follow a natural sequence relative to one another.*

RENÉ DESCARTES

Chapter 1: The ghost and the machine

It is a common habit among philosophers to choose a particular thinker with remarkable influence on the history of thought and attribute him the responsibility of the intellectual limits, pitfalls and disgraces of our time. An evil mind whose unfortunate efforts on the foundations of philosophy precluded a happy unfolding of thought in history and condemned humanity to inescapable conceptual prisons. Heidegger took Plato and Descartes to be the great beast of metaphysics, for Nietzsche it was the Christian tradition, and Jerry Fodor focused his wrath on Hume. But when it comes to philosophy of mind, the favourite target today is, doubtless, René Descartes. His legacy still stands with us, as Michael Wheeler (2005) has compellingly uncovered recently, and its consequences extend far beyond the, nowadays unacceptable and unaccepted, metaphysical dualism.

According to the Cartesian inheritance at the core of the received view in cognitive science¹, the mark of the mental is, characterized by a number of key assumptions. Following Michael Wheeler's synthesis (2005) the pervading Cartesian mindset conceives cognition as: (i) a subject-object dichotomy in which the cognizer (the subject) manipulates (ii) according to the rules of reason (logical, linguistic, etc.) (iii) inner representational and states, tokens of an immaterial nature: conscious-phenomenological, computational, or otherwise, (iv) whose content is acquired by inferential procedures and (v) used to plan (process, deduce, transform) in order to execute actions in the world or to satisfy its mindfulness with the delight of truthful transformations.

This abstract, rationalist and disembodied conception of mind is what Gilbert Ryle termed the intellectualist myth in philosophy, not just confined to a small group of pro-Cartesian fans but “the official doctrine”, as Ryle labelled it, shared by philosophers and the people in the street. As he noted in *The Concept of Mind* (which stands as one of the earliest and surgically more precise exercises of analytic philosophy against the Cartesian mindset) this legacy stands on “the idea that the capacity to attain knowledge of truths was the defining property of a mind. Other human powers could be classed as mental only if they could be shown to be somehow piloted by the intellectual grasp of

1 That of computational functionalism and symbolic Artificial Intelligence.

true propositions. To be rational was to be able to recognise truths and the connections between them. To act rationally was, therefore, to have one's non-theoretical propensities controlled by one's apprehension of truths about the conduct of life" (Ryle, 1949:26). On the contrary, Ryle continues, the mind is characterized by the behavioural dispositions towards a *know-how* involving a non-theoretical *capacity for behaving correctly on everyday activity* rather than an intellectual *know-what* determining action (as if the boxer was to be theoretically planning a correct and truthful defence before getting punched by the reality of the face-crunching musculo-skeletal ability of his opponent).

Following Ryle, the paradoxes of the Cartesian legacy stand on a *category mistake*. This category mistake originated due to the difficulty or impossibility of describing mental phenomena under the terms of the mechanical regularities that all bodies were discovered to be subject to. The Cartesian mindset assumed a para-mechanical analogy by which the mind is a non-mechanical non-material object, subject to non-material cause and effect, undergoing non-material processes. Interestingly terms like "object", "cause and effect" or "process" were uncritically reconstructed creating a para-mechanical mental domain. The paradoxes arise, according to Ryle, when we take both domains (the corporeal and the mental) to belong to the same type of categories (the same logical class) and we start combining them as opposed (the mind can be reduced to the body) or as causing each other (my intention caused my hand to move). Since the negation of a logical type (mind = non mechanism) belongs to the same logical category, the construction of the mind as the "non-mechanical" created a fundamental mistake, whose consequences, Ryle argues, we are still suffering.

A similar category mistake arises if we intend "right-hand and left-hand gloves" to belong to the same logical type as "a pair of gloves" and ask, for example, what do I have to buy other than the right-hand globe and the left-hand glove in order to get a pair of gloves. Another example of category mistake would be to claim (at the same logical level) that "there is an enjoyable mood in the room" and that "there is a couple of chairs in the corner of the room" where the first claim belongs to the class of social "atmospheres" while the second relates to the disposition of physical objects in space. Wondering how could an enjoyable mood cause the existence or the movement of a couple of chairs involves a category mistake, as much as it does the claim that the mind is the cause of intelligent behaviour. Ryle's proposal is precisely that intelligent behaviour *is* the mind: "overt intelligent performances are not clues to the workings of minds; they are those workings" (Ryle, 1949: 58).

The mistake involved on conceiving minds and mechanisms as belonging to the same conceptual category leads to what Ryle baptised as "the ghost in the machine": the idea that the mind is somehow a mysterious kind of sub-

stance operating within our bodies and being responsible for our actions. If we are to judge someone as intelligent or mindful we cannot do that by presupposing a ghost, responsible for its intelligence, but by looking at his or her actions directly:

[W]hen we characterise people by mental predicates, we are not making untestable inferences to any ghostly processes occurring in streams of consciousness which we are debarred from visiting; we are describing the ways in which those people conduct parts of their predominantly public behaviour. True, we go beyond what we see them do and hear them say, but this going beyond is not a going behind, in the sense of making inferences to occult causes; it is going beyond in the sense of considering, in the first instance, the power and propensities of which their actions are exercises. (Ryle, 1949: 51).

But how should we understand this particular description of public behaviour that we call intelligence? It is certainly not, Ryle continues, a singular action that we can call intelligent. A unique action could occur randomly. Someone that has never used a gun could (by chance) shot in the middle of the target. A single shoot will never be enough to determine if she is a good shooter not even sufficient to say that she intended to do so. It is rather to capacities, dispositions, abilities, habits, inclinations, aptitudes, propensities, powers (of which specific actions are but actualizations, realizations) that, Ryle argues, we refer when talking about the mind.

Fundamentally the Rylean approach states that mental properties are dispositional attributes. It is often considered that dispositional terms refer to states of things and that, consequently, there must be a mental state that bears the causal responsibility of doing and saying certain things. On the contrary, Ryle proposes to interpret dispositional terms as a set of determinable counterfactuals (that can take different values). Dispositional terms are used as open hypothesis of the form *if ... then ...*. For instance, if we say that “Aznar knows English” by which we mean an open set of hypothesis of the type “If Aznar reads a newspaper in English he will understand it”. Dispositional propositions permit to draw inferences of the above kind. But such inferences do not necessarily imply a third term that comes to justify it. For instance, it would be a mistake to imply that the deduction of “Aznar will understand the newspaper” from “Aznar knows English” requires it to be something in “his mind” that is the cause of his understanding the newspaper. The postulate of a third term is, Ryle defends, absolutely unnecessary: “to know English” is to satisfy a number of counterfactuals.

But not all the dispositional propositions are of the mental type. “Sugar is soluble” is a dispositional proposition (“If you put a lump of sugar in your coffee *then* it will dilute”). A crucial difference stands for Ryle on the fact that mental capacities are learned, whereas the mechanical or biological are not. The mind is, according to Ryle, the set of dispositional attributes (attitudes, abilities, capacities, etc.) generally acquired by learning or training and the facts or

occurrences that fulfil them (the acts). And all of this does not require the postulation of mechanisms that intend to justify their causes. We do not need postulate “solubility” as a cause of dilution to infer that the lump will dilute from the dispositional proposition that “sugar is soluble”.

The Rylean analysis of the cathedral, or rather the bazaar, of linguistic uses and practices concerning mental vocabulary is overwhelming. The distinction between two classes of incommensurable logical categories seems safe. Until we need/want to change or modify the counterfactual properties that we attribute to the sugar lump or to our own minds. If we want to avoid a lump of sugar to dilute under certain conditions, then, highlighting the unnecessary appeal to intermediary causes becomes an obstacle. What we need to do is to modify the internal molecular structure of sugar so that we can transform its disposition to dissolve. It is certainly true that this molecular structure is not an “intermediary cause” in the sense of a billiard ball between other two, in a linear chain of causes, between “solubility-cause” and “dilution-effect”. We can say that Susan can fall into depression. For Ryle that means that *if* some conditions are met, e.g. Susan loses her job, she will fall into depression. It would be misleading to say that there is a hidden mental cause of “depressibility” that is the cause of its capacity to fall into depression (like when mental theories affirm that “It is John’s desire to have a child” and postulate that John has a mental, ghost-like, desire, as if he had an incorporeal object inside that is the cause of John having a child if the appropriate conditions were met). When we deal with a system whose internal organization is unknown, or remains in principle inaccessible to us, it is epistemically dangerous to assign internal states to it, internal states that are postulated as reasons or causes of its actions or properties. The Rylean message is that in such cases we need to use the language carefully to avoid paradoxes and conceptual mistakes. The conceptual *vademecum* is to limit ourselves to dispositional propositions, inferred on the basis of observed behaviour. But for those systems of which we know and understand their internal organization (and that means that we might in principle be able to operate upon its dispositions and properties), we can and we *must* refer to its organization as a cause, albeit a special kind of cause (not a sequential intermediary mechanical object “pushing” effects outside the system). If Susan shows signs of falling into depression, her philosopher friends will be of little help with their conceptual analysis precluding intermediate causes. If Susan wants to avoid falling into depression she would better find a remedy on the neurochemical or psychological mechanisms that modulate and generate her mood.

When one needs to operate in order to modify the dispositions of a system it becomes manifest that mechanisms and dispositions need to relate each other through the notion of organization. Keeping dispositions and mechanisms apart, within the domains of incommensurable categorical domains, has

a price to be paid. Ryle paid the price throwing, so to speak, the machine out with the ghost's fog, implicitly assuming the ultimate Cartesian divide: a radical separation between mechanisms and minds. This is notorious in the following passage:

He is bodily active and he is mentally active, but he is not being synchronously active in two different 'places', or with two different 'engines'. There is one activity, but it is one susceptible of and requiring more than one kind of explanatory description. Somewhat as there is no aerodynamical or physiological difference between the description of one bird as 'flying south' and of another as 'migrating', though there is a big biological difference between these descriptions, so there need be no physical or physiological differences between the descriptions of one man as gabbling and another talking sense, though the rhetorical and logical differences are enormous. (Ryle 1949:50—51)

Ryle remains silent on anything that has to do with neurophysiology, and when he mentions it, as in the example above, he makes an unfortunate move claiming that there needs to be no physiological difference between two different (even opposing) mental acts. One can translate mental vocabulary into behavioural vocabulary by testing different counterfactuals. Minds will be thus described in the right dispositional terms, without provoking any category mistake. And yet, the *mechanistic substrate* of any behavioural disposition that the disposition itself vanishes as soon as the neurophysiologist's scalpel makes its lobotomic philosophical refutation. The alternative that I shall explore throughout this work is that the concept of mind does not refer to an open set of counterfactuals but to the set of *organizational* conditions that permits to predict and manipulate those counterfactual cases on the specific mechanisms that make them possible. Denying or ignoring the machine, just because the ghost is a refutable entity, does not lead to a categorically definitive view of the mind, on the contrary. As Dennett himself confesses in the introduction to *The concept of mind*²:

It seems to have been a point of unexamined faith for Ryle that whatever the scientists might learn about mechanisms of the brain, however necessary these were in grounding our behavioural dispositions, they would shed scant light on the questions that interested him. This might had been true, had brains been so much more complicated than banks. (Dennett 2000³: xi)

Paradoxically it seems that Ryle and Descartes shared a common and profound assumption: that there is no possibility that science will illuminate (in terms of discovering mechanisms), in any significant way, what needs to be known about the mind, however important physiology turns out to be for grounding the behavioural dispositions that are characteristic of minds.

² The British 2000 Penguin edition and the Spanish Paidós edition of 2005.

³ Dennett (2000) Re-introducing the concept of mind. Introduction to Ryle (1949/2000) *The concept of Mind*. Penguin Edition.

Interestingly, the very same year as *The concept of mind* appeared, the neurobiologist Donald Hebb published *The organization of behavior*, one of the most influential and pioneering treatises of neurobiology. This book included (for the Rylean's discomfort) a detailed account of the neural mechanisms underlying learning, which was, as we saw earlier, one of the defining characteristics of mental dispositions as opposed to the mechanistic or biological ones. Neuroscience is today one of the big allies of the philosophy of mind (Churchland 2002, Bickle 2003) and I shall repeatedly turn into neuroscience throughout this work. But the Rylean message needs to be retained and recall in the neurophilosophical era: (i) his careful refutation of the intellectualist tradition, the "Official Doctrine", (ii) the emphasis on the primacy of the *know-how* of everyday *life* activity as a behavioural practice and (iii) the care not to postulate anything like mental or spiritual causes for intelligent action, unless we provide a satisfactory account of the mechanisms that support it as a disposition and in the context of behaviour. The scalpel in our hands, I shall now abandon his careful linguistic-category analysis and behaviourist reformulation of mind. For us, post-Hebbian neurophilosophers, reference to the underlying organization is required if the concept of Mind is to have any place in a world of mechanisms.

Descartes was forced to shelter the mind into the realm of the immaterial, non-mechanical, mirroring the logical and causal structure of the mechanical into the domain of *res-cogitans*. He could not envision any imaginable mechanism capable of generating the abstract and content-independent intelligence that the human mind was embedded with. The substance dualism that he therefore adopted has remained for long as an insurmountable cliff between mind and world, cognitive behaviour and mechanism. Whereas for some (Ryleans and behaviourists) the solution involved throwing out the machine with the ghost's fog others embraced the ghostly computational machine to regain credibility for the "intellectualist myth". But mechanistic explanations have undergone considerable transformations through the advent of sciences of complexity⁴ and the time is ready to reconsider a concept for minds that takes seriously into account the complex biological and neural mechanisms that make possible the generation of mindful behaviour. This is the object of this thesis.

Chapter 2 deals with the epistemology of complex mechanisms, it analyses the limitations of the Cartesian decompositional approach to mechanistic explanations and considers how simulation models can approach complex systems in a generative manner. Chapter 3 deals with the relationship between concepts, models and theories and defines the epistemological requirements that a concept for minds should satisfy departing from a generic notion of agency. Chapter 4 explores the minimal organization of life through a simula-

4 We shall shortly see how this occurred and what is meant by the term "sciences of complexity".

tion model of self-organized chemical reactions networks endowed with a self-produced selective membrane. Chapter 5 explores unicellular motility through the case study of *E. coli* bacteria and the set of transitions that lead to multicellular agency. Chapter 6 is focused on the origin and characterization of the nervous system and analyses the notion of adaptive behaviour through two case studies: the jellyfish *A. digitale* and the nematode *C. elegans*. Chapter 7 discusses in detail whether adaptive behaviour is sufficient for intentional agency finding a set of interpretative problems when comparing phenomenological insights into the structure of intentionality-in-action with behaving organisms that fail to be adaptive in specific circumstances. Chapter 8 is devoted to provide a model of Mental Life, exploring the evolution of cognition, developmental cognitive neuroscience and current models of large-scale brain activity. Throughout the previous chapters the reader may want to refer to the schematic summary of the argumentative line of the thesis provided in Chapter 9.

Chapter 2: Complex mechanism, some methodological considerations

When one attempts to visualize the figure that is formed by these two curves and their infinite intersections, (...) they create a sort of network, of web, of infinitely intertwined tissue; (...) One gets perplex by the complexity of this figure that I don't even venture to draw.
POINCARÉ 1890¹

1. BEHAVIOURIST, FUNCTIONALIST AND MECHANISTIC EXPLANATORY PARADIGMS²

The methodological principles of mechanistic explanations can be traced back, once again, to Descartes, whose substance dualism was widely rejected by the time Ryle wrote his book, but whose mechanist methodological principles still permeated the fields of psychology and philosophy of mind (Wright & Bechtel 2006) and continue to do so. It is perhaps the Cartesian conception of mechanism, a conception that required to throw the mind to the nonmechanical realm, what has probably been more influential and difficult to overcome on the quest of understanding the living mind. In his *Discourse on the*

1 Quote borrowed and translated to English from: <http://laberintos.itam.mx/despliegaphp?idart=197>
Original Spanish quote: “Cuando uno intenta visualizar la figura formada por estas dos curvas y su infinidad de intersecciones, (...) éstas forman una suerte de red, telaraña, tejido infinitamente entrelazado; (...). Uno queda perplejo por la complejidad de esta figura que ni siquiera me atrevo a dibujar.” (Poincaré 1890)

2 The following two sections include material from an early draft of Moreno, Ruiz-Mirazo & Barandiaran (2008). The material reused here was written by myself for the early draft although some ideas have been further elaborated by the three of us for the final version of the publication. Explicit reference to the final publication are made for the results of the joint investigation.

Method of Rightly Conducting the Reason and Seeking for Truth in the Sciences

Descartes proposed to conduct scientific explanation by decomposing a system into as many parts as required so that every part could be fully understood in isolation and then assembled again little by little to achieve an understanding of the whole. As we shall see, Descartes' method had a long lasting influence on scientific explanation but also imposed some severe limitations. Current versions of mechanicism are less demanding on decomposition and open up the way for a richer organization of relationships between parts. What remains central to mechanicism is that explanations need to assemble component parts and their operations to reconstruct their "orchestrated" functioning:

A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena. (Bechtel & Abrahamsen 2005: 423)

Broadly speaking mechanistic explanations entail the identification of the parts of a system and their operations so as to provide an explanation of the functioning of the system. This entails the requirement that the system be decomposed and that each component be related to a particular functional contribution to the behaviour of the system. Familiar as it is for all of us who have ever opened up our computer, car or coffee machine to fix it, the mechanistic strategy is not without its problems. The complexity of biological and cognitive systems imposes a number of difficulties on the attribution of functions to structures and their decomposition in ways that become particularly relevant when attempting to deliver a mechanistic understanding that makes justice to the adaptive, flexible and creative capacities that are so fascinating and characteristic of mindful systems.

Throughout this section I shall illustrate how the sciences of complexity, a family of models and modelling tools, have transformed the notion of mechanism, re-conceptualizing the relationship between structure and function. As a result we have nowadays access to the notion of "emergent integrated organization" made possible through a new set of modelling tools that computer simulations have provided. The limits of Cartesian top-down analytic decomposition will open up the way for a synthetic bottom-up methodology that might come to fill those gaps on the machine where the ghost was thought to have a privileged status.

In order to approach the nature of mechanistic explanations in more detail a set of distinctions need to be introduced beforehand. First, there are two levels of description and decomposition that shall occupy us: structural and functional. One might proceed by a structural decomposition of parts isolating the components that may physically be identifiable within a system (e.g. atoms in a molecule, planets in the solar system, tissues in an organism, agents in social organization, etc.). At this level, the identification of the parts

is usually achieved directly by their separability (due to well defined boundaries, distance between parts, differential bounding forces and alike) or by types of properties that distinguish types of units (shape, size, reactivity, staining, etc.). In biological and cognitive sciences the branch of knowledge that deals with this decomposition is anatomy, where the main task is to distinguish and label structural parts.

The other mode of decomposition is functional, it is carried out at a higher abstract level and involves the division of labour of the overall behaviour of the system into component sub-operations or functions. Behavioural descriptions precede functional ones and are somehow required or assumed by any functional explanation or decomposition. Both behavioural descriptions and functional decomposition lead to two distinct paradigms within the understanding of the mind: behaviourism and functionalism. **Behaviourist “explanations”** are limited to establish the set of input-output correlations on a system’s behaviour (i.e. relationships between observed changes on the environment, conceived as “stimuli”, and observed changes in the system’s surface, conceived as “response”). In psychology, behaviourism (Watson 1914, Skinner 1938) had a strong influence, from the 20s to the 70s, against the phenomenological or introspective tradition that performed poorly against the strong epistemological standards of the nascent philosophy of science of the time. In this sense the early behaviourism (strongly influenced by the physicists, operational and a law-based conception of science) consisted on the establishment of quantifiable and metaphysically “neutral” stimulus/response correlations and regularities. However, behaviourism can vaguely be said to provide a full sense of explanations on its own, since the methodology is, a priori, constrained to state observed regularities on a history of stimulus-response (input-output) relationships, independently of how these regularities appear due the structures that generate them. Yet, behavioural descriptions (even better when classified and systematized under quantitative regularities) are required for functional descriptions and, ultimately, for any explanation. In particular, for mechanistic explanations, an adequate behavioural description provides the framework of what is to be explained: the context or container for an analysis of parts and functions.

Functional explanations, on the other hand, involve the subdivision of behaviour (or the capacity³ of a system) into intermediate sub-operations so that the overall behaviour is the combination of such component sub-operations (Cummins 1975). We can take the example of a coffee machine. The behavioural description states that there is a causal correlation between the input of coffee, water, electricity and the pushing of some buttons, that leads to

3 The term capacity is here to denote the set of dispositions (potentially available behaviours) that a system can exhibit. The object of functional analysis and explanation is not limited to the range of shown behaviours but could, in principle, also explain capacities.

the output of a certain amount of brownish liquid through a small tube. In other words: given the appropriate conditions, the Coffee machine has the capacity to make coffee. Assuming certain previous knowledge on how to make coffee (or by analysing the properties of the input and the properties of the output) one can make a tentative functional decomposition of what set of sub-operations the machine must carry on in order to output drinkable coffee. One could assume the following functional decomposition: a) coffee beans need to be ground, b) water needs to be heat so that ground coffee can be properly mixed with it and c) filtering is required to get the final liquid clean of dregs. One might assume that these three functional steps are required necessarily and independently of how the machine proceeds on its details, i.e. on the mechanistic structural implementation of the overall functionality of the coffee machine. For functionalism, localization of function into structural (physically identifiable and distinct) parts is somehow superfluous; what matters is the functional causal organization, decomposed into abstracted operations and operators.

Some have even convincingly argued for purely functional explanations that abstract away from the underlying structure, often under the assumption that functional relationships are multiply realizable, thus rendering the underlying supporting structure of functional relationships irrelevant. In cognitive science, the computer metaphor provided a strong support for purely functional explanations of the mind (Putnam 1960, Fodor 1968, Block 1996). Two key discoveries made possible a purely functionalist research program in cognitive science. First, the Turing-Church thesis in which a single physical mechanism is proven to be able to carry on any mathematical or formalizable function that can be effectively calculated (Turing 1936, Church 1936). The construction of *Universal Turing Machines* (i.e. computers), using the von Newman architecture, came soon and made possible for a single mechanisms to be capable of implementing the functions of any other mechanism (thus the adjective “universal”). The second discovery involves the hypothesis that neurons act as logical gates or boolean functions able to implement a Universal Turing Machine (McCulloch & Pitts 1943). As a result, the view that psychological phenomena could be studied just at the functional level of description and that explanations could be delivered as computational operations between functional states, became a widespread assumption. This view came to be reinforced by the formalization of human reasoning into propositional logic and the possibility of its implementation in computational machines, together with the development of generative grammars based on Turing Machine equivalent transformation rules (Chomsky 1965). Reason and language, the most superior capacities characteristic of minds, became susceptible to computational-functionalist explanations.

However, the problem of functionalism remains to discover which is the functional decomposition and organization that a system really has. Let me recall the case of our coffee machine's functional decomposition that we previously composed of three operations: ground, dilution and filtering. If one looks up more carefully, it might be possible that coffee beans be ground through high temperature and pressured water, and that a filtering mechanism not really required if the final liquid is left to stand and only the top liquid is taken as output. What appeared as a reasonable decomposition of behaviour in three functional sub-operations might well be achievable by just two sub-operations. Without opening the machine and looking at its structural composition it becomes extremely difficult to infer which among the possible functional decompositions is really taking place. This means that functional decompositions, without a corresponding structural analysis, implies a high degree of speculation. Independently of its relationship with structural decomposition it is difficult (if not impossible) to find a criterion that determines which among the whole possible sets of functional decompositions is the appropriate one to explain the behaviour of a system. Simply put, functional analysis is under-determined by behaviour: the same behaviour can be achieved by different functionally decomposable sub-operations.

One can try to escape from this indeterminacy. Functional decomposition can be guided by historical, optimality or logical constraints that might be assumed to determine the functioning of a system, or by disambiguating between possible functional decompositions through behavioural error patterns or reaction times (Wright & Bechtel 2006). For instance, if the coffee making process is very fast one might infer that there is no sufficient time for coffee dregs to be separated spontaneously from water and that some kind of active filtering process needs to take place. In addition, optimality considerations might suggest that coffee beans are mechanically ground, since it is much more efficient (e.g. in terms of energy and resources) than doing it through pressured water. One can thus conclude that although other functional decompositions may be logically possible the three-step decomposition must be the right one. But, ultimately, none of these heuristics will provide a definitive argument for what the case really is. Optimality considerations rest on the assumption that the system under study has the possibility to explore (through its designer, evolution or ontogenetic adaptation) the full space for optimal solutions. In addition, optimality calculations cannot predict that formally optimal mechanisms may suffer material constraints of various sorts and get trapped into some suboptimal configuration. Even even assuming that no material constraints apply, the system might need to optimize competing functions whose interplay may well lead to a wide range of equally stable (but never totally optimal) configurations. Only the real system has the answer.

To avoid functional decompositional uncertainty, scientists look for mechanistic explanation in which identifiable structures permit an accurate and empirically testable reference for functional analysis and provide the ultimate ground for a satisfactory explanation. In contrast with behaviourism or functionalism, mechanistic explanations have a much stronger compromise with the underlying structure. In fact, *mechanistic explanations* can be best viewed as a mapping between the more abstract functional decomposition and the more concrete and empirically accessible structural decomposition. As a result, the relevant functional operations can be related to distinguishable structural components, what Bechtel and Richardson (1993) have named “localization”. However, structural decomposition (deciding upon what constitutes a part) might not always be easy to achieve, there are often an undefined set of possible decompositions of a system into parts (Wimsatt 1974). On the other hand, as seen earlier, a parallel problem is faced by functional decomposition since a given behaviour (without additional information) might be decomposed into an open set of functional sub-operations. A coupled functional and structural process of investigation into the mechanisms that give rise to a particular phenomena can solve the problem, providing the means for a mutual constraint between each mode of decomposition. A tentative functional decomposition might inform a possible decomposition of parts. In turn, observable parts and their local operations can constraint and narrow the set of potential functional decompositions. Mechanistic explanations thus involve a circular process of decomposition and localization of parts-functions, so that an adequate decomposition and localization of the internal operating processes is achieved to explain the functioning of the system under study. This circulation between structural and functional decomposition and mapping is further facilitated by the possibility to manipulate and measure variations on local and global behaviour of parts and wholes.

Biological and cognitive sciences have adopted different forms of mechanistic explanation with differing degrees and possibilities for decomposition, localization and manipulation. Different levels and degrees of complexity transform the methods and theoretical frameworks by which we come to understand such systems. I shall take care of this transformation and modes of mechanistic explanation in the following section.

2. FROM ATOMISTIC REDUCTION TO EMERGENT ORGANIZATIONS

As we noted above, it is the Cartesian method itself that introduced a powerful, yet limited, procedure to follow in scientific explanations: analytic decomposition of parts and its (linear) aggregation. I will call this basic mode of

mechanistic explanation *atomistic reductionism*⁴; where the properties/behaviour of the whole are taken to be reducible to the sum of the properties/behaviour of its component parts. It is precisely in the domains of biology and psychology (nowadays including cognitive science) where this method had a greater impact over centuries (e.g. early understanding of reflex arcs and circulatory system). In terms of the mapping between structural and functional descriptions stated above, atomistic reductionism might be understood as a class of mechanistic explanation where a one-to-one mapping between structure and function is established and where there is just a linear aggregation of the part's functions on the composition of the whole. As a result, properties of the whole can be reduced to the sum of the properties of the parts or identified with a single localized component property or a combination of them. What is meant by linear aggregation (or sum) might be best understood as the concatenation of a sequence of operations performed by well identifiable parts. Think on a billiard table scenario: a player inputs momentum to ball A which moves and collides with ball B which in turn transfers momentum to adjacent ball C which moves and falls under a hole until it stops (halts) at the ball repository⁵. The causal structure of the system is captured by a concatenation of operations of one to one momentum transfer starting by the input of the player and terminating when the last ball stops. Many human mechanisms operate in a similar way: a lever moves a gear that in turn makes another gear rotate which in turn displaces a latch and the door opens. The same logic applies to the mechanisms of a coffee machine: coffee beans are inserted, then ground, then a flow of boiling water is passed through the ground coffee and passed through a filter; the espresso comes out of the machine ready to be drunk. Some natural mechanisms might also be described and modelled in a similar vein: a reflex arc, the snap trap of carnivore plants, or the process of digestion from ingestion to secretion, etc.⁶. That the explanation be linear

4 The philosophical debate around reductionism is extraordinarily large and many different ramifications have been discussed in detail (e.g. inter theoretical reduction, emergence of levels of organization, etc.). On what follows, I shall focus on the relationship between structural parts and how their relationships are understood in terms of providing explanations of the functioning or behaviour of the systems. Within this context reductionism will be understood as an extreme in which parts are taken to be “responsible” or local causes of certain aspects of functioning (thus the label “atomistic” to name this type of reductionism). And non-reductionism, or holism, will be considered as the opposite extreme in which it is the global integration between parts what plays the major explanatory role without possible decomposition.

5 Note that a purely functional explanation of the billiard ball could be impossible: i.e. without knowing the exact position of the component balls on the billiard table. One could, for instance, infer the number balls on the table and even gain some knowledge of their distribution just by looking at the input momentum and the time of arrival of balls to the output repository. There would, however, always rest some uncertainty since an indefinite number of trajectories and collisions may satisfy those constraints.

6 Atomistic reduction is favoured by human linear thinking, which in turn may be interpreted as an internalization of language (whose linearity is a direct consequence of sound's single dimensionality).

does not rule out the capacity to explain circular systems like the blood circulatory system. But the explanations of circular systems, under atomistic reductionism, is exhausted in the linear explanation of a single cycle that is then assumed to be repeating itself almost uniformly. Circularity is never a sub-functional process but just the concatenation of the end of the linear explanation with a new beginning.

Some forms of atomistic reduction might also proceed by abstracting away intermediate operations and establishing a functional mapping between specific components of the system and some of the system's behaviour at a higher level of description. For instance, one might get an understanding of the functioning of a car by abstracting away all the mechanisms of the engine and just focusing on the abstract function of the accelerator pedal and the steering wheel. Assuming that the internal mechanism and some contextual conditions do not change (what is generally called a *ceteris paribus* clause) there is in fact a correlation between the acceleration of the car and the pressure on the accelerator pedal. Thus, for certain purposes (such as driving the car) one might carry out a one-to-one atomistic reduction of the relevant functions (turning, acceleration and deceleration) to certain component parts (steering wheel, accelerator pedal and break). A similar reductive explanation has also been popular in biology when assuming, for instance, a one-to-one mapping between genes and phenotypic traits or the reduction of certain cognitive functions (such as memory) to specific components of the brain (e.g. certain types of molecules acting on synaptic connections).

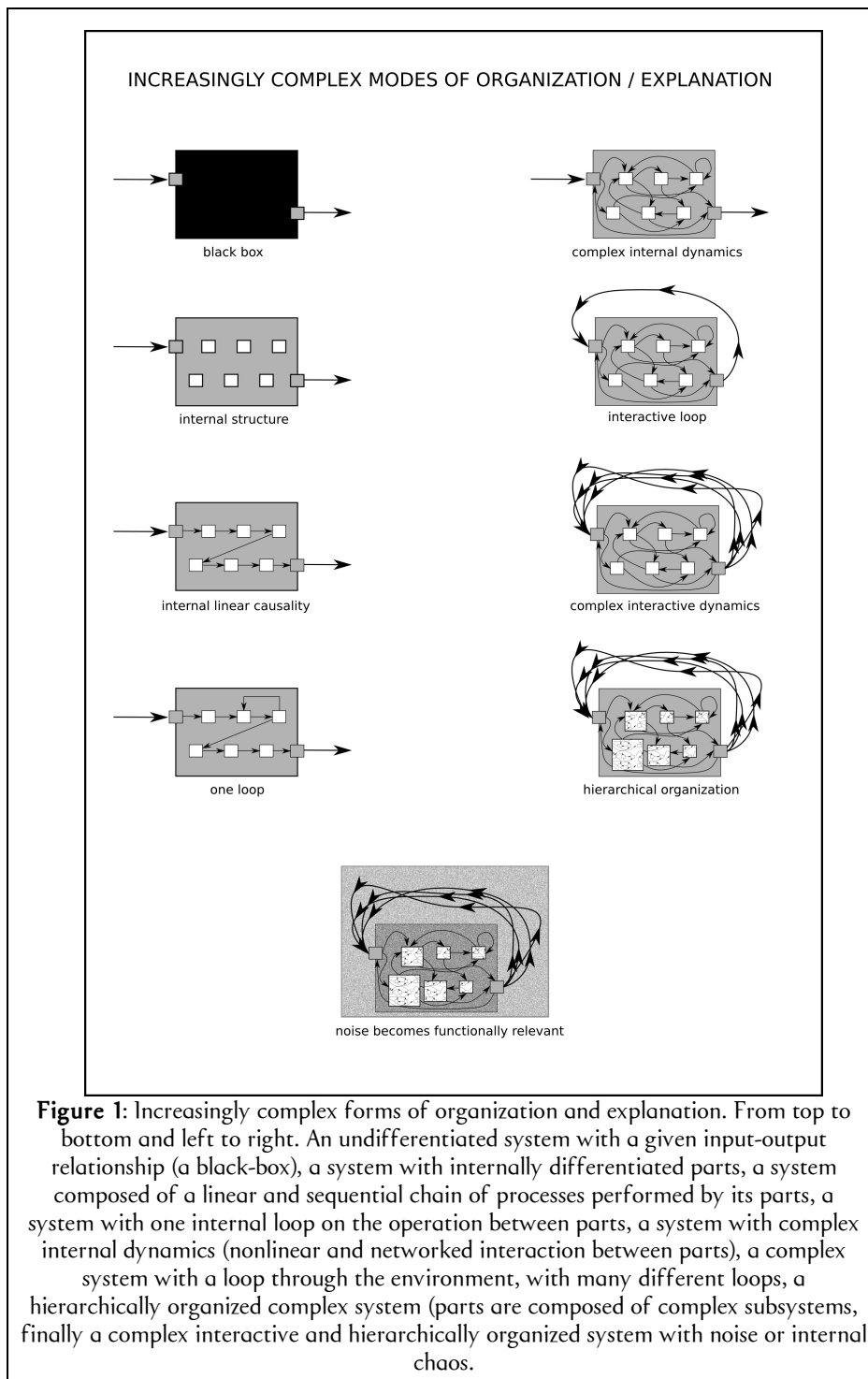
Things, however, might become much more complicated like, for instance, in a sand pile or a chemical solution. In such cases there is a huge number of components and attributing specific functional relevance to a single element (e.g. a molecular unit in a dilution) is impossible or inappropriate. Yet the system might not really be very complex if the contribution of the many parts to the properties or behaviour of the whole can be averaged out (what Weaver called "disorganized complexity"—1948). In a chemical solution, for instance, where millions of molecules might be moving around at different velocities, individual details might be disregarded to deduce the reactive property of the solution: mean statistical values such as concentration and temperature might be sufficient to fully specify the reaction function of the dilution. The same holds for the pressure or temperature of a gas or other macroscopic properties. In such cases there is a many-to-one mapping between structural components and functional properties so that, again, properties of the whole can be reduced to statistical properties of the parts (e.g. molecular shape) and some additional information coming from statistical averages (such as temperature and concentration).

In sum, what the strategy of decomposition and localization permits in atomistic reductionism is to isolate parts and to study their functional contri-

bution separately so that an adequate understanding of the system can be, or is expected to be, attained. When an unknown or complex process mediates the relationship between certain parts and their functional contribution in a regular manner one can ignore the mediating processes (under a *ceteribus paribus* clause) and opt for localizing the function to their correlated parts. When parts interact massively and stochastically one can ignore the interactions and provide statistical explanation of the behaviour of the whole. Unlike observational behaviourism mechanistic explanations permits to gain some knowledge of the internal component-processes that produce behaviour. Unlike functionalism, it permits to constraint the space of possible functional decompositions by localizing them in distinct structural components that constraint and enable such functions. This strategy of *divide and conquer* works for a certain type of systems. The division or decomposition might be carried out at the level of functions (top-down decomposition) or parts (bottom-up decomposition), as noted above, and successful mechanistic explanations generally involve a combination of top-down and bottom-up approaches. I should call these “top-down analytic” and “bottom-up analytic” approaches; where the term analytic denotes the decompositional aspect. When a system has a high number of parts and functions (when it is structurally complex), the process of achieving a satisfactory mechanistic explanation might be long and costly but might not require a change of methodological principles.

There are however two main ways in which this approach might get increasingly complex rendering atomistic or statistical reduction inappropriate or requiring a revision of the epistemological and ontological assumptions underlying the explanatory strategy. One mode of complexification occurs at the structural level, the other at the functional level. Structurally complex systems involve a high number of types of components arranged in such a way that no higher level simplified structural description is always available. This is what some measures of complexity (such as Kolmogorov complexity) are meant to capture: how much a string of symbols describing a structure can be compressed by substituting or transforming redundant substructures and relationships between them. For instance, a chromosome might have a complex structure in the Kolmogorov sense if the description of the sequence of amino-acids (the parts of the chromosome) cannot be compressed. However, this type of structural complexity might turn out to be trivially complex if for every part of the structure (e.g. every gene) there is a well identifiable function (e.g. a phenotypic trait)⁷. Thus, it might be the case that the system is composed of many parts performing their functions in a large sequence of processes. But,

⁷ We know this not to be the case for the genomes, which have been shown to operate within the cellular, organic and environmental context in complex regulatory networks with plenty of pleiotropic and redundant effects.



however complicated, such systems might not threaten the traditional atomistic reductionist methodology. The process of achieving a satisfactory explanation will turn out to be long and tedious but methodologically speaking what-the-explanation-consists-of remains the same: functional and structural decomposition and localization of parts-functions. In such cases structural complexity entails a very long and non-compressible explanation of each part and its assigned functions but no genuine methodological thread.

The true challenge of complexity arises when considering dynamical complexity at the functional level. This is what Weaver called “organized complexity” in which modelling a system requires “dealing simultaneously with a seizable *number of factors which are interrelated into an organic whole*” (Weaver 1948:539). I will distinguish conceptually four different aspects of dynamic complexity that threaten the traditional atomistic reductionism: internal emergent functionality, interactive emergent functionality, chaotic fluctuations and, finally, dynamical hierarchies and organismic totalities. Figure 1 captures a succession of levels of complexification in mechanistic systems.

2.1. Internal emergent functionality

One of the earliest and clearest formulations of a theory of emergence dates back to Broad (1925). He stated that a property of systemic relationships R is emergent if and only if it cannot be deduced from the most complete possible knowledge of the properties of its component parts taken in isolation or integrated in systems different to R :

Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents A , B , and C in a relation R to each other; that all wholes composed of constituents of the same kind as A , B , and C in relations of the same kind as R have certain characteristic properties; that A , B , and C are capable of occurring in other kinds of complex where the relation is not of the same kind as R ; and that the characteristic properties of the whole $R(A, B, C)$ cannot, even in theory, be deduced from the most complete knowledge of the properties of A , B , and C in isolation or in other wholes which are not of the form $R(A, B, C)$. (Broad 1925: Ch.2)

More recent debates around reductionism and emergence have shifted the focus from properties to functions and from relations to interactions. But Broad’s formulation remains, in general terms, still valid. What the atomistic-reductionist mechanicism assumes is that the system under study is *decomposable* or *nearly-decomposable* (to use Simon’s terminology—1962/1996): i.e. that interactions between component parts A , B , C are weak or minimal so that decomposition of the system in parts and their isolated study outside R does not seriously compromise their characteristic functioning when put them back together. Unfortunately, many (if not most) of natural systems (specially biological and cognitive ones) fail to satisfy this assumption: they are, to some extent, *integrated systems*. Integrated systems are those where parts strongly

interact with each other so that recurrent interactions between them become essential to achieve the overall system behaviour (Simon 1962/1996). In such cases functional decomposition is not always possible or, more precisely, certain relevant properties and functionalities of the system cannot neatly be reduced to an aggregation of its component parts. In other words, the mapping between functions and structures is not a one-to-one mapping, nor a statistical many-to-one but a one-to-many mapping (a component takes part in different functions) or a many-to-many mapping (different parts participate in multiple functions with different configurations) due to the recursive, non-linear and redundant nature of the interactions between parts. Parts cannot thus be linearly aggregated to understand the functioning of the system as a whole: “the whole is more than the sum of the parts” as runs one slogan of the sciences of complexity.

The Cartesian mechanicism undervalued this aspect of complex mechanisms. The third step of his method required to assume a sequential order, even if a fictitious one, among the components of a system:

The third was to carry on my reflections in due order, commencing with objects that were the most simple and easy to understand, in order to rise little by little, or by degrees, to knowledge of the most complex, *assuming an order, even if a fictitious one, among those which do not follow a natural sequence relative to one another.* (Descartes 1637/1996: 13, italics added)

But when components interact recursively, the “natural” order and relationship between parts is crucial to understand the system’s behaviour. In addition, the sequential order cannot capture the relevant organizational architecture of a complex system.

Consider first the case of a *single loop* in contrast with the sequential relationship between the operations of the parts considered above. If the effect of the operations of component A feeds back to itself through another component B the functional/dynamical contribution of component A cannot be isolated from component B; i.e. functioning of A cannot be isolated and defined as a set of definite transformations between well defined inputs and outputs since the output of the system will in turn (through B) become an input and affect its functioning in a recursive manner. Not only is a mistake to assume an order “even if a fictitious one, among those which do not follow a natural sequence relative to one another” (e.g. analysing the operations of A after those of B) but the very conception of the system as a sequence of operations leads to a limited conception of systemic functioning. The functional organization of systems with many feed-back loops (i.e. integrated systems) is that of a *non-linear network of interactions* as opposed to the *linear sequence of operation* of aggregate systems. In the most simple of such systems overall functionality is achieved through a big number of relatively simple but networked components through their recurrent parallel interactions. The strategy of decomposition and localization cannot succeed to study emergent functionality

The early cyberneticians soon became aware of this problem of the traditional method:

Science stands today on something of a divide. For two centuries it has been exploring systems that are either intrinsically simple or that are capable of being analysed into simple components. The fact that such a dogma as *vary the factors one at a time* [which may be considered equivalent to a parts-function decomposition] could be accepted for a century, shows that scientists were largely concerned in investigating such systems as allowed this method; for this method is often fundamentally impossible in complex systems. (Ashby 1956: 5)⁸

A paradigmatic example of the problems of decomposition and localization in complex systems is provided by self-organized processes whose occurrence spreads over physical and chemical to biological, cognitive and social systems. In physics the self-organized properties of collective patterns found in far-from-equilibrium systems like lasers, spin glasses or Benard cells have been recognized and well studied for long. More familiar examples of self-organization in physics are tornadoes, swirls or candle flames. In chemistry, Turing's pioneering theoretical work on reaction-diffusion patterns (Turing 1952) and Illya Prigogine's work on dissipative structures (Nicolis & Prigogine 1977) were the most sound approaches to self-organization. In biology Waddington's notion of epigenetic landscapes (Waddington 1957) stands as a major reference that addressed how embryonic developmental processes self-organizes while genes shape developmental pathways acting like order parameters. In cognitive science Turing also advanced some distributed models in the early 50s (Copeland & Proudfoot 1999), Ashby (1952) draw one of the first comprehensive pictures of the brain as a homeostatic dynamical system, while Rosenblatt (1958) created the *Perceptron* (a neural network capable of classifying input patterns with a distributed architecture) and Hopfield (1982) developed one of the most important notions of self-organization in recurrent neural networks inspired on the physical approach to the self-organized processes of spin-glasses.

What all these systems have in common is that a global pattern or behaviour emerges out of the local and distributed (often stochastic) interactions between components. Yet, this components do not perform any well defined component sub-operation that could be said to be decomposed from the overall behaviour, nor can they be averaged out so that the global pattern be considered just a statistical property of the lower level components. In the "simple" cases every part performs a very simple and uniform operation but its contribution to the overall pattern changes contextually and rapidly in relation to the global or higher level pattern. It is then said that function emerges from a distributed and networked structure: "Roughly, the notion of emergence refers to the fact that the system's global behavior is not only com-

8 Quoted in Wimsatt (1974).

plex but arises from the collective actions of the simple components, and that the mapping from individual actions to collective behavior is non-trivial.” (Mitchell 2006: 1195).

This type of systems are generally conceptualized at two different levels: the micro and macro levels. The macro level patterns is often said to constraint the activity of the micro level. Proponents of *strong emergence* (Campbell 1974, Moreno, A. & Umerez, J., 2000) defend that the macro level presents a causal power of its own (downward causation) whereas proponents of weak emergence (Bedau 1997) conceive the macro level as nothing more than a global collective perspective on the system. Discussion around emergence has mostly focused on issues of reduction of the emergent pattern to lower level interactions. What remains important for our discussion is that the whole systemic context in which parts are situated needs to be taken into account in order to provide an explanation of the system’s behaviour.

Thus, in many forms of self-organization at physical, chemical, biological and cognitive levels, long-range correlations and stability lead to global functional patterns. And these functional patterns cannot be reduced to local interactions between components nor to an organizing subsystem whose regulatory or instructive action could be said to act as a program or source of functional structuring. In other words: the relationship between structural elements and behaviour cannot be captured under the form of a map or a design chart (e.g. like the design of an engine) but it rather takes the form of a dynamic generator. When translated into a graph the representation of the causal organization of the system takes the form of a landscape (a state space) where every part (or variable) is a dimension of space, not a point or a decomposable “part” of that landscape. The map is not a map of parts but of dynamic relationships between all the parts.

2.2. Interactive emergence

Another, not less important aspect, of complexity comes from the *interactive emergence* of some functional properties of systems. Feed-back loops may not only be introduced between components of the system but also between the system and its environment so that some characteristic behaviour of the system may be impossible to study by isolating the system from its environment (e.g. in niche construction, adaptive behaviour or developmental processes). This feature of strongly interacting systems renders direct and unique causal attribution of observed behaviour to internal components problematic. Think for instance on the task of reaching a cup of coffee. The perception of the cup and the desire to reach it initiates the movement of you arm. Yet, far from having a coordinate mapping module and a trajectory calculator module that

outputs a definite motor command to your arm, your movement is continuously adapting to the optimal trajectory through a continuous process of visual and proprioceptive feedback. Your behaviour cannot be decomposed into isolated detection, planning and execution subcomponent functions but needs to be recast through recurrent environmental interactions that need to be integrated in order to explain the observed behaviour. The same holds for a bacteria climbing up a sugar gradient or the roots of a tree searching for a water source. Grey-Walter's robotic tortoise (1950) and latter work by Valentino Braitenberg (1984) demonstrated how very simple mechanisms could lead, when coupled to certain environments, to complex and intricate behaviours that results from a continuous feedback between agent and environment (including other agents).

In addition, both types of emergence (the interactive and the internal) may appear nested. Not only can the internal organization of the system be distributed among strongly interacting parts or the causal structure of its behaviour can be distributed through the environmental feedbacks, but internal component parts may also get intertwined through the environment. First proposed by Graccé (1959) *stigmergic* collective niche-construction (in ants, termites and bees) is a paradigmatic example: individuals (part of the colony-system) get local feedback from the action of other individuals on the environment so that each part of the system (each individual) is coordinated with other parts *through* the environment (Bonabeau 1999). In locomotion, Randall Beer (1990) has explored how feedback through the environment helps different legs in an hexapod to be coordinated without centralized internal control exploiting their physical embodiment. Brooks (1991) and Steels (1991) provided nice examples of how a robot could perform behavioural functions (such as wall following) without appeal to an explicitly representational functionalism. This distributed causal spread has come to challenge functionalist and mechanistic modes of explanation particularly within the framework of representational cognitive science (Steels 1991, Clark 1996, Wheeler 2005).

2.3. Chaos and fluctuations

In addition to the interactive and internal causal spread that leads to functional emergence, the dynamics of complex systems often show chaotic properties; introducing severe transformations on the epistemic relationship between such systems and their models. In short, *chaotic behaviour* is defined by the characteristic that small variations on the initial conditions give rise to huge changes as the behaviour of the system unfolds. The consequence is that even if we had a perfect model of the system (with a detailed account of the relationships and operations of the parts) the measuring of the state of the target system does not guarantee a successful prediction of its evolution since

microscopic inaccuracies on measurement will lead to enormous divergence on prediction. The stereotypical example is that of weather modelling and the popularized idea that the flap of a Butterfly's wing in Brazil can be responsible for a tornado in Texas⁹. One of the most celebrated references goes back to Lorenz's work on chaotic attractors (Lorenz 1963) where a completely deterministic system was shown to exhibit unpredictable non-periodic behaviour (yet the system could behave in characteristic regions and drawing stereotyped shapes). The rise of computersystems able to calculate or to systematically reproduce the behaviour of chaotic systems, lead to an explosion of their simulation and qualitative study since 60s to present day. But the issue was not new. The very foundations of celestial mechanics faced the problem for the first time. It was probably Poincaré in 1890 the first to note that the solution to the model of the behaviour of certain systems (in this case the three body problem) lead to "a sort of network, of web, of infinitely intertwined tissue" whose complexity was overwhelming but still tractable to some degree.

However, biological and cognitive systems need to satisfy adaptive constraints and their functioning cannot be purely chaotic; exponentially divergent evolution of behaviour would inevitably lead to an incapacity to respond adequately to different situations and to maintain a stable and robust performance. Despite their chaotic properties, the behaviour of biological and cognitive systems shows higher order regularities; functional stability is maintained over environmental and internal perturbations together with the capacity for flexible and plastic adaptability to new circumstances.

In turn, internal and environmental fluctuations or "perturbations" are not an unpleasant and negligible noise to get rid off but can become essential for the functional organization of a system. Another slogan of the sciences of complexity applies here: "order out of chaos"; meaning that structured patterns emerge out of underlying chaotic dynamics. Adaptive flexibility requires that structurally limited systems quickly reconfigure their behaviour to satisfy changing demands. This is why complex systems are said to lie at the "edge of chaos" (Langton 1990) or in self-organized criticality (Bak 1997) where complex patterns can appear and disappear to give rise to new ones making the system capable of satisfying the demands for a balance between stability and flexibility. Chaotic fluctuations have been proven to be crucial to produce some relevant functional properties in both chemical (Nicolis & Prigogine 1977) and neurodynamic (see Freeman et al. 2001) regimes: "[the] self-sustaining, randomized, steady state background activity is the source from which ordered states of macroscopic neural activity emerge" (Freeman et al. 2001:109). This significant contribution of chaos (often conceptualized as

9 This, nowadays famous, phrase is attributed to Lorenz who gave a talk, at the 1972 meeting of the American Association for the Advancement of Science in Washington, entitled: "Predictability: Does the Flap of a Butterfly's Wings in Brazil set off a Tornado in Texas?" (Hilborn 1994).

randomness or noise) for functionality can hardly be integrated in traditional mechanistic reductionist accounts. If randomness cannot be something to be ignored or averaged out (in order to provide a clear explanatory picture of the mechanisms) but provides lower level fluctuations and variations that are “re-cruited” by the system to generate changing global configurations, then, the picture of a phenomenon that is the result of well identifiable parts subject to functional analysis gets severely compromised. For many (Mitchell 2006) this chaotic or random fluctuation of some lower level elements (such as random search in ants) is a crucial feature of all complex phenomena. At the most abstract level, evolution itself can be considered an instance of this type of “mechanism” of blind stochastic variation and selective retention. And this very same generic mechanism has been argued to operate, not only in biological phylogenetic evolution, but in the immune system (Burnet 1956), in neural development (Edelman 1987) or even in all types of cognitive or epistemic process (Campbell 1974).

2.4. Hierarchies, levels, mechanisms and organization

The sciences of complexity have mostly focused on what Evelyn Fox Keller has named “one shot-order-for-free” types of reactively simple or poorly organized systems (Keller 2007). But complex biological and cognitive systems show a considerable degree of hierarchical organization (atoms-molecules-organelles-cells-tissues-organs-systems) and modularized structures. Simon (1962/1996) defended that hierarchical and modular organization is a principle of complex systems. On the one hand, hierarchical levels will form almost spontaneously due to difference in spatio-temporal scale and inter-component cohesive forces (stronger within atoms than between molecules). On the other hand, he argued, modularized structures would provide the means for easier evolvability since semi-autonomous parts (modules) could be combined to form aggregates of different types and number of components. Among other arguments, the modularity and the hierarchical organization of what he considered complex systems made Simon and others believe that the aggregativity assumption of the classical mechanistic and functional analytic decomposition could be maintained. Along this line, mechanistic explanations can be viewed as affording for increasingly fine grained details: mechanisms are made of functionally specific parts whose behaviour can be analysed as a mechanism, which in turn is composed of parts and so on in a hierarchy of explanatory levels that mirrors the architecture of hierarchical organization of complex systems.

Things, however, may not turn out to be so simple specially if we take into account that living systems are homeostatic systems capable of self-construction and repair and also show a high degree of flexibility and plasticity. It turns out that complex systems show a nested hierarchy of spatial and tem-

poral scales of self-organized processes in which a variety of structure-function relationships may appear. At many levels and between certain time-windows, traditional mechanistic assumptions will hold and it is within these contexts that most of the mechanistic and explanatory advances in biological and cognitive science has taken place till recently (Moreno, Ruiz-Mirazo & Barandiaran 2008). But this early success on disclosing some relevant mechanisms does not imply an unlimited success of the traditional method whose limits (even among those that have endorsed and successfully applied it) are being increasingly made apparent (Klipp et al. 2005) .

A crucial aspect that escapes traditional mechanistic explanations is the complex relationship that is established between levels and scales that appear nested in the same phenomenon. For instance, any living creature is the result of the interplay between at least two major temporal scales: the evolutionary and the developmental. How developmental constraints affect evolutionary dynamics is difficult to capture by traditional explanatory paradigms that show difficulties on integrating mechanisms operating at different temporal scales into the same model. A similar situation arises when considering individual and collective scales of functioning. But the relationship between local and global patterns that the concept of emergence was meant to capture undergoes an additional complexification when nested hierarchies of emergent levels appear. And even the most simple of cognitive behaviours (e.g. habituation) will include causal contributions from a wide range of levels: genetic, metabolic, neural, musculo-skeletal and environmental, etc. In Polanyi's words:

Each level relies for its operations on all the levels below it. Each reduces the scope of the one immediately below it by imposing on it a boundary that harnesses it to the service of the next-higher level, and this control is transmitted stage by stage, down to the basic inanimate level. (...) When examining any higher level, we must remain subsidiarily aware of its grounds in lower levels and, turning our attention to the latter, we must continue to see them as bearing on the levels above them. (Polanyi 1968: 1311—1312).

A mechanistic explanation need often cut across different levels and big part of the explanatory work involves precisely to make explicit how causal interactions do occur between levels (e.g. behavioural, inter-cellular and genetic mechanisms involved in learning). Under extremely simplified conditions (e.g. a habituation of a withdrawal reflex in an anaesthetized *Aplysia*) some reductionist assumptions may hold and one can neglect the threats of self-organization to classical mechanistic account (e.g. the cascade of internal biochemical transformations inside the neural cell can be simplified at a higher level of description and the underlying complexity ignored—Kandel 2001). But more sophisticated cases may well involve taking seriously into account self-organizing processes occurring at different scales and having crucial causal consequences across levels.

In particular, biological and cognitive systems present second (or higher) orders of holistic organization made possible by regulatory control (Moreno, Ruiz-Mirazo & Barandiaran 2008) and large scale distributed coordination. Regulatory systems (e.g. genetic, neural or endocrine systems) operate modulating locally self-organized or modular processes to generate coherent functional processes. Typically, a living system presents different hierarchies of regulatory control and yet, unlike classical control systems, the regulated processes maintain and repair the regulatory subsystem introducing an inter-level dependency that is difficult to unpack when holding traditional decompositional assumptions. Large scale distributed synchronization is also ubiquitous in complex living mechanisms (e.g. cell signalling networks, immune systems and large-scale neural dynamics) yielding higher orders of functional integration among different subsystems.

In addition, biological and cognitive systems are not only functional but intrinsically *normatively* functional systems: i.e., not only they function in a certain (more or less complex way) but they must function on that way in order to assure their continuing existence. The key idea here is that the structure of living systems is not (on its stability and functioning) independent of its functioning but, on the contrary, that the functional level feeds-back (so to speak) to the maintenance of the structural level. There is thus a circular co-dependence between the stability or self-maintenance of structures and their functions. The concept of a mechanism assumes that the system is not broken, it assumes a *ceteris paribus* clause about the context and the stability or invariance of the components and their relationships. Interestingly, many biological and cognitive phenomena are characterized by self-repairing, adaptive and homeostatic regimes so that the stability of the structures and functions is recursively maintained. Even more, the overall “function” of the system is to maintain itself. It is the fundamental structure-function dichotomy that is at risk here and the concept of a mechanism, even if expanded to accommodate emergent functionality, may fall short to capture this type of complexity. I will, instead, use the term organization to refer to it. Thus, the term organization will be used to refer to holistic mechanisms that maintain invariant or stable the structures and relationships that define a system (Maturana & Varela 1973/1980). Whereas mechanisms describe and explain the causal structure at a particular instant, the organization describes a more encompassing context in which mechanistic models may hold.

An adequate explanation of certain phenomena (e.g. embryonic development) will ultimately require the merging together of multiple mechanisms operating at different spatial and temporal scales into a coherent ensemble that is able to integrate local processes of self-organization, recurrent environmental interactions, random fluctuations, and also causal processes subject to the traditional atomistic mechanicism. The challenge today is to generate hol-

istic models that are able to provide what may be called a many-to-many-to-all relationship between structures and functions within an organic self-sustaining and adaptive totality (Ruiz-Mirazo, Moreno & Barandiaran 2008). In a similar vein, mechanistic explanations in psychology will ultimately require to put them in the context of increasingly higher levels of organization of behaviour expanding across different spatial and temporal scales and different degrees of inclusion of the bodily and social environments. After reviewing the evolution of mechanistic explanation of motivation, Wright and Bechtel conclude:

[A] full explanation of motivation itself—especially beyond that of immediate attributions of incentive salience—must eventually involve models of mechanistic systems governing the production of planning and decision-making, the regulation of emotion and long-term memory, creativity, social role formation, and so forth (...). In sum, explaining motivation mechanistically requires illuminating the organizational collusion and interaction of these various composite systems that engage their environment at increasingly higher levels. (Wright & Bechtel 2006: 38, on-line version)

The question remains, however, over how should this integration of mechanisms into an increasingly complex organizational context be made. What are the available modelling tools, once the reductionist-localizationist program is shown severely limited on its capacity to account for emergent distributed processes, to integrate various levels and scales and to model the resulting organizations?

3. MODELLING COMPLEXITY: A SYNTHETIC BOTTOM-UP APPROACH¹⁰

We have seen how many of the target systems that biological and cognitive sciences deal with are integrated systems, endowed with a non-linear networked structure with chaotic properties and where functionality emerges internally and interactively. The complexity of the natural phenomena does not, however, rule out the possibility for mechanistic explanations. It just transforms some of the early assumptions of localizationist versions of mechanism and its atomistic reductionist consequences. I will liberally borrow from Glennan a more encompassing definition of mechanisms that may be put at work to model complex systems:

A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations. (Glennan 2005: 445).

¹⁰ This section is mostly based on previous work (Barandiaran and Moreno 2006a, Barandiaran and Moreno 2008).

Invariant, change-relating generalizations will appear stabilized at certain levels in biological and cognitive systems and permit to model a system whose local dynamic and causal structure is often far away from more basic physical or chemical laws or principles. Following Glennan, in this context, a mechanical model “consists of (i) a description of the mechanism’s behavior (the behavioral description); and (ii) a description of the mechanism that accounts for that behavior (the mechanical description)” (Glennan 2005: 446). Although Glennan remains unclear about the status of models in which functional localization fails (stating that in such systems only functional decomposition is mechanistically relevant) the sciences of complexity have provided tools by which the (network) structure of such systems can be systematically related to their global functionality, providing means for their quantitative and qualitative modelling (see Strogatz 2001 for a recent review). In turn, a mere description of the lower level will not do. When the traditional assumptions hold, a description of the mechanisms involves the narrative of decomposition and localization which amounts to an explanation. Complex systems however require a simulation (a dynamic reproduction of the phenomenon) that cannot be reduced to localizationist mapping and its description.

When parts are functionally integrated, the modelling process involves the description of the infinitesimal change of a component in relation to the change of the others. Observable parts or components are represented as variables and the “change-relating generalizations” or interactions between parts are represented as functions governing the rate of change of variables (Ashby 1956). These models usually take the form of a set of differential or difference equations. Interestingly, complex systems are not analytically tractable, not just in the sense of analysis as decomposition as we have seen, but in the strict mathematical sense. One cannot solve the equations to determine the state of a variable at a given time. Numerical methods are used instead to draw the state space of the system. Numerical methods are also called qualitative (as opposed to the quantitative nature of analytically tractable systems) consisting on a fine grained step-by-step update and record of the state of all the interrelated variables of the system. Thus, the evolution of the system needs to be *reproduced* on the simulation, rather than formally deduced for a given time value or decomposed in functionally specific subsystems. Since sensitivity to initial conditions might be critical (due to potentially chaotic properties of the system), the simulation process is repeated for a wide range of initial conditions and the behaviour of the system recorded. A full record of different simulations allows to draw the state space of the system (given certain boundary conditions and specific parameters) so that regular or stable patterns of behaviour can be observed. Different network configuration parameters, the rules governing components or boundary conditions, can be systematically studied on the simulation so that a deep understanding of the

structure-function mapping can be gained without a strict decomposition and a one-to-one localization.

Complex mechanisms can thus be modelled using simulation models in which the specification of local invariant interactions, global constraints and conditions, leads to the unfolding of global emergent behaviour. This type of complex-mechanistic approach takes the form of *generative models* that produce the behaviour under study without implying that a one-to-one mapping may be possible between structures and functions. Nevertheless, it is assumed that nothing more than interactions between parts is necessary in order to provide a satisfactory explanation of a given phenomenon. This amounts to a *synthetic bottom-up* methodology since knowledge of the mechanisms that produce global behaviour is achieved by synthesis (putting simulated component processes interacting together) and reproduced from the bottom-up: i.e. functionality is not imposed or attributed from a more abstract descriptive level but emerges from the interaction of the components of the simulation. This methodology is characteristic of Artificial Life (Langton 1989, Boden 1996) and, more generally, what Simon (1969/1996) called “The sciences of the artificial”. What distinguishes Artificial Life models from other sciences of the artificial (particularly from traditional AI) is the emphasis on emergent functionality so that the lower level implementation of interactions between parts is rarely imposed or constrained from the top-down (component x is created to implement pre-specified function f) but emerges bottom-up (components x_n are created and their iterated and networked interactions give rise to function f). Thus, in typical Artificial Life simulation models, component parts x_i only become functionally significant in the context of recurrently interacting parts x_n and/or the interaction that all components together, or different subsets, establish with its (simulated or physical) environment.

Due to its defiance to functional decomposition this kind of models are difficult (when not impossible) to design by hand and their creation is made possible by additional modelling tools. These tools generally involve the automatic optimization of parameters, components, and local rules to achieve a given global functionality. Genetic algorithms implementing artificial evolution (Holland 1992, Mitchell 1996), simulated annealing (Kirkpatrick 1983) or different distributed algorithms based on local “learning” rules (e.g. Hopfield 1982) are among the most popular tools to achieve a given behavioural outcome from an integrated system.

The main interest and methodological novelty of Artificial Life models lies in their capacity to develop an experimental research program in the computational domain. The epistemic status of this computational domain does not easily fit within traditional concepts of philosophy of science. An adequate interpretation of the functional meaning of the different processes and patterns, obtained when running the model, becomes much more problematic than

what traditional epistemology of scientific methodology has established. Artificial Life simulations acquire a hybrid status; they are, as Sismondo has pointed out, at the same time tools, representations, objects and ideas:

These models and simulations easily cross categories, such as “theory” and “experiment”, the bounds of which are otherwise well established. And modelling and simulation sit uncomfortably in science both socially and epistemically, because the boundaries they cross (Sismondo 1999: 247).

In addition, even the traditional relationship between more abstract theoretical statements (laws and principles) and models (specific instances of such principles under different combinations, constraints and conditions) is itself complex and opaque:

A third point, one that is characteristic of the trial and error methods often involved in solution procedures, is that a clear and transparent understanding of the deductive relations between fundamental theory and specific application is not necessary in order for a good fit to ultimately be achieved between theory and data. To put it in a characteristically philosophical way, there has been a long tradition in epistemology that every step in a derivation from fundamental principles should be individually open to inspection to ensure correctness. What is characteristic of many solution methods is that such transparency is not available and that the virtues of a model, such as stability under perturbations of boundary conditions, scale invariance, and conformity to analytic solutions where available can be achieved by trial and error procedures treating the connections between the computational model and its solutions as an opaque process that has to be run on a real machine for the solution to emerge. (Humphreys 2002: S8—S9)

There are no standardized methods to achieve fruitful explanations from Artificial Life models. This condition (the lack of a standardised procedure) requires that, together with the model, researchers carefully design and specify how it shall be used in explanatory terms, making explicit the interpretative framework that makes the model an explanatory tool. Four different classes of such frameworks can be distinguished, according to the level of abstraction of the model or, in other words, according to the position models are made to occupy between theories and empirical data: mechanistic-empirical models, functional models, generic models and conceptual models. This classification in four levels is certainly not exhaustive, but reflects somehow a clustering of existing models and their underlying epistemic logic.

A valuable criteria for classification of models is to think on the evaluative framework to which a model belongs: how and what are they meant to represent. I distinguish those models whose object of modelling is theoretical (whose evaluation is done against theoretical principles or formalisms), and those whose object is empirical (which are evaluated against empirical data coming from specific natural phenomena). In the first class I distinguish, on the one hand, those generic systems that serve to discover or classify generic properties of complex systems; I call these *generic models*. On the other side

of the theoretical realm, we find what I call *conceptual models*, whose explanatory framework is given by the dissonance or adequacy with the relationships established among concepts of a given theory. Abstract conceptual models are used to formalize or compare definitions of generic concepts (such as emergence, complexity or hierarchy) while domain specific conceptual models are used to explore the role and interaction between more specific concepts (such as learning, plasticity, evolvability, etc.). The third and fourth classes of epistemic models directly involve their matching or interaction with empirical data. On the one hand, *functional models* are defined as those that must adjust to the particular behaviour or functionality exhibited by certain natural system. *Mechanistic models*, on the other hand, are those which act as functional *and* structural models of a particular phenomena. The model is meant to be realistic: variables of the computational system represent observables of the empirical system, including the internal mechanisms producing the phenomena under study.

Note that the same artificial construct can be used at these four different levels of abstraction, so that it is not the simulation itself what specifies its level of abstraction, but the use and interpretation of it. For instance CTRNNs (Continuous Time Recurrent Neural Networks) can be used at different levels: to explore abstract systemic properties of dynamic systems at an abstract generic level, to explore the concepts of learning and memory at a conceptual level, to achieve a desired functionality found in ant behaviour at a functional level, or to model the pattern generator circuit of Aplysia at a more realistic mechanistic-empirical level.

3.1. Mechanistic-empirical models

In complete mechanistic models there is a correspondence between the variables in the model and a set of observables of the modelled natural system (synaptic connections, metabolic pathways, number of genes, etc.). The model is meant to be empirically realistic, at least to a particular level of mechanistic accuracy. As Glennan puts it:

Whether a state space model is a mechanical model depends upon what state variables are chosen, and whether the laws of succession used to characterize the state changes represent direct causal interactions between the parts of the mechanism. (Glennan 2005: 448)

This mechanistic correspondence with the modelled object is exploited to disclose its underlying complexity and extract intermediate explanatory patterns that are hardly discernible under limited experimental conditions and access to empirical data. The following quote illustrates the way in which scientists make use of simulation models as surrogates of the target system:

The simulated bacterium thus pursues a biased random walk over the computer screen in a manner responsive to the local concentrations of aspartate. At any in-

stant of time, one can read out the concentrations of extracellular aspartate and intracellular signalling molecules in an individual bacterium as well as other parameters such as the rotational bias of the flagellar motor. The same approach can be applied to populations of bacteria, revealing their changing distribution with time. Evidently, these simulations have many obvious differences to real bacteria, as we discuss below. Nevertheless, the virtual bacteria reproduce a broad range of experimental data relating to the sensitivity and rates of response and adaptation, for both wild-type cells and chemotactic mutants. Because their internal biochemistry is firmly based on decades of experimental data, we believe it is legitimate to treat these representations as experimental objects in their own right. The advantage is that we can expose cells of any specified genotype to precisely defined stable gradients of any required shape (including those that are difficult, or impossible, to achieve in the real world) and observe their behavioral responses. (Bray et. al. 2007: 13)

A parallel project in neuroscience is the *Blue Brain Project*¹¹ supported by IBM. With up to 35.000 processors is capable of simulating with a molecular level of detail more than 10.000 neurons and 30.000.000 synapse models of the cortical columns of a mouse brain performing 18.7 trillion calculations per second (Markram 2006). The project website states: “We have achieved biological fidelity such that the model itself now serves as a primary tool for evaluating the consistency and relevance of neurobiological data, while providing guidance for new experimental efforts.” (Blue Brain Project¹²).

These types of models are thus used to manipulate the systemic variables and relationships in ways not accessible to the direct manipulation of the modelled system and, particularly, to make intensive systematic and automatized manipulations. They are becoming increasingly widespread in different areas of contemporary biological research, particularly in the new emerging field of System's Biology (Kitano 2001, Klipp et al. 2005, Wolkenhauer 2006). As Klipp puts it:

“Recently, researchers working in different fields of biology have expressed the need for systematic approaches. They have frequently demanded the establishment of computer models of biochemical and signalling networks in order to arrive at testable quantitative predictions despite the complexity of these networks” (Klipp, et al. 2005:v)

Since the simulation is able to integrate many different contingencies and parallel mechanisms that altogether contribute to the production of the phenomena under study, it permits to discover which combination of variables and parameter values are crucial to achieve a particular behaviour or functionality. Some of the causally relevant variables, however, will not belong to the set of observables but will emerge as collective variables or higher order patterns that are a combination of local and distributed parts. It is particularly in rela-

11 <http://bluebrain.epfl.ch/> [last accessed 28 March 2008]

12 <http://bluebrain.epfl.ch/Jahia/site/bluebrain/op/edit/pid/19094> [last accessed 28 March 2008]

tion to these intermediate emergent patterns that the simulation becomes indispensable as a tool for discovery and explanatory grounding.

At times, the very model represents the most complete understanding of a given natural system we can nowadays hope to achieve: maximally simplified (but still complex), manipulable and predictive artefacts that unfold the causal structure of a particular living phenomenon. This is probably why Morowitz has claimed that “computers are to biology what mathematics is to physics”¹³. Examples of realistic and mechanistic models can be found at Webb and colleagues’ detailed model of a cricket’s phonotaxis (Horchler et al. 2004) or Bray et al.’s (2007) model of bacterial chemotaxis mentioned above. Due to the complexity of the target system the very simulation model becomes an object of exploration and a valuable source of data, to the extent that the simulation itself has occasionally been used to correct empirical data (Abouhamad et al. 1998).

3.2. Functional models

These models are evaluated against a particular behaviour or functionality exhibited by their target systems when the underlying mechanisms are unknown, controversial or incompletely understood. Thus, whereas functional models need not or do not fit with underlying mechanisms (at least at the level of direct observational adequacy of modelled parts) they include higher level constraints which are specific of the phenomena under investigation. Functional simulation models are part of a wider mechanistic explanatory project and, precisely due to their lack of mechanistic correspondence, these models can be used to explore candidate mechanisms that produce or contribute to the observed and simulated global pattern or behaviour. When mechanisms are complex the simulation permits to assess the performance of existing hypothesized candidate mechanisms (whose complexity eschews a logical or diagrammatic deduction of consequences) or to assess which environmental factors participate in the interactively emergent causal structure of behaviour, etc. (e.g. landmark navigations).

Particularly interesting instances of functional models are given by the use of CTRNNs (and its variants) as universal smooth dynamical system approximators (independently of any attributed resemblance with real neuronal architectures). Artificial evolution is applied to a CTRNN control system to achieve a particular embodied behaviour on a simulated robot (Beer 2003). Using this technique, Vickerstaff and Di Paolo (2005) artificially evolved path integration behaviour to match homing behavioural data from *Cataglyphis fortis* ants, making use of a CTRNN as a dynamical generative mechanism that only vaguely resembles real neurons. Path integration is a form of finding the

¹³ Quote borrowed from Hunter (1993).

return home path using only a compass and an odometer to monitor the path travelled. Proposed models of path integration (PI) in *Cataglyphis fortis* lacked mechanistic accuracy and were developed under different assumptions regarding how the Home Vector (HV), that integrates information about distance and angle to home position, was encoded on ants' brains. Acknowledging the complexity of internal and interactive emergent functionality, Vickerstaff and Di Paolo attempted a different approach:

We approach the problem of the neuronal mechanisms of PI [Path Integration] here without making any apriori assumptions about the way the HV [Home Vector] is encoded but rather use an automated network generation technique, the GA [Genetic Algorithm], to explore a set of possible networks that enable an agent to home using PI after an excursion. This procedure permits a less-biased exploration of possible neural mechanisms and also has other advantageous features for system design. We must fully and explicitly specify the sensory and locomotory capabilities of the agent, the information-processing capabilities of individual neurons and synapses, the sources of noise present and the behaviours we wish the agent to display before we can use the GA. The simulation code must explicitly generate the behaviour of each candidate PI model and agent in order to evaluate it. This reduces the chance of producing a flawed or fragile model. The GA is also able to prune the evolving networks to favour simplicity of neural structures. Finally, we can readily have the GA produce models under variations of the assumptions, such as the class of network model or the nature of the animal's internal compass representation. (Vickerstaff and Di Paolo 2005: 3352)

Vickerstaff and Di Paolo's models are considerably rich in sophistication and detail to be treated extensively here. It shall suffice, however, to highlight a number of relevant modelling issues raised on their approach. First, their evolved control system reproduced previously observed errors without the system having been optimized to match the error data (i.e. without having to introduce ad-hoc artefacts to fit empirical adequacy—as some alternative models had done before). The resulting control architecture was further simplified and analysed to understand the role of relevant intermediate dynamic patterns or functions (like leaky integration and cosine-shaped compass response) and used to disambiguate between previously proposed alternative theories of path integration. In addition, path integration was generally modelled as a long distance integration mechanism, under the observation that the real strategy of the ants involved an additional capacity to navigate the local vicinity of their home with a different behavioural strategy (usually an 8 shaped walk around the estimated position of the home). Interestingly, Vickerstaff and Di Paolo's model required no additional mechanisms to carry on local search in the vicinity of the home. This last finding suggested the hypothesis that the very same mechanisms that operate to generate homing behaviour through path integration was able to generate local search (something previously unexpected and assumed to be done by a different and functionally specialized mechanism). This result uncovers a characteristic feature of a

complexity approach: when functionality is left to emerge from the bottom up, previous assumptions regarding functional decomposition are often discovered false or, at least, not necessary

A key feature of functional models is that dynamic causal structure of the behaviour generating system (achieved through automated optimization techniques) can then be analysed independently of the specific lower level mechanism that could support it. Thus, halfway between empirical and abstract theoretical modelling, this research strategy leads to a kind of emergent functionalism in which it is the dynamic organization of behaviour what captures the essential features of natural phenomena (unlike computational or adaptationist functionalism).

3.3. Generic Models

At the most abstract level we find certain computational constructions or *template* (Humphreys 2002) with no particular reference to any specific object of study, but whose formal structure has been selected in virtue of their resemblance with a wide range of natural phenomena. Their generality is very high and exploration of these artificial systems leads to the discovery of generic abstract properties of complex systems. Such is the case of explorations in Random Boolean Networks, Cellular Automata (CA), Networks Graphs or architectures, some research in Neural Networks, Coupled Oscillators or Dynamical Systems Theory more generally. Often generic models originate on empirical or functional models when a particular structure is found to have properties which may be generalizable to other domains. Then, particular and domain specific details of the model are abstracted away and the empirical model is transformed into a generic one. It is this type of modelling that has yielded to state that laws of self-organization occur at different time-scales and domains.

A very common methodology to extract valuable knowledge from these models is the exhaustive statistical analysis of different configurations of the system as well as measurements and operations performed on the resulting patterns. As a consequence, the space of resulting behaviours or structures may be classified and a set of internal relationships between parameters and resulting patterns are discovered to hold invariant under certain conditions (e.g. the number and distribution of connections within a network). This knowledge of the properties of generic models can feed back to the empirical domain both at predictive and constructive levels, given the appropriate adaptation of the abstract system to the real one.

A paradigmatic example of generic systems is provided by scale-free networks. Although the theory of scale-free networks was originally developed by physicist analysing the structure of the WWW (Barabási and Reka 1999, Barabási 2002) it has gained increasing attention in those scientific domains that deal with complex networks. Research in scale-free networks has discovered a

number of properties of the structure of non-random networks found in nature and the dynamics that such network architecture may generate (Strogatz 2001). Similar properties and structure apply to social networks, protein-interaction networks, computer networks, economic networks, and provide useful principles across levels (close to the use that model-based philosophers of science have attributed to physicals “laws” and principles). Stuart Kauffman’s explorations into generic properties of Random Boolean Networks (Kauffman 1969, 1993, Gershenson 2004) are another instance of generic models aiming to discover some general principles of self-organization in complex systems (independently of the social, genetic, metabolic, immune or neural nature of these systems). Recent systematic examination of natural networks (Milo et al. 2002) has delivered a picture of recurrent “building blocks” or network-architecture units, called motifs, whose distribution is much higher than what would be expected from a random distribution. Some of these motifs may perform stereotyped functions and their systematic classification and combinatorial analysis through generic simulations stands as a promising tool to understand generic network properties that can then be applied to specific cases

3.4. Conceptual models

The early enthusiasm within the field of Artificial Life about the genuine instantiation of living phenomena in the computer gave rise to a whole set of artificial worlds. Discussion about the plausibility, similarity, and adequacy of these models with existing theories of life produced both a fruitful debate about the underlying assumptions in theoretical biology and the status of such models as instances of living beings. As a result, models came to be used as a tool to question and reorganize theoretical assumptions and concepts rather than to the creation of so claimed artificial living systems without a clear epistemic purpose. Some of these models became conceptual ones which are, probably, Artificial Life’s most specific and original use of simulation models.

What I call a conceptual model involves the simulation of processes which are, in virtue of some dynamic or structural analogy with theoretical notions, *conceptualized* under a certain theory of the living, cognitive, social or, in general, complex systems. Conceptual models can be very abstract or very specific depending on the theory under which they are interpreted/constructed. For instance, at the abstract level, the model could work to illustrate, formalize or compare one or more theories of the concept of emergence using, let’s say, cellular automata patterns. On the other hand, a domain specific conceptual model can be exemplified by a simulation of active perception in situated agents.

Unlike generic models (whose applicability to the empirical realm is more standard due to its similarity with abstract mathematical research) conceptual

models have a more heterodox epistemic status: their relation to theories and empirical data is more complex and intricate. As Emmeche (1994) has pointed out, these models (which he calls “second-order simulacra”) are not elaborated as abstractions of the biological empirical domain, but from the biological theories themselves (i.e. from generalization and abstractions of other models and applications of generic principles). Such kind of models are not attempts to capture some empirical reality but are abstract worlds in their own right (e.g. Tierra—Ray 1992). Maynard Smith called Artificial Life into question arguing that it is a “science without facts”¹⁴, referring to the problem of how to assess a set of computational models whose (potential) empirical references are imprecise or non-existent. However, it would be an error to evaluate conceptual-simulation-model based research by traditional empiricist or observational standards. The main interest (and methodological novelty) of conceptual simulation models lies probably in their capacity to develop experimental research on the internal conceptual relationships within theories of biological organization. This computational research allows what Dennett (1994) calls the realization of highly rigorous and far-reaching thought experiments, which the naked human mind could never perform on its own.

Bedau (1998) and, particularly, Di Paolo and colleagues (Di Paolo et al. 2000) have elaborated a more detailed account of the role and methodology of Artificial Life as “opaque thought experiments”. The opacity of the thought experiment lies on the complexity of the model. The unfolding of properties and patterns from a set of premises (local rules or differential equations) are not always predictable in the absence of a computer simulation that performs recursive calculations, integrates random perturbations, visualizes the results and so on. As it happens with traditional thought experiments, the epistemic value of conceptual simulation models does not lie on their adequacy with some empirical phenomena (since the thought experiment involves hypothetical and idealised situations). On the contrary, the model operates on the hidden assumptions of the theories used to design and interpret the model and on the conceptual relationships between these assumptions.

When concepts of a theory are related to each other through relationships which cannot always be derived on logical grounds, computer simulations become cognitive tools for theoretical development (Casti 1997). For instance, learning and ontogenetic or phenotypic plasticity have intricate effects on evolution. The interaction between these two concepts (learning and evolution) is difficult to generalize and study through natural fossil records or other empirical means. Additionally, it turns out to be extremely difficult to theorise about the interplay between dynamics that occur at two radically different time scales and with consequences that only appear through repeated interactions and infinitesimal (and often non-linear) cumulative changes. An al-

14 Quoted in Horgan (1995).

ternative is to develop artificial worlds (whose local rules are abstractions of the generic mechanisms that evolutionary theory takes to be essential for natural evolution) where simplified forms of evolution and learning can be studied. The Baldwin effect (Baldwin 1896), for example, was nicely illustrated by a computer model by Hinton and Nowlan (1987) and gave rise to a revival of the subject (Weber & Depew 2003). Subsequent Artificial Life research has made explicit many other properties and dynamic relationships between learning and evolution (Ackley and Littman 1992, Suzuki and Arita 2004, Mills and Watson 2005) that remained opaque or hidden to the naked human thinking or analytic mathematical techniques. For instance, Mills and Watson (2006), making use of a simulation model, argued that genetic *assimilation* is, at least conceptually, sufficient for the Baldwin effect to occur and that *canalization* is not necessary.

The conceptual relationships that the models uncover, illustrate or deny are rarely the result of just running the simulations. Most of these models require a careful exploration and experimentation under different conditions in order to generalize results and find intermediate explanatory patterns to extract conceptually useful knowledge from the dynamics of the simulation. The resulting conceptual achievement shall later be used to configure explanatory patterns of specific cognitive, evolutionary, metabolic or social models subject to empirical manipulation and introduced on the traditional scientific nomologico-deductive or hypothetico-deductive method.

Conceptual models are used in a number of stereotyped ways. *Proofs of concept* are a typical use of these models in which the possibility to produce a particular behaviour is demonstrated in the model, given a set of mechanisms previously considered incapable of producing such behaviour or functionality. A good example is provided by Harvey et al. (2005). Whereas it is generally accepted that learning requires changes on synaptic connections in neural network research, Harvey and colleagues artificially evolved a dynamic neural network that was capable of learning without synaptic plasticity. Other considerations (such as the capacity of real networks to retain long term changes robustly without long term potentiation of synaptic connections) may be invoked to justify the necessity of synaptic modification for learning and memory. However, Harvey et al.'s simulation model provided a proof of a concept showing that the case is not, in principle, necessary.

On the other hand, models are often used to *illustrate, formalize or quantify a previously ill-defined concept* such as emergence, hierarchy, autopoiesis, representation, self-organization, etc. Here the model acts as a simplified environment in which (given that the competing theoreticians accept the assumptions on which the model is based) disputes over a conceptual definition can be solved in virtue of their viability, accuracy, correlation with expected classification, applicability to known phenomena, etc. In turn once an artifi-

cial patterns is taken to satisfy the requirements as a stereotypical example of the concept under investigation, different formalisms can be applied to further develop the concept (e.g. hyperset diagrammatic formalisms applied to different artificial models of minimal living organization, like tessellation automata or autocatalytic networks—Chemero & Turvey 2008).

Conceptual models are also used to study *interactions happening between different levels of organization* or different temporal and spatial scales, which are often studied by different disciplines, with different theories and tools. Examples of these levels that appear merged in the same simulation are neural mechanism and behaviour, genotype and phenotype, ontogenetic development and evolution, individual and collective behaviour, etc. Finally, conceptual simulation models can also be used to provide *reductio ad absurdum* arguments by implementing a set of assumptions and initial conditions and leaving the simulation unfold the dynamic consequences that are not apparent and contradict some of the premisses.

3.5. Animal models and *in vivo* synthetic approaches

Natural systems are far too complex to ask certain questions, particularly those that would require taking into account their holistic properties. To some extent the *raison d'être* of simulation models is the understanding of a theoretical domain that involves mechanistic models but some of whose key concepts (life, cognition, adaptation, development, etc.) are in need of an holistic integrated and often multi-scale approach. Simulation models provide the means to integrate various types of mechanisms with different degrees of abstraction and idealization (often at a purely conceptual level as we saw). In order to be able to extract relevant explanatory patterns and workable concepts, holistic models tend to minimalism: i.e. to integrate only necessary and sufficient conditions to generate a given phenomenon. But the ultimate benchmark for our theories of complex systems will come from simulating “the whole organism” or, more feasibly, creating *in vivo* minimal systems (like proto-cellular systems or robotic architectures with cultured cell) that, while “real”, may provide tractable objects to achieve knowledge about the most complex. Obviously, science has not been waiting for the possibility to create such artificial “living” models. In this sense, model organisms such as *E. coli*, *C. elegans*, or *D. melanogaster* have played a crucial role providing not-so-minimal but increasingly tractable models. Decades of empirical research and data collection regarding genetic, cellular, molecular, developmental and physiological aspects are opening the way to achieve integrated models of these organisms. It is by assembling the data that different scientific disciplines provide for such model organisms that more holistic questions can be empirically addressed: the interplay between genetic and behavioural pro-

cesses is a characteristic example. These integrated models provide, in turn, functional and conceptual breakthroughs to discover and understand the organization of more complex organisms (like humans).

4. COMPLEX THINKING AND ITS FOUNDATIONAL IMPACT ON BIOLOGICAL AND COGNITIVE SCIENCES

Before the advent of the new insights that simulation models and complexity theory made possible, the foundations of biological and cognitive sciences had to be largely based on those assumptions that could meet the demands of the mechanistic decompositional methodology. These assumptions had enormous consequences on the way cognitive and biological systems were conceptualized, particularly at higher level theories (such as principles of biological evolution or philosophy of mind). The one-to-one mapping assumption, for instance, somehow forced to jump over many levels of complexity to establish a workable one-to-one representational relationship at different scales. Consequently, as the synthetic modelling approach to complex mechanisms came progressively available, a set of transformations also followed on the conceptual frameworks of biological and cognitive sciences.

In biological systems this atomistic representationalism took the form of a one-to-one genotype-phenotype mapping whereby organisms were conceptualized as phenotypes encoded on genetic strings and evolution as a process of genetic variation and retention due to a selective forces that propagated linearly from phenotypes to selfish genes (Dawkins 1974). Research on genetic regulatory networks (Lewis 1992) has completely transformed the foundational framework of evolutionary biology together with the study of self-organization in developmental processes (Goodwin et. al. 1983, Kauffman 1993). Genes are no anymore seen as a freely recombinable encoding of biological forms but as a central but never unique responsible operator among the processes that produce organisms, immersed in a network of internal and interactive processes that constitutes the life-cycle of an organism. As a result, evolution does not operate upon an abstract functional phenotypic space but along a highly constrained (occasionally discontinuous) space of possible morphologies (Alberch 1982) whose formation requires acknowledging the environmental, material, self-organized and often random processes that appear networked at different scales (García-Azkonobieta 2005). Another assumption that had great consequences for theoretical biology was the sharp system-environment distinction that was made necessary in order to map internal structures to environmental conditions. When this sharp distinction is blurred by models of interactive emergence it is not any more the environment that poses problems for genetically instructed phenotypes to solve, but a

complex organism-environment interaction process that defines the adaptation and survival of an organism (Lewontin 2000), including ecological (Gilbert 2001) and niche-construction (Laland et al. 1999) processes.

The set of transformations that the sciences of complexity had in the foundations of biology were paralleled in cognitive science. Symbolic computational units and their rule based manipulation were postulated as theoretical primitives by functionalist approaches to the mind (Fodor 1975, Newell 1980, Block 1980). Cognition was primarily conceptualized as a set of inner symbol manipulation procedures on the side of the subject, mirroring external states of affairs (an a one-to-one mapping between internal abstract functions and environmental structures) in order to deliver planned actions to be executed in the world. The rise of artificial neural networks and parallel distributed processing in the 80's started to undermine the atomistic symbolic assumption, leading to a subsymbolic processing paradigm where cognitive processing became biologically grounded on the distributed nature of brain architectures (Rumelhart et al. 1986, Hopfield 1982). During the 90's the subject-object dichotomy became challenged by situated robotics whose engineering principles exploited recurrent system-environment interactions to generate emergent functional behaviour (Brooks 1991, Steels 1991). By the same time dynamical system approaches to cognition (Kelso 1995, Port and van Gelder 1995) started to conceptualize cognition as a dynamical process where cognitive behaviour and development emerged from dynamically coupled brain-body-environment systems (Beer 1995, Thelen and Smith 1994) challenging previous assumptions about the innateness of cognitive capacities (Elman et. al. 1996) and their purely internalist and disembodied computational basis. Where previous mainstream neuroscience was mostly focused on finding neural correlates of cognitive representations and localizing functions, brain activity started to be modelled as emerging from the collective dynamics of distributed neural ensembles at the edge of chaos (Skarda & Freeman 1986, Tsuda 2001) integrating its biological embodiment as a hierarchical interplay between internal bioregulatory functions and sensorimotor activity (Damasio 1994, Lewis 2005). Thus, the clear cut subject-object dichotomy and the abstract atomistic computational framework came to be progressively substituted by an internally and interactively distributed network of dynamical processes capable of giving rise to context-sensitive, flexible and multi-scale adaptive capacities, equally constrained and enabled by its biological embodiment.

This is the context in which current attempts to develop a concept for minds needs to be carried out. From Descartes to Ryle we were long missing the theoretical and empirical resources to conceptualize the mind as a complex mechanism. The question cannot wait for a complete realistic and detailed

simulation of the mechanisms that give rise to minds (e.g. to the full brain model promised by the Blue Brain Project for the next decade—Markram 2006). On the one hand, we are still really far from being capable of achieving any satisfactory and empirically valid simulation of the systems that we intuitively take to be cognitive (not even *C. elegans*' 302 neuron "brain" is accessible to present day real time recording and simulation techniques). But, more fundamentally a question precedes any simulation effort: What is the whole that needs to be modelled in order to achieve a successful concept of mind? The answer involves a set of feed-backs between partial mechanistic, functional and conceptual models that progressively delivers an integrated answer of the processes and contexts that should be included on a concept for minds. On the other hand, we need not even wait for a "total simulation" whose achievement, if it ever happens, could eventually become as complex and impenetrable as its explanandum. The challenge that remains achievable is to provide, at least at a conceptual level, a modelling framework able to start addressing, along the theoretical landscape opened by the sciences of complexity, what a model of mind would look like. This is the main goal of the present work. But it first requires some additional methodological clarifications: a) that we analyse in more detail the relationship between models, concepts and theories, b) the justification of the type of generic form to which cognition belongs as a "natural kind" within the wider class of complex systems and c) that we make explicit the set of epistemological constraints that the whole enterprise requires to deliver a valuable model. These are the objects of the next section.

Chapter 3: What a Concept for Minds Requires

Philosophy is the art of forming, of inventing, of fabricating concepts.
DELEUZE & GUATTARI

1. MODELS, THEORIES AND CONCEPTS

The received view of the philosophy of science took physics as the paradigmatic scientific field and delivered a nomologico-deductive model of scientific explanation (Hempel & Oppenheim 1948). Theories were taken to be axiomatized formal systems, where some of the axioms were generally understood as laws (universal generalizations). If axioms were considered statements with a truth value (a referential semantics) and theories become a set of statements about the world. In turn, phenomena were considered as subsumed under a universal law from which a predictive model could be logically deduced given certain boundary and initial conditions and the predicted results empirically tested through measurements. Models were primarily deductive objects, special cases of laws, concrete cases of theoretical statements. Therefore, what came to be considered the main scientific enterprise was the search for the laws of nature (finding the set of statements that truly describes the world) and model-building was considered a secondary quest aiming the confirmation or rejection of theoretical statements.

Nowadays, however, there is an increasing awareness that the main mode of explanation in science is captured, rather than by the concept of “law”, by the concept of “model” (Giere 1999, Giere 1979/1997, Cartwright 1983, Morrison 2000), and particularly in biology and cognitive sciences by mechanistic models (Bechtel & Richardson 1993, Glennan 1996, Glennan 2002, Bechtel & Abrahamsen 2005, Wright & Bechtel 2006). Models, then, become essential components of scientific knowledge-structures. Grossly, Godfrey-Smith provides the following summary of how models are conceived from the model-based approach to scientific theorizing:

A model-builder’s usual goal is to construct and describe various hypothetical structures. These structures are used to help us understand some actual target

system or systems. Generally, the understanding is supposed to be achieved via a resemblance relationship between the hypothetical and the real system. But both the degree and kind of resemblance that is sought are adjustable. (...) [T]he ability to describe and develop model systems in some detail, while remaining cautious or flexible about the particular respects in which the model might resemble the target system, is an essential tool. (Godfrey-Smith 2005:3)

This conception of models as representing or subsuming target systems by means of similarity or resemblance relationship involves an important dose of realism; which is not without its problems (Knuuttila 2005). A detailed discussion about the issue of realism and or, to use Giere's expression, "how models are used to represent reality" (Giere 2004) is out of the scope of this work. It shall suffice to make a set of provisory remarks. First, no model represents anything by itself, without a context of use that involves not only users (in the sense of "neutral" manipulators) but also the interpretative background of knowledge and intentions in which these users are involved. Second, models are not "there" in the abstract space of language, mathematics, computers or otherwise, ready to be used. Models are epistemic artefacts (Knuuttila 2005) that involve a full process of construction or production with important consequences for way in which they are put to explanatory work:

Treating models as artefacts makes their production processes visible. Seen from this angle a large part of the work of representation that is taking place in sciences is conveying into another form that is already represented and modelled somehow. Looking at representation from this point of view stresses the methods, ingredients and various representative devices that are needed in producing models. The production point of view on representation seems to me an important complement to the use point of view. It shows how any ready-made model is already a complex representative achievement in itself and not an isolated theoretical entity. I think that this has a certain sobering effect: one needs not be puzzled about what connects the model and its supposed target system since the model is from the very outset a result of various procedures of connecting. (Knuuttila 2005: 69—70).

What underlies Knuuttila's claim is that the boundaries of processes in reality are not rigid and absolutely predetermined so that models can be considered as mirrors of boxed reality bits. To a big extent, reality appears malleable and subject to different and overlapping moulding or modelling approaches. In turn, how models relate or connect with phenomena is a complex process in which the construction of the model bears a significant weight on its interpretation and therefore on its use as an epistemic artefact. A philosophical investigation about modelling minds requires to make explicit this constructive process.

Humphreys (2002) has proposed that the modelling process involves the following components: computational template, construction assumptions, interpretation, initial justification, correction set and output representation. The *computational template* is the artificial construct or artefact itself that is used as a model. Whereas Humphreys focuses on computational mod-

els of wide generality (thus the term “computational template”) his analysis can be extended to whatever is used as a model: a set of equations, a simulation, a physical three dimensional display or a linguistic structure. *Construction assumptions* involve approximations, idealizations, constraints and abstractions across which *ceteris paribus* clauses are included on the model. The construction process also provides a *justification* of the model even before it is applied or tested. In turn, an *interpretation* of the model is required to use it as a model in relation to its target systems and in relation to the inferences that can be drawn from it. This interpretation however, Humphreys argues, is not something to be added after or additionally to the model but is specified by the very construction process and, I would add, involves a much wider cultural and scientific background of practices or know-hows and preconceptions about reality. On the other hand, the *correction set* involves a somewhat inverse process to that of construction: the relaxation of the construction assumptions to meet the demands of particular cases of application of the model to empirical phenomena. Finally, the *output representation*, as Humphreys calls it, is the delivered explanation (the result of the simulation in a computer graphic, the analytic solution of an equation, the inferred consequences of a linguistic descriptive explanation, etc.) for which an interpretative process is also required.

Within this conception of science, theories may be considered as *families of models* across which interpretative frameworks and construction assumptions (including more abstract models or principles) hold also familiar¹. The term “familiar” is meant to denote a family-resemblance in the Wittgensteinian sense; a not fully formalizable nor uniform but common and analogous background context that brings all models to share essential features. For instance, the representational theory of the mind, can be said to include a family of models of perception, categorization, reasoning, decision-making, etc. The family resemblance comes from the shared background framework of internal representations as vehicles of cognitive organization. This is not to say that all models of the family share exactly the same working definition of representation or, for instance, that types of memory are equally treated in all the models of this family. But, certainly some interpretative frameworks and construction assumptions (centrality of representations, the abstract model of computational machines, the subject-object dichotomy, etc.) are shared among the various models of this theoretical-family.

1 Giere speaks of “collections of models”, but the term “collection” connotes a kind of arbitrary aggregate or a very poor relationship between them whereas my emphasis is put on a familiar relationship between them.

Assuming this position regarding the nature of models and theories, how would concepts (and particularly a concept for minds²) fit into the whole picture?

Philosophers of science are accustomed to thinking of categorization as preliminary to theoretical science, as natural history is preliminary to the theory of natural selection. At best, it is commonly thought, classification leads to empirical generalizations which are then systematized in theories. So, there is thought to be some relationship between categorizing and theorizing, but not one that promises much insight into the nature of theorizing. This reaction however, is itself grounded in the view that theories are sets of statements, perhaps sets of universal generalizations. (Giere 1999: 99–100).

Contrary to the received view, model-based philosophy of science considers that, couched in linguistic terms, scientific theories are composed of two types of linguistic entities: predicates and statements (Giere 1999). Models act as predicates, like the model of ‘two-body Newtonian gravitational system’ or a ‘computational machine’, and statements are attributions of models to target phenomena: such as ‘The Mind is a computational machine’, or ‘The Earth and the Moon form a two-body Newtonian gravitational system’. Interestingly, Giere notes, “the linguistic structure corresponding to a predicate is not that of an axiomatic system, but of a definition” (Giere 1999:98). In particular, most predicates take the form of concepts or categories.

Following the work by cognitive psychologists, Giere considers how psychological and scientific concepts and categorizations display an inclusive hierarchical structure: e.g. living systems, plants, trees, evergreens, pine, white pine. Along the hierarchy there is generally a basic level category (e.g. tree). Objects of this basic level category share more common features between them than do the objects included on the level above. The basic level category is more informative than the lower level categories on that the gain in similarity of the lower levels is smaller. For example, the category “tree” is more informative than “evergreens” in the sense that there are more distinctive features in common between trees (in comparison with other subcategories of plants) than there are between evergreens in comparison with other subcategories of trees. Basic level categories, in turn, are pictured as prototypes or idealized models from which cases “radiate” with a higher or lesser degree of similarity.

Giere applies this schema to scientific modelling and shows how the concept of a pendulum becomes a basic level model in classical mechanics, depicted by the prototypic model of a simple pendulum. The simple pendulum is subsumed under the higher category of “harmonic motion by linear restoring force” and can be further specified into lower level categories by in-

2 I shall repeatedly use the term “a concept for minds” instead of “the concept of mind”. The favoured expression is meant to be less committed to a extreme referential realism and more sensible with a conception of models as moulds-for certain, multifaceted and often complexly evasive, phenomena.

troducing additional constraints that give rise to models of dumped pendulums, etc. Interestingly, basic concepts or categories do not apply directly to real-world objects. This is not to say they have an undetermined epistemic status. On the contrary, they are both idealized cases of directly empirically applicable models and can be inferred or deduced (given appropriate constraints) from higher level models or general principles. It is this basic level of modelling that I shall consider characteristically *conceptual* and the domain in which the present work is to be developed.

We can now picture the task of developing a concept for minds in relation to the broader context of scientific theories and models. Models are hypothetical constructions that involve a set of assumptions (approximations, idealizations, constraints and abstractions) and interpretative frameworks that are essential to put them at explanatory or epistemic work. Theories, in turn, are families of models that share construction assumptions and interpretative frameworks and that appear hierarchically structured. Within this hierarchy some concepts occupy a privileged status as basic level models, whose construction involves a circularity between empirically connected models and higher level theoretical principles. The task of developing a concept for minds will thus involve the making explicit of the process of production of a conceptual model within a theory of complexity (as sketched on the previous chapter).

Before starting with the construction process we shall, through the rest of this chapter, make more explicit what is the form that a concept for minds should take, what set of epistemological constraints are going to guide our construction and what should be taken as an adequate departure point. These tasks constitute the rest of this chapter.

2. CONCEPTUAL MODELS AS EXPLANATORY AND GENERATIVE DEFINITIONS

In the model-based approach to the philosophy of science, a concept for minds, its definition, would naturally take the form of a model, whose construction involves assumptions that are shared along a theoretical family of models. In particular, a concept for minds would resemble the conceptual simulation models used in Artificial Life: they do not connect directly with experimental domains but are evaluated against other (generally abstract) theoretical models and more specific empirical models. But conceptual models of this sort need not necessarily be formalized or computerized. Even simulation models (as characteristic objects of complexity theory) are generally preceded by a linguistic-conceptual formulation (often accompanied by diagrams and illustrations) that already integrates much of the construction assumptions

and fixes the interpretative framework. Our task would not involve so much the construction of a definite computational or mathematical construct but rather a linguistically expressed conceptual model. In turn, linguistically expressed conceptual models can be connected with more tractable formal, simulation or animal models engaging with empirical evaluative procedures. In this sense the construction of a conceptual model for minds will involve a circulation along the hierarchical structure of scientific theories: from principles of self-organization to mechanistic models of bacterial chemotaxis.

I have chosen to depart from the hypothesis that minds are dispositional behaviours produced by complex organized mechanisms and therefore it is a model for this type of systems that would be required on the construction of a conceptual model for minds. Complex systems require special kinds of mechanistic models, ultimately models of emergent hierarchical organization, as we saw in the previous chapter. Thus, defining cognition would require the construction of a conceptual generative model that captures the causal organization that makes possible and specifies the nature of Mind as a specific phenomena, different from the generic classes biological or complex systems. This specification involves at least two major aspects. The first involves to make explicit the causal domain in which Minds are to be pictured out and the relationship of this domain with its underlying hierarchy and context. The second is to find a characteristic mode of organization within this domain capable to generate the phenomenology of minds and cognition. We shall analyse them separately

Domain specificity: When building models, scientists and engineers tend to slice reality in the causal domain that becomes relevant for their object of study, isolated from the rest of processes that might underlay it and surround it. For instance, an engineer will tend to find and use materials that remain stable for the task they need to satisfy (or flexible just in those aspects that are contained on their assigned functionality). The aeroplane engineer does not want the wing to start growing nor the wheels of the plane to melt, thus she uses components whose properties (other than those required by the design) remain stable. This is why when working with a model scientists rarely pay attention to the processes underlying their domain of modelling. Either the scientific field has already been built around a domain (in a historical feedback process by which scientific disciplines and studied phenomena constraint each other isolating a causal domain in the process) or the components have been chosen to isolate a causal domain in which modelling or design are possible. In sum, all activity of modelling involves, implicitly or explicitly, the specification of a level of observation and manipulation in which certain (modelled) regularities take place. Therefore the process of making explicit the production of a model requires to isolate and specify a domain of modelling (the types of variables and relationships involved). So, for instance,

before we build a model of the functioning of the circulatory system (even before we understand that there is a circulatory system as we know it today) it is of fundamental importance that we understand what the underlying components are and how they relate to the rest of the body. The engineer might not need to know the molecular details of his building material; it is enough to know that it has the required resistance, conductivity or flexibility. The chemist might not be interested on the details of the MIT bag model of the atomic nucleus if the chemical reactions at hand are not causally affected by sub-atomic processes. But the physiologist needs to know that tissues are composed of cells, that cells need oxygen and that oxygen needs to be carried out from the lung to every cell, if she wants to build an adequate model of the circulatory systems. Since mindful systems appear within hierarchically organized systems (where top-down and bottom-up fine grained interactions take place) it is of fundamental importance to make explicit what the relationship is between the domain of mind-modelling and the underlying processes that sustain and interact with it: e.g. the living body and the environment. How this level or domain of observation and modelling is constituted in the first place (how it appears in nature) and the analysis of the set of organizational hierarchies in which it is embedded on (and from which some properties might be inherited or some relevant constraint might appear essential) will become a big part of what is to come.

Organizational specificity: It is here where the Rylean program meets the modelling process. A model for minds should be capable of generating the behavioural dispositions that he took essential for the concept of minds. Thus a complex mechanistic model that defines a concept for minds involves the construction of a set of interrelated parts capable to *generate* a set of characteristic features. This is indeed a feature shared by most of models in complex sciences: models are built so that interactions between lower level components or structures give rise to emergent functional/behavioural properties. Interestingly, a definition becomes also an *explanation* through a generative model whose dynamical simulation reproduces the emergent properties of the target systems. To define a very complex system (in the Kolmogorovian sense) implies to reduce it to a manageable generative description. If this description can be produced as a model of the causal organization or mechanisms of the system, a description that defines a system becomes also an explanation. In this sense, defining and justifying the constraints, functions, structures and relationships that give rise to the model becomes a big part of the modelling and its explanatory enterprise. Imagine that you make a foreign friend that asks you to define what a football match is. You will most probably start describing what is going on during a stereotyped football match. This description would involve an explanation of what the components are (different types of players) and how they relate to each other and the ball (the rules

that govern their behaviour both in terms of intentions and goals and also in terms of constraints—forbidden moves, boundaries of the camp, etc.). The overall picture that your friend would get, if you succeed on the definition, is a generative model where components and rules produce a characteristic game. If you want your friend to understand football in more detail you might move to use a computer simulation (like a typical Play Station game) so that your friend can “experiment” different situations, tactics, and strategies.

I can now recapitulate by saying that the goal of this thesis is that of *making explicit the construction of a model for minds as a complex generative organization: i.e. as a class of hypothetical systems in which a set of component processes relate to each other in a specific interdependent manner capable to generate a set of characteristic features that distinguish mindful systems.*

3. MINIMALISM, UNIVERSALITY AND NATURALISM: A MUNDANE DECLARATION OF PRINCIPLES.

In order to attain the goal of fabricating a concept for minds we will make use of other models; some of them will be more general and abstract (like principles of self-organization) other more concrete (like the detailed description of the mechanisms involved in bacterial chemotaxis). We shall also make explicit the construction assumptions that the model involves and we shall specify the modelling domain. But first, I would like to stop and propose a number of epistemological constraints that should bound and guarantee that the construction process meets some epistemological standards in order to transfer the model to the more empirically testable domain of scientific discourse. Let me start with Naturalism, then move to Universality and finally tackle the Minimalist constraints. I shall hereafter refer to these constraints as the MUN constraints.

3.1. Naturalism

Naturalism is a widely spread philosophical position stating that *ad hoc* substances are not to be introduced in a model in order to explain the target phenomena (Hooker 1995). Thus, for instance, consciousness or information as theoretical primitives (Chalmers 1995), representations or logical structures, should be avoided as foundationally privileged departure points. However, what constitutes an *ad hoc* substance or property is a difficult matter. The way out of this dilemma is to think of naturalism as a scientifically embodied philosophical practice both at its sensory and motor surfaces (so to speak): i.e. it should be grounded on available scientific knowledge and be able to feed-back to scientific practice (through its capacity to generate new hypo-

theses, to provide principles to reorganize knowledge, clarify concepts, uncover fallacies, etc.). In addition, one should wonder in which type of scientific field must one be embedded on. In this sense, I shall expand our naturalist constraint to encompass an additional requirement: a *bottom-up* approach. By bottom-up I mean that concepts and components of our model should be built from the most simple and elementary (in relation to a given level of organization of empirical research) to the most complex and higher order ones. In particular, we will defend a biological *grounding* by which components of our model should be derived or closely related to more fundamental biological processes³. Thus, if we were to include meaning, for instance, as a characteristic feature of minds, it would be inappropriate to attribute semantic properties to component neural ensembles if meaning is pre-defined strictly in higher level terms (for example, as linguistic performance or in reference to the use of dictionaries, with no reference to its biological and neural grounding). Higher level descriptions should be accompanied and grounded on bottom-up explanations of how those phenomena can be sustained and emerge from lower level organizational principles. In this sense, naturalism does not require that we connect our conceptual model directly with specific empirical studies, but that we make explicit the route of an indirect connection and build the model so that such a route can be traversed.

A bottom-up approach includes a final naturalist constraint: that observer dependent properties should not be attributed to the model itself if no specific procedure is established to reconstruct them in a bottom-up observer-independent manner. Observer-dependent properties are relational properties that may be accessible to the external observer's privileged position but do not arise from the intrinsic relations of the model. They include, for instance, system-environment correlations (like certain areas of the visual cortex responding in correlation with certain environmental features), the designer's intentions (like the introduction of a new function into a model *to represent* emotions) or the evolutionary history of the system under study. These properties are somewhat to play an explanatory role on the model they should emerge or be directly present on it. Otherwise we projecting observer-dependent properties to the model and from the model to the explanation of the target system. This form of projection of observer-dependent properties is a common mistake that appears on many human made and interpreted devices such as computer programs or robots (Clancey 1989).

3 The bottom-up approach does not forbid to use top-down methods, in fact a top-down bottom-up circulation will be of significant importance. What the bottom-up approach emphasizes is that the role of top-down methodologies is better understood as a form of heuristics and guidance of the bottom-up grounding.

3.2. Universalism

Our first MUN constraint mandates that we build our model from a naturalist perspective, which entails a bottom-up biological grounding of the concepts and components belonging to our model for minds. Our second constraint is *Universalism*. Currently available biological systems amenable to experimentation and study are the result of a set of historical (evolutionary) contingencies. But knowledge has universalistic aspirations. As Artificial Life founder Christopher Langton (1989) claimed: it is not life-as-we-know-it but rather life-as-it-could-be that is of interest to the field. We could equally define our object of study the-mind-as-it-could-be rather than the-mind-as-we-know-it. This forces us to define generalizable patterns of life and mind rather than focusing on particular anatomical details of present mind-supporting brains and bodies. For instance, if emotions are to be part of our final model it would be inappropriate to say that emotions are defined by the signals coming from a particular neural pathway, as if human brain anatomy was to determine what emotions are to be in the Universe. It might be the case that some psychological processes be unambiguously identified or correlated with certain brain areas, but this is not to say that what that process is equivalent with a certain anatomical component or set of components that happen to be the locus of such emotions in planet-earth vertebrates (Rohde & Di Paolo 2005). Therefore the universality constraint requires that we make models as generalizable as possible.

3.3. Minimalism

Minimalism is our third and final epistemological constraint. Minimalism might be seen as a direct consequence of our first naturalist bottom-up constraint but it is worth making it explicit as a specific requirement in itself. It states that our model must contain all but no more than those features that are necessary and sufficient to define the class of systems that it targets. So, rather than taking the higher level epistemic properties or language like sophisticated mental phenomena as a departure point, minimalism states that we should proceed making use of the simplest and more amenable components in order to build a model. This being said, and this is an important remark, the upper boundary for complexity increase must remain open. For instance, Neils Bohr, inspired by Rutherford, proposed the planetary model of the atom taking as a departure point “a simple system consisting of a positively charged nucleus of very small dimensions and an electron describing closed orbits around it” (Bohr, 1913:3). Bohr's model, although focused on a simple system (the Hydrogen atom) to start with, was built with the rest of atomic forms in mind, so that components (electron orbits, nuclear forces and their relationships) could be aggregated to form more complex models once the minimal

one was satisfactorily constructed and tested. Equally, we should tend to generalizable and expansible models, where a minimalist core stands as a foundational first step that permits to organize and discuss conceptual and empirical relationships. In the absence of a complete model, some properties might be studied on partial property-specific models. In this sense, formalization and computer simulation permit a common language to recombine and integrate achievements and components from different local or particular modelling experiments.

But, unlike Bohr's case, we face a situation where there is no generally accepted and empirically available minimal target object to model. We are lacking the Hydrogen atom of the mind. What constitutes a genuine example of minimal cognition (not to speak of minimal mindfulness) remains an open issue which deserves much more attention than what it currently receives⁴. In such a situation, minimalism is a methodological remedy for the study of complex systems. So let us imagine that there was nothing like a one electron + one single proton atom left in the universe: only complex macromolecules to experiment with. In such a case we could proceed by creating something like an artificial atomic-physics that was to construct complex simulation models of non-existing atoms. Out of this model an artificial chemistry could be constructed which, finally, could be compared with experimentally available target macromolecules. The Hydrogen atom of life and mind must be reconstructed from what we take to be coherent with our present knowledge of biological and neuro-psychological phenomena.

3.4. MUN-modelling: the epistemic engine

The model-based generative modelling approach sketched in the previous section “naturally satisfies” the MUN constraints: a) *Naturalism*: a.1. if it satisfies the principle of scientific embodiment if it is operationally grounded (if it provides the means to relate variables in the model with variables in the natural objects or justifies their abstraction or idealization), a.2. it shall satisfy biological grounding if it makes specifies the modelling domain: if it justifies the level of modelling and its relation to the rest of the processes that sustain and surround it; b) *Minimalism*: if the model is simplified to capture just those necessary and sufficient features of cognitive organization and c) *Universality*: if we provide the means by which the model can be, in principle, applied to any new system we encounter. This matrix of epistemically normative boundaries is what should be used to evaluate the result of the present work.

4 There are a number of recent exceptions like Randall Beer's target article and its commentaries (Beer 2003) on minimally cognitive robotic agents or van Duijn and colleagues' exploration into the principles of minimal cognition (Duijn *et al.* 2006). Together with Alvaro Moreno I have also addressed this question somewhere else (Barandiaran & Moreno 2006b).

The conceptual modelling approach, together with the MUN-constraints constitute a sort of epistemic engine. As any machine or organism, the constraints are also enabling (like the pipes of an engine, the rules of chess or the joints of your arm) and will guide the production of a concept for minds with some methodological guarantees. What rests to be done is to find a departure point, an input for the machine. We do not stand alone in this task. Philosophers and scientists have long dedicated efforts to produce mechanistic and functionalist enclosures to delimit cognition as a constitutive feature of minds. In the following sections we shall travel along some of these attempts to end up with agency as a departure point and propose a *morphophylogenetic* approach as a scaffolding for our construction process.

4. ENCLOSING COGNITION: THE PROBLEM OF DELIMITING BOUNDARIES

Aristotle divided the soul on three main parts or components: the vegetative soul, the appetitive soul and the rational soul; characterizing respectively the plant, the animal and the human kingdoms. By the term soul nothing homologous to a ghost was invoked by Aristotle but, on the contrary, a kind of form, of functional organization of matter, responsible for the type of order and flexibility shown by living systems. Unfortunately the rational soul (first divine, always anthropocentric) captured most part of the philosophical attention throughout the history of western thought, with the Cartesian extreme dualism at its climax (in which two separated substances demarcated the mental and the material/mechanistic realms). The conscious realm of cognitive entities was postulated as the locus of general purpose intelligence and rational thought. But as the Cartesian mindset started to be shown inadequate to analyse and create intelligent systems (see Wheeler 2005 for a detailed account of the Cartesian legacy and its decay) the enclosure of cognition has become substantially challenged. Today, any of the classical Aristotelian souls can be found to be the locus of cognitive research within natural sciences: from the characterization of the most simple (unicellular) forms of life in terms of information processing (signal transduction mechanism, gradient perception in chemotaxis, etc.) and the study of animal intelligence and adaptive behaviour (from ants to apes, through octopus and molluscs), to the mathematical and computational study of logical (rational) thinking, including the study of machines with computational capacities. The anthropocentric monopoly of cognitive powers is nowadays redistributed into an open cognitive market composed of bacterial chemotaxis, computer networks, insect colonies, techno-scientific institutions, brains, rats and robots (among others). A varied set of philosophical, methodological and experimental turns have

been responsible for this aperture. Let us consider them in more detail as they unfold the present situation concerning the specificity of cognition.

For a considerable period of time (from the 50's to the late 80's), cognitive science enjoyed a peaceful period, taking human intelligence and its linguistic and logical capacities as the centre and measure of cognitive research and taking the computer metaphor as the methodological and philosophical *vade-mecum* against an always threatening metaphysical dualism. During this modern period⁵, cognitive enclosure was somehow intuitively provided by the paradigmatic case of human abstract intelligence. But this enclosure soon became threatened under the possibility of machine intelligence and the so-called Strong AI program. If the paradigmatic case of intelligence and cognition was provided by human abstract reasoning capacities, such as chess playing, deductive inference or linguistic performance, it soon became manifest that such capacities were (at least partially) reducible to disembodied computational manipulation of representational symbols. Under these premises, there was nothing left to avoid sufficiently sophisticated computer programs to become instances of genuine cognitive processes. This became the core of a long standing debate between practitioners of Strong AI (defending that, up to a certain—not well defined—level of complexity, computer programs could become genuine cognizers) and Weak AI (stating that computer programs lacked a certain essential feature of cognition—consciousness, semantics, embodiment, etc.—and should not be considered other than valuable tools to study certain aspects of cognition).

Another source of threat to the anthropocentric cognitive enclosure came from cognitive science research on animal behaviour. The rise of cognitive science, with its well established foundations on the information processing paradigm, permitted to apply its methods to other realms of the living. The information processing talk was applied to molecular signalling between the cells forming a tissue, to bacteria swimming up a glucose gradient, to neural signals in worms, to adaptive plant growth, to immune recognition of antibodies or to pheromones in an ant colony. A too wide definition of computational states and information processing allows to interpret almost any changing higher level regularity in biological phenomena as cognitive, including that of plants (Godfrey-Smith 1996, Trewavas 2003, Balusk et al. 2005).

5 The key baptismal justification for labelling this the *modern* period of Cognitive Science relies on its rationalistic optimism, its cognitive anthropocentrism and its representational adequationism (the idea that mental-cognitive states are mirror images of worldly states of affairs). Less scientifically minded but more philosophically oriented versions of these three characteristics were considered by post-modern continental philosophers as defining the core architecture of modern western thought. The post-modern turn, on its linguistic and socio-historical dimensions, downplayed the idea of an absolute criteria for truth and reason as context-dependent phenomenon, the overcoming of anthropocentrism by super-individual structures (language, history, paradigms and social technologies) and the world as a socio-historically embodied domain. The similarities between post-modern philosophy and some philosophical implications of embodied, situated and distributed approaches in cognitive science is difficult to obviate.

In addition, another long standing boundary of cognitive processes, the skull, also became threatened. What has come to be called extended and distributed cognition (Hutchins 1995, Clark & Chalmers 1998) discovered that much of the information processing load in cognitive processes actually happens outside the limits of the skull (or any other centre of signal convergence within the limits of the individual organism). Cognition spreads non-trivially into distributed social processes, external technologies (from pen and paper to computers), natural landmarks and a set of developmental scaffoldings (including language, interpreted as an external cultural technology). Thus, although originally developed as an abstraction and generalization of human rational logic-linguistic capacities, this “modern” paradigm in cognitive science has seen a progressive liberalization of the cognitivist methodology and conceptual apparatus to encompass a wide range of phenomena.

The turn of millennium brought about the rise of a new paradigm within cognitive science. This time the emphasis was put in biological, embodied and dynamical aspects of cognition, taking biological adaptive behaviour as a paradigmatic example of cognition (Brooks 1991, Beer 1990, Hendriks-Jansen 1996). The lessons learned from animal intelligence and distributed-extended cognition, including a set of paradigmatic critiques and accompanying successes in robotics and developmental psychology, were integrated with a, still more encompassing, theoretical and methodological framework: that provided by dynamical system theory and its accompanying philosophical new foundations for cognitive science (Port & van Gelder 1995, van Gelder 1998, Beer 2000). Expressed in terms of the philosopher Tim van Gelder, the dynamicist turn proposed an ontological claim (van Gelder 1998) stating that cognitive systems are instances of a dynamical causal organization and a methodological claim stating that cognitive systems are better understood within Dynamical System Theory (DST) in clear opposition to computational approaches and the computer metaphor.

Against the abstract and disembodied character of the computational paradigm, embodiment and situatedness became also a central focus of attention for cognitive scientists. The *dynamical hypothesis* (as labelled by van Gelder) permitted to shift the information processing talk into the terminology of self-organization thus allowing for a more integrated study of brain, body and world as a continuum in the production of cognitive behaviour. But neither the dynamical hypothesis nor DST itself offers any criteria to delimit cognitive phenomena. As van Gelder puts it (and the question is similarly addressed or ignored by many other dynamicists): “This paper simply takes an intuitive grasp of the issue for granted. Crudely put, the question here is not what makes something cognitive, but how cognitive agents work” (van Gelder 1998: 819). The new situation regarding the enclosure of cognition as a specific phenomenon, appears in still a weaker position than its predecessor’s

computational framework. If the computational paradigm came to be too close to generic computing machines and many biological phenomena interpretable in terms of information processing, the dynamicist paradigm is dangerously close to the wider range of the physical realm (not surprisingly, dynamical system theory was born to solve an early challenge of Newtonian mechanics—i.e. Poincaré's solution to the three body problem). Self-organizing and, in general, dynamic phenomena are found in all scientific realms: for instance, the bifurcation phenomenon that can be seen in a model of a decision making process can equally be found in models of chemical, economic or physical systems. Surely, the "intuitive grasp of the issue" that van Gelder takes for granted is not willing to accept generic physical, chemical or economic dynamics as cognitive and yet, it fails to propose specific criteria capable to demarcate the cognitive from the non-cognitive.

Thus, we are faced with two main paradigms of cognitive science (the boundaries of which are nevertheless permeable). On the one hand, cognition is characterized as *information processing*, as an abstract, disembodied and computational phenomenon. On the other hand, cognition is characterized as *behavioural adaptive flexibility* as a rather biological, embodied and dynamical phenomenon⁶. And both paradigms risk of pan-cognitivism: the possibility of (almost) everything becoming cognitive if the boundaries are not sufficiently well established. And yet, cognition appears as distinguishing itself from its surroundings on its continuous defiance to the language of physics, on its comfortable accommodation to the intentional stance (Dennett 1987), on its creativity as well as stupidity, worry and anticipation, desire and frustration. However, its demarcation is probably not amenable to the fencing, enclosing and defence of a determinate territory. It may turn out that a more productive approach be found if we open the landscape and try to define a gradient, a tendency, the perspective of a mountain chain whose peaks denote a genuinely cognitive character and yet provide a path for climbing from the ground level of the well established physical sea to the Olympic crest where cognitive divinities live. But... what could permit to redefine the terrain so as to find cognition on its peaks and not everywhere around the ground or in a transcendent sky (be it consciousness or pure reason) that lies beyond the accessibility of a naturalist research program?

6 A third approach might also be considered relevant here, not as a branch of cognitive science *per se*. We are talking of molecular biology and the *ruthless reductionist* claims (Bickle 2003) that some are willing to promote on its basis. This field aims to describe "how neural science is attempting to link molecules to mind—how proteins responsible for the activities of individual nerve cells are related to the complexity of mental processes" (Kandel 2000:2). But molecular biology is not without its own problems to provide a rigorous and explicit definition of what does delimit and specify the mind and its complexities. An issue that is often neglected as irrelevant by reductionist approaches without acknowledging how much the field itself depends on implicit (and often poorly developed) assumptions about the question.

In line with some dynamicist approaches, an alternative path to that followed by anthropocentric (logical-linguistic or phenomenological) enclosures of cognition is to define a bottom-level generic form that all cognitive systems share and that permits a gradual approach towards the mind. A naturalist hill-climbing strategy may be envisioned that allows for an adequate characterization of the cognitive peaks of nature (and define its accessibility or inaccessibility to the artificial).

Thus, if we are to rule out language, thought or consciousness as a departure point, a good alternative might be provided by the wide set of “systems capable to perform actions”; i.e. agents. This is the path taken by many dynamicists, specially by those centred in autonomous robotics, ethology and adaptive behaviour. Most of them, however, do not directly address the issue of the specificity of cognition: they are happy to consider simple adaptive agents as genuinely cognitive or assume, rather uncritically, that cognition or intelligence is nothing more than the progressive complexification of the adaptive capacities. What remains interesting from this perspective is that grounding cognition in agency provides a gradual path towards increasingly complex forms of agency while it permits to naturalize this gradualism through the set of examples and models of agency widespread in nature (from bacteria to monkeys, from jelly-fish to flat-worms). The idea is not really a new one. Evolutionary and biological epistemology (Lorenz 1977, Campbell 1974, Piaget 1967/1969) or adaptive and control systems approaches to mind in cybernetics (Ashby 1952), to mention but a few, had previously suggested approaching the mind through its evolutionary origins on biological grounding. Today, the huge amount of data that has been gathered from animal models and their articulation in complex simulation models permit a much detailed reconstruction of agency as a generic template for minds. The way is open for what Pamela Lyon (2006) has termed a *biogenic* approach to cognition that takes agency as a non-anthropocentric reference to ground cognitive capacities. In turn, agency, on its broader sense, remains open to include increasingly complex properties and capacities.

5. WHAT AGENCY REQUIRES

What we finally need is to bound a generic model of agency that serves as a template for our construction process for the more concrete model of mind. Working-definitions are good candidates to start delimiting this generic form, and the rise of synthetic robotic approaches to cognitive science gave rise to an explosion of definitions of agency. For instance, Russell and Norvig on their classical AI handbook (1995: 33) propose that “an agent is anything that can be viewed as perceiving its environment through sensors and acting upon that

environment through effectors”. Robotist Pattie Maes (1994), on the other hand defines an agent as “a system that tries to fulfil a set of goals in a complex, dynamic environment”; in turn, Beer (1995) considers an agent “any embodied system [that pursues] internal or external goals by its own actions while in continuous long-term interaction with the environment in which it is situated”, while Smithers (1995: 97) states that “agent systems are systems that can initiate, sustain, and maintain an ongoing and continuous interaction with their environment as an essential part of their normal functioning”. After an extensive review of different definitions of agency (including some of the previously mentioned ones), Franklin and Graesser (1996) conclude that “an autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future”. Yet, most of these definitions strongly rely on the intuitive notions of sensing and perception. In fact, perception and action can be seen as the cornerstone that distinguishes agency from mere happening or physical coupling between a system and its environment. Specially aware of this crucial transition from the realm of the physical to that of the agential, Stuart Kauffman (2000) has defined an agent as a system that “can act on its own behalf in an environment”. Following his work, Ruiz-Mirazo and Moreno defend that minimal autonomous agents are those chemical systems capable to actively constraint their boundary conditions for self-maintenance (Ruiz-Mirazo & Moreno 2000).

Abstracting away from the particularities of the above definitions and from the most descriptive approach (i.e. without hypothesizing internal mechanism), we can generalize that agency involves, at least, *a system doing something by itself according to certain goals or norms within a specific environment*. From this early characterization, three different, although interrelated, aspects can already be distinguished: (i) there is a system as a *distinguishable entity on its environment*, (ii) this system is *doing something by itself* on that environment and (iii) it does so according to a certain goal or *norm*⁷. As we shall see, all three aspects will become crucial to fully specify what the minimal form of agency is and to define the gradient towards the cognitive. Eventually, cognition might unfold as a particular kind of agency: either specified by a type of agent, or by some specific kind of norm or goal that the agent follows, or by a type of action, or by a particular kind of relationship with its environment, or by special type of environment (call it a cognitive world), or, finally, by a mixture of some (or all) of the above possibilities. In any case, if we are to proceed from the bottom-up, not taking any distinction or presupposition for granted, we shall examine in detail what agency

⁷ Note that, although it does not in itself constitute a requirement, the environment appears as a key notion in two respects: a) as that from which the system distinguishes and b) as the domain in which actions occur.

involves in the first place, what anything subject to the generic form of agency should require and how those requirements already determine a set of relationships and properties.

5.1. Individuality

First of all, and most immediately, for a cognitive process to be in place, there must be a system-environment distinction, where we attribute some capacities to an entity in relation to its environment. This we shall call the **individuality condition**. Cognition and, most fundamentally agency, implies a basic duality between system and environment. This fundamental duality is very often taken for granted, and the characterization of cognition is reduced to the establishment of the *kind* of relationship (representational, informational, intentional, adaptive, etc.) between agent and world. But, since no specific environment neither cognitive relation with it could exist before or apart from the constitution of an agent, the first condition for the appearance of cognition must be the presence of a system capable to define its own identity, thus distinguishing itself from the rest of the world and, in doing so, cutting up a specific environment with which to establish a “cognitive relationship” or within which its actions are to be carried out. Note that the system be required to distinguish *itself* from its environment is a very strong condition and it deserves a careful analysis. If we were describing a continuous and undifferentiable process, in which no distinctions could be made so as to distinguish some part from the rest (e.g. an homogeneous hectometre of gas), we could not even be able to determine what might constitute the system and what its environment (not to speak about *identifying* a locus of agency). So, the first condition seems that there be parts or distinguishable structures that can be said to be components of a system or that constitute the system in some way. But we are still left out with the question of which components are going to be part of the system and not of the environment. In fact, as observers, we can establish any arbitrary separation in order to define a system and its environment. The *problem of identity* shows up as the following question: from the set of possible and arbitrary separations between system and environment, which is the one we choose and why? We can coherently justify the distinction with reference to certain functionality or usefulness that the composition might present for us. So, for example, we might agree on declaring that the table, pen, paper, computer, lamp, etc. constitute a system (a technological device that allows me to develop my academic work). But, in this case, what belongs to the system and what does not is determined by the intentions of an external agent (myself) and the set of useful possibilities that the full device affords for my abilities. No intrinsic force or process is lumping the components together, nor the system as a whole (independently of me) seems to have a specific way of functioning and demarcating itself from the rest of my office.

Even if some form of cohesion is present, lumping together component parts, a similar problem remains for the ascription of identity:

[T]he decisive observation, of course, is that to the artefact the identity is *accorded*; and, insofar as this requires a continuity of memory and tradition in those who do accord it, the identity is the function of another identity, namely, that established in memory, individual and social. This originaive identity of the cognitive subject is a prerequisite for the accorded identity of the object. (Jonas 1968: 240).

But unlike other entities, agents appear as unified systems in themselves and do not depend on their being useful for an external agent or accorded on their identity by a community of other agents in order to become what they are. If a model of agency cannot account for the way in which an agent defines its own identity, the system-environment distinction will require a regress to infinity. If the agent's identification depended, in turn, on another agent, we would be required to justify the identity of this second agent by means of another one and so on *ad infinitum*. In addition, and this will constitute a big part of what is to come, agents seem to define themselves through the actions they generate; and this brings us to the next necessary condition that constitutes agency

5.2. Causal asymmetry

A system, with a definite identity (however this might be mechanistically achievable), must be doing something (or being capable to perform some kind of operation on its environment) by *itself* in order to be an agent, i.e. we assume some kind of asymmetry on the causal structure of the agent-environment interaction. In other words, and although it may seem to be rather obvious (but this is precisely the job we are involved on: to make the obvious, and the not-so-obvious, explicit), we presuppose that the agent is *an active source of action*, not just a passive entity suffering the effect of some other forces or origins of action. Let us call this, the **causal asymmetry condition**. So, for instance, even if we had a well identified entity, let's say a philosopher, if the system-environment interaction involves a wall falling down and smashing our friend, on that particular interaction, we would have no reason to call the philosopher an agent whatsoever but, rather, a patient of the event and, most probably, a patient again, later on, in a recovery-hospital (suffering the painful treatments of a medical agent). In addition, although the catastrophic event was somehow originated on the wall-supporting structure, we certainly don't want to call it an agent either. The event was the result of physical laws and processes (gravity, tension, friction, etc.) affecting a passive structure (the wall), so that the wall cannot be said to be the *active source* of the event either. That is why we usually label this kind of events as "accidents", where no agency is involved as such. Being an active source of action somehow in-

volves that there is a kind of complexity asymmetry on the system-environment causal continuum so that a goal or a functional relationship between the agent, and the events that it triggers on the environment, is established. This brings us to the last condition.

5.3. Normativity

Finally, even if we had a well identified system and that system could be said to be the active source of its interaction with the environment, in order to call it an agent, we are still missing an extra ingredient. Someone suffering an epileptic attack is not considered to be an agent. The epileptic person is a well identifiable entity and, in addition, she is the genuine source of her interactions with the environment. But something is preventing us from calling her a genuine agent on this particular scene. When considering agency we presuppose that the interaction is not random or arbitrary itself but makes some “sense” for the agent itself. Agents present some kind of internal goal or norm according to which they are acting. Systems whose dynamics can be specified completely by applying universal laws of physics are not called agents. Norms or goals are not deducible from natural laws *alone*, they show up as (behavioural) regularities that have some sense of *ought-to-be* in themselves in addition to being in place or just happening as they do. A sense of normativity unfolds: the norm *must* be followed, not achieving the goal becomes a *failure*. Planets cannot fail to “follow” the laws of nature, while agents regulate their interactions, which are subject to failure and success according to some (implicit or explicit) norm. This I shall call the **normativity condition**. Normativity challenges physicalist scientific approaches to the understanding of our world because it introduces a value asymmetry (good/bad, true/false, adapted/maladapted) in the description of nature; an asymmetry that is not present in any of the fundamental laws of physics or the models that typically compose physical and chemical sciences. But, although alien to fundamental physics, normativity is an essential component of agency. It may not appear as an all-or-nothing property but rather as a continuum, as a degree of improvement, of increasing/decreasing adequacy, of gradual functional achievement, etc. Agents act so that they somehow define their own well or bad acting, valuing and evaluating their interactions; over and above how things (interactions), in fact, occur. A naturalist and minimalist account of agency requires that the agent, in a sense or another, define its own normativity. Otherwise, as in the case of the identity condition above, we are condemned to an infinite regress. We, as observers, can make judgements about the “adequacy” of an agent’s behaviour in relation to some norm, standard or goal (be it epistemic, artistic, ethical, functional or otherwise). But if we are to adopt a naturalist approach we must, somehow, be able to justify this normativity on the very “nature” of the agent.

All these characteristics appear somehow related to internal properties. That the system be defined by itself, that the system be active or that it be normatively regulating its interactions, requires that we look inside, that we somehow give reason of these features in terms of the way in which the system is organized and organizes its interactions with the environment. They all seem to depend on a particular internal organization, or arrangement between its constituent components, that provides it with these interrelated properties defining the system as an agent. Now, even the most sceptical externalist will agree that it is worth trying to find whether there is any kind of organization of matter able to provide the minimal and most fundamental model for agency. Our goal will be to pick up this basic organization as a departure point so as to move, progressively, towards more complex and cognitive-like forms of agency. On the journey we shall stop and analyse, in more detail, those three conditions that specify agency (i.e. identity, causal asymmetry and normativity) making explicit what does exactly mean that there is *a system doing something by itself according to certain goal or norm within a specific environment*.

As we travel from the bottom-up we shall see that increasingly complex forms of agency (specially those leading to cognition) will require the introduction of additional components or processes into our minimalist model. Those additional components might alter substantially the organization that supports agency and its relationship with the environment. As a result of these changes a host of relevant properties (adaptivity, teleology, situatedness, embodiment, behaviour, etc.) will start to unfold.

6. A MORPHOPHYLOGENETIC APPROACH: THE ROUTE FOR RECONSTRUCTION

We are now in place to depart, the epistemic machinery in gear and a good template (that of agency) ready to be reconstructed towards minds; we only need a preliminary route for our reconstruction. Big part of the inspiration for a biologically grounded approach to agency, cognition and mind comes from evolution. If there was no mind at the beginning of the universe, minds must have appeared as a result of a long evolutionary process, somewhere between the origins of life and the birth of hominids. If nature has produced minds departing from a Big-Bang through the emergence of life to the appearance of mammals, we can most probably greatly benefit from our knowledge of evolution to use it as a scaffolding for our conceptual reconstruction.

But not any evolutionary story would do. I will call *morphophylogeny* the evolutionary scaffolding that permits us to build conceptually distinct modes of organization from the bottom-up. A morphophylogenetic approach may

permit a gradual satisfaction of the MUN epistemological constraints reconstructing the phylogenetic transitions that nature had to traverse towards minds. Strictly speaking a morphophylogeny is not a morphology nor a phylogeny *per se*. First, as a morphology the interest is not put on specific molecular mechanisms that could lead to cognitive agency or specific environmental factors that could have forced the appearance of cognitive capacities. On the contrary, a morphophylogenetic approach has a strong commitment with an Aristotelian sense of form as organization, with the morpho-logic of what kinds of organismic forms require and are capable to display increasingly mindful capacities. Second, a morphophylogenetic approach is not a precise phylogeny on that it is not a specific evolutionary path that we are interested on but rather the set of bottlenecks and transitions that lead from simpler to more sophisticated forms of living agency (relatively independently of some lineages and specific adaptive strategies). It could easily be the case that types of bacteria that evolved after the formation of eukariotes present more sophisticated forms of organization (in agential terms) than that shown by eukaryotic species, or that early eukaryotes show more complex forms than what some plants can achieve. What remains important is not the temporal sequence but the set of limitations and opportunities that different (functional and dynamic) forms of organization afford for the genesis and complexification of agency.

The morphophylogenetic scaffolding will largely be composed of complex mechanistic models of biological organization, distilled on their minimal form, often accompanied by simulation models and illustrated by model-organisms that provide an empirical benchmark test and a source of generalization. The dialogue between mechanisms and organization, evolutionary bottlenecks and transitions will be critical to uncover necessary and sufficient conditions, lower level constraints and higher level openness to arrive to our final destination. A morphophylogeny of agency (and the progressive construction of increasingly complex models of agency) is the object of the next three chapters, Part II, of this work.

PART II

The Morphophylogeny of Agency

The strength of conceptions does not, therefore, depend on their degree of truth, but on their antiquity, their embodiment, their character as conditions of life.

FRIEDRICH NIETZSCHE

Chapter 4: Life: minimal agency in basic autonomous systems

The main concept which, I consider, underlies every observation of a living being, and from which we should never move away, is that it be autonomous in itself, that its parts hold a necessary relationship, that nothing be mechanic, so to speak built and produced from outside, even though the parts act towards the outside and are affected from outside.

JOHANN WOLFGANG GOETHE¹

I shall now attempt to render explicit some conceptually relevant aspects of the appearance the most minimal form of agency. Starting from the domains of physics and chemistry, in which no object can be predicated as the locus of actions, I will try to reconstruct the emergence self-generated identity or individuality, norms and environments (worlds) characteristic of agents. The question is close to that of the origins of living organization and I will take advantage of some recent advances in this field.

The fast growing interest on the minimal and artificial forms of life (under the label, among others, of ‘synthetic protocell biology’—Solé et al. 2007) provides valuable models to work with. In addition, this new scientific discipline merges with some traditions in philosophy of biology, complex systems and cognitive science that have long elaborated the relationship between minimal forms of living organization, agency and cognition. Among those we find the autopoietic theory of life and cognition (Maturana and Varela 1973/1980, Varela 1979, Weber and Varela 2002, Thompson 2007, Stewart 1996, Bourgine and Stewart 2004, Luisi 2003) and some closely related developments on the field of autonomous systems (Kauffman 2000, Ruiz-Mirazo 2001, Etxeberria, Moreno & Umerez 2000, Barandiaran & Ruiz-Mirazo 2008, Ruiz-Mirazo & Moreno 2000, Ruiz-Mirazo & Moreno 2004, Moreno, Etxeberria & Umerez

¹ Translated from Goethe (2002:102).

2008, Collier 1999, Collier & Hooker 1999, Christensen & Bickhard 2002, Bickhard 1993, Di Paolo 2005), together with some other philosophical trends such as Jonas' philosophical biology (Jonas 1966/2001, 1968), von Uexküll's theory of meaning (von Uexküll 1940/1982) followed by the biosemiotic school (Hoffmeyer 1996) and some aspects of evolutionary theory and its origins (Dawkins 1976/2006, Pross 2004, Maynard-Smith & Szathmáry 1995). Although still far from a full-fledged scientific formulation, from a conceptual point of view, much of the work contained in this chapter has already been carried out, although in a fragmentary way, by some of the authors mentioned above (particularly by Kauffman 2000, Ruiz-Mirazo 2001, and Ruiz-Mirazo & Moreno 2000, 2004).

Living systems cut out themselves from a background of physical or chemical processes. They have the capacity to generate and re-generate themselves, to be the locus of a perspective in relation to their environment, to increase their complexity through and open evolutionary process of evolution and, undoubtedly, living systems can be considered genuine sources of agency. These features of life have attracted the attention of scientific research for centuries and cellular and molecular biology can nowadays offer a picture of the mechanisms and processes that underlie such fascinating properties. Throughout this chapter we shall reconstruct a model that is capable to generate them on their minimal form.

The structure of such reconstruction will be as follows. First, I explore some physical requirements for systems able to support agential processes and introduce the notion of self-generated identity (individuality) by way of the notion of self-organization. Second, in order to overcome the limits of the simplest forms of self-organization, I introduce component production chemical networks (metabolism) and discuss the alternative of replicator molecules as the foundational field for a theory of agency. Third, basic autonomous systems will be characterized and a simulation model proposed to illustrate its minimal form of organization. Next, I expand on how the conditions of individuality, normative functionality and causal asymmetry that are already present in this model. Finally, the notion of a minimal agent will be reconstructed analysing how it meets the epistemological MUN constraints stated in previous chapters.

1. TYPES OF COHESION

A solid departure point to ground the full project of a morphophylogenetic reconstruction of the concept of agency is the *persistence* of a system. Only relatively stable, enduring, entities will be able to support agency. Surely, this condition is not exclusive for agency, for it extends to almost any kind of en-

tity that we can rely on as part of our world, yet, however obvious, this departure point already helps to define a number of key notions.

The crucial feature for persistence is precisely that of cohesion. We can start defining cohesion as the property of a system by which it achieves unity in spite of internal or external fluctuations (Collier 1988, 2004). A system must have a degree of cohesion in order to become stable and a locus of individuality and agency. A quick look at the kind of processes that surround us shows that the universe has evolved producing forms of order in some places (like rocks or galaxies), while, in others, matter presents no cohesion at all (such is the case of gases, for instance). The ordered matter takes in turn two different forms: in some cases, basic components appear lumped together forming *conservative structures* and in others, they constitute *dissipative structures*.

The first type (conservative structures) refers to spatially ordered forms of assemblage of material sub-units, where order is temporally instantaneous, like in rocks or some crystals or temporally unfolded, like in atoms or planetary systems. In both cases the cohesive form exhibited is just an expression of gross physical forces acting between components that, interacting under certain conditions, fall into a stable dynamic or structural stability. These systems will stay as they are indefinitely once created: i.e. they are energetically conservative or quasi-conservative. Either chemical bonds (in molecules) or gravitational forces (in planetary systems) lump parts together. The cohesion of such systems will only be destroyed if internal or external fluctuations/perturbations can counteract these forces. However, and precisely because of their form of cohesion, conservative systems are not good candidates to support agency since: 1) they are not capable of generating variable internal states by themselves; if left alone they tend to maximal entropy (i.e. maximum level of disorder) and all their ordered complexity has been externally pre-defined (in terms of initial and boundary conditions, not as a result of the activity of the system), and 2) these systems can perform no work (which, as will be argued later, becomes central to characterize agency)².

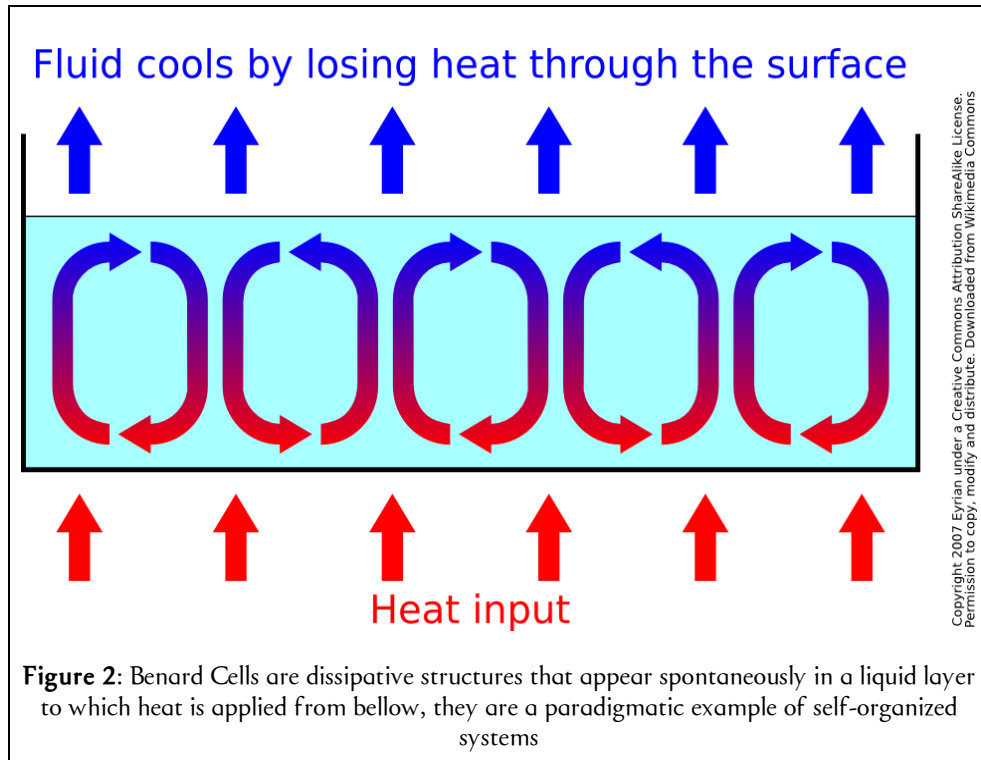
The other form of stability is that shown by dissipative structures (Nicolis & Prigogine 1977). It appears in far-from-equilibrium thermodynamic conditions (FEE hereafter). In a dissipative structure a set of interacting elements generate a cohesive dynamical pattern under an energy gradient in FFE conditions (i.e. far from the state of maximum disorder to which the second law of

2 The case of human made machines, such as robots, that present an organized conservative order capable of channelling energy to produce changes on their environments is a particular case that will receive the attention it deserves at the end of this work. By now it shall be enough to note that these systems require other agents (humans) for their design and continued existence, which shall prevent us from taking them as a departure point. Note also, following Collier and Hooker (1998) on this topic, that such conservative systems are fully determined by externally imposed constraints on their design and thus escape what can be said to be a genuine source of autonomous agency.

thermodynamics brings the system in the absence of the continuous flow of matter and energy). Examples of this type of system are whirls, hurricanes, Benard cells or lasers. In all these different systems a huge amount of microscopic elements adopt a global, macroscopically ordered pattern in the presence of a specific flow of matter and energy. Interestingly, their internal dynamic cohesion is not only a consequence of the material features of their components but also, and most importantly, of a process of circular causality (Haken 1977). The resulting macroscopic pattern itself contributes to maintain the dynamical cohesion at the microscopic level: by dissipating energy, the pattern contributes to its own stability and cohesion. Thus, these systems are able to generate and maintain, through recursive dynamics, a new way of correlation among their constitutive elements that otherwise would remain disconnected.

2. DISSIPATIVE ORDER AS SELF-ORGANIZATION

A dissipative structure is often called a self-organizing system. Yet, the term self-organization has been controversial since its very origins in cybernetics (Ashby 1962) and it is important to note that models and definitions of self-organization are generally domain specific (collective behaviour, chemistry, Artificial Life, biology, ...). In our case we are interested on the domain that deals with the lower boundary of physical sciences and we shall adopt, for our purpose, a preliminary definition that is due to Ruiz-Mirazo: “By self-organization we mean a phenomenon by which local non-linear interactions between elementary units generate a global behaviour (e.g., a spatial/temporal pattern) that is maintained through a certain number of constraints, among which—at least—one is a product of that very phenomenon” (Ruiz-Mirazo 2001). A paradigmatic example of this type of self-organization are Bernad’s convection cells (see figure 2). These cells appear spontaneously in a liquid layer to which heat is applied from bellow. The initial state of the layer is that of equilibrium and a thermal conduction that extends from the hot area (bottom of the layer) to the top. As the heat increases, however, the uniformity of the fluid breaks, small fluctuations are amplified and convection cells start to form. These cells remain stable and dissipate heat more quickly than thermal conduction does in the equilibrium phase.



Note that the global pattern (the convection cell) is not instructed (dynamically specified) from the outside, nor can it be reduced to or predicted from the activity of any of its local components (molecules) alone. As Collier notes (2004), for self-organization to occur three conditions must be met: a) an energy gradient across the boundaries of the system must be present so that internal order can be generated without violating the second law of thermodynamics (this is an external boundary condition)³, b) components must interact recursively and non-linearly so that correlations between components can be established (this is an internal microscopic or local property); and c) these local interactions need to be able to create global attractors, defined by the critical points that separates different dynamic phases of the system (internal macro-global property). Under an appropriate combination of these three conditions a self-organizing pattern can emerge spontaneously: local fluctuations of interactions between components are amplified to achieve the global

³ Collier develops a more general notion of self-organization that can be applied across physical and formal or computational process thus the gradient is an entropy gradient in general and an available energy gradient in the case of physical self-organization. However at this point we are only dealing with physical self-organization.

configuration that minimizes local entropy production and energy dissipation (Nicolis & Prigogine 1977).

When such global patterns arise, the introduction of the term “self” can be considered appropriate and justifiable from a naturalist point of view. Two complementary factors determine this *selfness*: a) that the collective pattern is not instructed or configured from outside and b) that it is the result of “internal” activity. As for the first condition Collier concludes: “since the external gradient needs contain little organisation or information of other forms except entropy/exergy, which is statistical in nature, and undifferentiated relative to system organisation in self-organising cases, *the process is not externally modulated*”. As for the second aspect, it is the recursivity or circularity (of a pattern that constitutes the condition of its own stability) what provides a minimal form of self-created **individuality**. If, under certain boundary conditions (without any organizational or informational specificity) we observe a dynamic order emerging spontaneously within a very homogeneous system, we may say that, at least in a very primitive sense, there is here a form of stability that “actively” maintains its own cohesion through the interaction between its component parts.

In addition, some of these systems have causal powers not present before the self-organized process appeared. Think, for instance, on the formation of a tornado where, although its boundaries might not be fully distinguishable, there is a well identifiable phenomenon with a genuine causal asymmetry in terms of the production of changes on its environment (e.g. its destructive capacity). Benard convection cells have also dynamic properties which the system rests on its previous homogeneous thermal conduction phase. But surely, we do not want to call the tornado and alike full-fledged agents yet. However, FFE self-organized structures provide a first sense of self-maintained identity, a rudimentary form of a recursive *self* that results from the cohesive emergence of a higher order dissipative pattern.

Yet, purely physical self-organized systems of this kind are too simple to produce any interesting form of interactive process with their environments, or to keep themselves going without externally predefined and controlled boundary conditions. For instance, when the weather conditions change (temperature, pressure, etc.) a tornado disappears and there is nothing it can *do* in order to “survive”. The same holds for Benard cells, they vanish when the temperature gradient stops. Thus, in order to search for the origins of agency, we have to look for forms of self-organization able to *actively controls* some of their boundary conditions and thus begins to *act*.

3. TOWARDS ORGANIZED COMPLEXITY

Among the wide set of self-organized systems, those based on chemical processes are of particular interest, because they allow *the construction of complex recurrent organizations through the creation of local and selective constraints*. This is a crucial point and deserves a more detailed analysis. In the physical medium the process of self-organization is achieved by a huge number of microscopic components, where none of them contributes *specifically* to the formation of the macroscopic pattern. The emergent form of order is the result of undifferentiated and *stochastic* contribution of the components. For instance, in the case of Bénard cells there is no specific mode of contribution to the macroscopic pattern that results from *types* of molecular components. At most, the density of the water might affect the formation of the convection cells, but density is again a stochastic macroscopic property.

Although generally considered *self-organized* it is not easy to find a proper organization in these physical dissipative structures. The concept of organization usually refers to an integrated disposition of parts or processes within a system, each contributing differentially to different functions. In the most simple instances of dissipative structures there is only one part (the global or macroscopic pattern) and only one function (that of constitutive self-maintenance) and the concept of organization loses its meaning⁴. It is in self-organized chemical component production networks where this uniformity is broken and parts (reactions) can be distinguished and may come to contribute differentially to the maintenance of the whole. This differential contribution leads to a dissipative organization with different levels of kinetic and thermodynamic constraints.

3.1. Metabolism: chemical component production networks

Unlike the most generic physical medium, the chemical one permits the combination of both dissipative dynamic order and conservative structural order, with crucial consequences. On a nested set of chemical reactions the shape and combinatorial properties of the molecules (properties that belong to the conservative order) can contribute differentially to the self-maintenance of the global pattern of a network of reactions (dissipative order). Although chemical reactions are still stochastic processes they show a number of interesting properties:

1. The shape of their components (the molecules) determines their reactive capacities; they can thus generate different *types* of reactions that

⁴ Even if interpreted that each molecule is a part of the system the problem remains on that not specific individual contribution exists, the global pattern is the only "function" and components contribute in an undifferentiated way to it.

may contribute *differentially* to the macroscopic self-maintenance of a network of reactions.

2. The molecules can change or be created combinatorially thus permitting the *creation of new types* of reactions.
3. Since some molecules possess catalytic properties (due to the conservative order expressed on their shape and combinatorial properties) reactions might appear nested in *feed-back loops* affecting their reaction rates mutually
4. As a result, molecules that may be randomly created through stochastic collision can be *selectively retained* if they catalyse, or participate in, those reactions that produced them. In other words, if a new molecule enters the a network of chemical reactions (either from outside or newly created as a result of molecular collisions) its concentration will increase if it participates on the network by catalysing those reactions that lead to its production (i.e. if it generates a positive feed-back loop of reactions leading to its production).

These four properties that arise in a networked set of chemical reactions can potentially lead to increasingly complex FFE organizations whose stability is defined by a set of environmental conditions and concentrations internal to the network.

As Prigogine and Stengers (1984) have emphasized, in non-linear chemical reactions (those that appear under catalytic loops, which are ubiquitous in biological systems) microscopic fluctuations may give rise to highly specific behaviours of the system and permit to generate more complex dissipative structures than those found along the “purely physical” domain. Not only is temporal stability broken, leading to cyclic or strange attractors (defining an intrinsic rhythm), but also spatial homogeneity is broken, leading to spatial structuring. Thus the system determines, intrinsically, its own spacial structure. These systems escape from what Prigogine and Stengers have called Boltzmann's principle of order (i.e. that the behaviour of a macroscopic system is equal to the mean behaviour of its constituents). The system appears organized and we can start speaking in terms of local specific states that propagate across the full systems giving rise to differentiated global states. Mesoscopic relationships appear in which micro-macro relationships are not just many-to-one (many components interact to produce a single macroscopic pattern), but many-to-many (differentiated parts generate differentiated global states). We are thus closer to systems capable of having self-defined internal states while collectively generating a global cohesion that may ultimately be expressed in terms of agency

The idea of a network of reactions that, altogether, maintains a dissipative order through the production of part of the components of the very network is considered by some at the heart of the origin of life and living organization

(Maturana and Varela 1980, Kauffman 1993, Dyson 1982, 1999, Morowitz 1992, 1999), often grouped under the label “metabolism first school”. They keep, however, an ongoing debate with the so called “replication first school” regarding the origins of life and, most importantly for our task, the evolution of biological complexity at its origins. A debate that deserves some attention given the consequences it bears for the foundations of agency. So, before we proceed we shall consider this alternative departure point for our reconstruction of minimal agency

3.2. Replication: template molecules

Many researchers, specially motivated by the rise of molecular biology and the modern synthesis in evolutionary biology, have explored the possibility of taking chemical “replicators” as a point of departure to understand living systems (including their agential capacities—Dennett 1987, 1995). Dawkins’ *The selfish gene* stands along the most influential foundations of this approach. “What does matter is that suddenly a new kind of ‘stability’ came into the world” (Dawkins 1976/2006: 16). Dawkins refers here to the “*replicator*”. Note that replicators imply a type of stability very different to the one previously explored which already includes an agential capacity, that of replicating. Replicators or, more precisely, modular template molecules (if we are to avoid any agential attribution that is yet to be explained) amount to a rather abstract form of stability. It is not the conservative stability provided by the atomic bonds holding components together within the molecular structure, but the abstract sequence defining the molecular species. This form of stability occurs when the chemical medium and the template’s molecular properties (affinity to similar or complementary components) facilitate the proliferation of the same sequence within a chemical pool. To take this possibility into account is of fundamental importance since it involves an explicit reduction of what might be called organismic or metabolic agency to replicator agency. This reductive project is explicit when Dawkins claims that “we are machines created by our genes”, where the minimal form of agency is the gene, or the replicator, capable of creating and controlling derivative forms of agency (through the expression of phenotypic “envelopes for genes”):

Other replicators perhaps discovered how to protect themselves, either chemically, or by building a physical wall of protein around themselves. This may have been how the first living cells appeared. Replicators began not merely to exist, but to construct for themselves containers, vehicles for their continued existence. The replicators that survived were the ones that built survival machines for themselves to live in. (Dawkins 1976/2006: 19)

According to this view, “replicator” molecules epitomize the origins of agency, and their differential “replication” defines the ultimate logic of living agents. Dawkins is definitely not the only proponent of such view. Ady Pross

has recently summarized the “replication first” school under the following terms:

We argue that life’s emergence must have begun with chemical replication, (...). Thus despite life’s extraordinary complexity, this complexity is not inherent, but stems from an evolutionary process that involves continual complexification—a fundamental and ongoing process associated with replicative chemistry. The complexification process from single-cell entities to multi-cell ones is now established evolutionary ideology, but we believe, in line with the ‘replication first’ school of thought, that the evolutionary process of complexification extends all the way down to a primal replicating entity. Complexity in its various facets—structural, informational, and metabolic—are all adaptations that kinetic selection (the equivalent of natural selection operating at the molecular level) introduced to further the replicating ‘agenda’. (...) Metabolism is the particular adaptation that was employed by kinetic selection to keep the thermodynamic tiger at bay, and as such must have been preceded by replication; kinetic selection can only operate once a replication-mutation cycle is underway. Thus we believe that life began with a process of molecular replication and that activity has remained central to all its subsequent evolutionary explorations through ‘complexity space’. (Pross 2004: 318—319).

There are two main objections to this alternative. The first one is based on experimental grounds, the second conceptual, and both are intimately related. On the experimental side, the numerous experiments *in vitro* that have tried to reproduce the origins of life based on “replicator” template molecules have failed to generate complex chemo-dynamic (metabolic) organizations. Instead of producing increasingly complexer systems (as those observed in the biosphere and for whose explanation replication is called into scene) the tendency has rather been that of maximal reduction of the sequence of components of the replicator unit: a minimal form that maximizes the speed and copying accuracy of the molecule.

During the 60s Sol Spiegelman carried out a number of well known experiments demonstrating that RNA molecules were capable of replication in the absence of metabolism (Spiegelman 1967, 1971). Spiegelman and colleagues introduced into a pipet a template RNA from the virus Q_β together with Q_β replicase and various monomers. The mixture produced an exponential increase of the RNA template. In addition, the resulting RNA was different than the original one and repeated experiments led to the same final result. For the first time, an *in vitro* process of RNA template evolution was observed; giving rise (under the specified experimental conditions) to an “evolved” form of template RNA that came to be known as *Spiegelman's monster*. Not only did the name “monster” come to satisfy the expectation of an exciting *de novo* synthesis of living creatures but also the accompanying delusion of such expectations by a newly defeated creator. After 74 generations the experiment gets stuck at the same stage: from the 4500 bases that composed the original RNA template of the Q_β virus the “evolutionary result” was always the same: a

replicating “monster” of just 220 nucleotide bases. Eigen (Eigen et. al. 1981) repeated the experiments in 1974 and he even achieved to synthesize RNA *de novo*. But the results of the “replicator evolutionary processes” were always a minimalist string of 150 to 250 nucleotides. Some kind of “evolution” had in fact occurred, but in the opposite direction to that required for agency: the complexity of the molecule and the reactions involved decreased.

Interestingly, the proliferation of increasingly complex molecules, the necessary step to overcome the “monstrous” result described above, requires its embeddedness in a chemical reaction network (Orgel 2000)⁵. When Pross states that molecular replication remains “central to all its subsequent evolutionary explorations through *complexity space*” he seems to be somehow assuming that “mere replication” has, in itself, a “complexity space” to be explored. However “mere replication” can only adapt to faster, more reliable and more stable template molecules. In other words, speed, accuracy of the copy and stability of the molecule exhaust the “functional space” of “replicators” and consequently the “complexity space” that the replicator can traverse. At most, some forms of replicative parasitic functions might be envisioned (as it occurs in some Artificial Life models such as Tom Ray’s well known *Tierra*—Ray 1994), but nothing like agency can be derived solely from the logic of replication; at least in the absence of an “internal environment” in which the template molecule operates producing, activating or controlling different *functions*. A set of energetic and chemical processes must be in place in order for the replicating molecule to achieve higher complexity structures, even for its mere replication! In fact, Spiegelman’s experiments required the introduction of a highly complex organic molecule, the Q_β replicase, for the RNA to replicate. So, before any kind of “replication” process could be in place some form of proto-metabolic network needs to be assumed. Thus, the structural complexity of replicated molecules (the combinatorial complexity of the components of the molecule) relies deeply on the organizational complexity of the network of processes of which the “replicator” is part. Therefore, even if we were to take template molecules as important components of agential organization and its evolution (whose necessity is still to be shown), that could only happen in the context of reaction networks providing a “phenotypic complexity space” to be explored (as noted by Wicken 1987 and extensively argued by Ruiz-Mirazo, Umeretz and Moreno 2008). And this leads us to the conceptual argument against the primacy of replication for the project of grounding agency.

On the conceptual side, attribution of agency to these template molecules is a fallacy. The origin of the fallacy lies on the very notion of replicator. To say that something is a replicator (that it replicates *itself*) entails that it is the active source of the replicating process:

⁵ Quoted in Fernando & Rowe (2008).

At some point a particularly remarkable molecule was formed by accident. We will call it the Replicator. It may not necessarily have been the biggest or the most complex molecule around, but *it had the extraordinary property of being able to create copies of itself*. (Dawkins 1976/2006: 15, italics added).

If one looks up more carefully into Dawkins' hypothesis regarding the early capacity of "creating copies of itself", its lack of *selfness* is made apparent. Dawkins asks us to imagine a string of molecules floating on a soup of component molecules so that each component of the string had an affinity with molecules of its same type. As a consequence, if the appropriate conditions in its molecular medium are met, copies of the molecule will be formed. In such a case the expression "it replicates" might be used, but the linguistic ambiguity is here the source of the conceptual mistake: "it" cannot refer to the template molecule but to the chemical pool containing it, for there is nothing that the molecule *does* other than to be located (in virtue of external forces) surrounded by the appropriate molecular species.

Replicators cannot be said to be agents for the simple reason that they have no individuality, *they can only be said to act or have effects statistically as a molecular species within a wider context of chemical components*. There is nothing that the "replicating" molecule does as such. The reactions in which it is involved as a molecule are the result of fluctuations and Brownian motion and only facilitated by molecular forces and their spatial configuration (which is what molecular affinity involves). To say that a molecule is auto-catalytic and contributes to its reproduction is a *contextual and species predicate* not an individual one. On the contrary, in chemical component production networks to say that a system is self-maintaining and self-(re)-producing is an individual predicate, a predicate of the network as a unity (as will be made explicit shortly). One could argue that, since the network of reactions is ultimately constituted by molecules, it is inevitable to reduce this predicate to some kind of species predicate as well. That is not the case. A metabolic network is not a compound of molecular species but a specific distribution of concentrations and processes (reactions) between them, namely, an organization. The argument can be summarized under the slogan "copying does not suffice for copying" (in the sense required to ground agency). In other words, only certain kinds of self-maintaining dissipative organizations can be the locus of agency; replicative conservative-structures cannot (however structurally complex they turn out to be under certain contextual and historical conditions). And the minimal model of a self-maintained dissipative organization is a component production chemical network (a proto-metabolic system). Achieving its individuality as a self-maintaining network the system is organized into processes (reactions) that play specific thermodynamic and kinetic functions within the network.

The "replication first" approach has received numerous other critiques (Shapiro 2000), and the "metabolism first" school is not without its problems;

among them, the request for empirical evidence and consistency given primitive earth conditions (Anet 2004). The issue for us is not, however, to elucidate which was first, replication or (proto-)metabolism, but which one is capable of providing a solid foundation for agency; i.e. which logic, that of replication or that of metabolism, may satisfy (albeit in a minimal form) the three conditions for agency previously stated. As we have seen, the replicator logic cannot accommodate such requirements on its own, while the metabolic logic of the dissipative order found among chemical production networks is better suited to do it. Whether these types of networks require the presence of replicators as necessary predecessors remains an open issue together with the question of the conditions under which life began in the primitive Earth. But I hope to have shown (and it will become increasingly manifest as we proceed) how metabolic or proto-metabolic organization⁶ (with or without the presence of replicators) implies a logic of self-maintenance and process organization that cannot be reduced to molecular template replication processes alone, while it may provide a solid foundation for agency.

3.3. The early exploration of the complexity space

Now... what set of additional requirements need component production networks meet in to achieve agency? And what are the chances that these chemical networks increase their complexity to satisfy those demands? Following Fernando and Rowe (2007, 2008) we can envision a prebiotic (pre-genetic code) evolutionary scenario⁷. In order to reach this scenario we need the auto-catalytic system to become a cohesive unity and acquire some means of reproduction (understood on its most basic form as division with conservation of self-maintaining capacities). The earliest form of such a unity might be given by the reactants of the auto-catalytic set becoming hydrophobic and forming a spot (any other kind of phase separation will do). If auto-catalytic self-maintaining systems are spontaneously separated from their environments creating a unity (a spot) a rudimentary form of reproduction might be envisioned. At an early stage reproduction need not be very a sophisticated pro-

6 Metabolism, as we know it in current cells (with alternative pathways, regulatory mechanisms, spatial organization, specialized organelles, etc.) is a much more sophisticated and qualitatively complex organization than what we have here termed “metabolism” more generally. This is why the term proto-metabolism seems more adequate to label early predecessors of genuinely metabolic networks. In what follows, however, we shall refer to metabolism as the abstract organization that pictures a, more or less simplified, chemical reaction network that produces and repairs the components of the network. This abstract and holistic relationship between components of living cells can be said to be universal and that captures one of the most important and characteristic features of life (Cornish-Bowden et al. 2004).

7 This section is not meant to be an accurate development of what truly might have happened during an early prebiotic evolutionary scenario. The goal of this section is to make conceptually feasible that such scenario could in fact occur and, more importantly, to explore the its consequences that it bears for the way in which we shall think about evolution and complexity growth in what is to come.

cess; division of the spot due to external fluctuations or surface tension when growing should be enough.

At this point we can already imagine how our auto-catalytic networks may start evolving. From the molecular collisions involved on the reaction networks some will, occasionally, produce new types of molecules (or else, new molecules might enter the spot from the environment). Most of them will have no effect on the system (a single molecule may have no effect within a spot of millions of reacting molecules). Occasionally, part of these randomly created molecules may catalyse the reactions that produced the collision of which the new molecule is a result, thus increasing its concentration by direct or indirect auto-catalysis. In turn, it should be expected that in such cases the new molecule might produce a cascade of new reactions. In most cases such cascade of reactions will have devastating effects on the auto-catalytic network but... occasionally it might accelerate the auto-catalytic cycle or contribute to its robustness. Thus, random collisions between molecules provide the means for *variation* in such prebiotic evolutionary process while spot division provides reproduction. Given that when division happens internal concentrations are, more or less, homogeneously distributed within the spot, *inheritance* is given by the “transmission” of molecular species and their concentrations into newly created spots. *Selection* will occur insofar new variations contribute to the stability, increase and robustness of the spot (a similar scenario is devised by Segré et al 2000). Note, thus, that in order to trigger a rudimentary form of pre-biotic evolution, complex molecular replicators need not be introduced necessarily. The growth-rate of the proto-metabolic network and the robustness of its dissipative order already define a functional complexity space. This space, defined by how new molecular species can contribute differentially to increasingly robust and proliferating units, is already there for this mode of pre-Darwinian evolution to explore. Artificial-chemistry simulation models (like Fernando and Rowe’s—2008) have shown that this type of spots are likely to increase their organizational complexity towards proto-agential capacities (unlike in vitro and in silico models of “self”-replicating molecular templates).

Yet, a network of reactions enclosed in a phase-separated spot (with or without the presence of replicator molecules) is lacking an additional feature that shall become a necessary step toward autonomous agency: a physical border.

[The physical border or membrane is] the only way, on the one hand, (i) to assure the control of energy flow required for the robust maintenance of the network, and on the other, (ii) to solve the problem of diffusion and dilution (control of concentrations). Furthermore, if the precursor reaction network does not become self-enclosed, it will not be able to create a particular—and minimally stable—chemical micro-environment, being directly exposed to all changes taking place in the milieu. In other words, the system will not have any control over the boundary conditions that bring about its distinctive, far-from-equilibrium dy-

namics (and, thus, it will be extremely fragile). (Ruiz-Mirazo & Moreno 2004: 243)

This way, the formation of a membrane is crucial for the appearance of systems that actively produce and regulate a self-generated inside-outside dichotomy (Moreno & Barandiaran 2004) and anticipates the appearance of basic autonomous systems.

4. BASIC AUTONOMOUS SYSTEMS: CONSTRUCTIVE AND INTERACTIVE CYCLES

A fundamental next step on our trip, then, arises with the appearance of chemical self-organized auto-catalytic systems whose productive activity includes the construction of components that ensemble together creating a *selective membrane*. The formation of this membrane may well be the result of the early evolutionary process mentioned in the previous section or some replicator precursor may have also been involved. In any case, what turns out to be crucial is that this membrane encapsulates the chemical organization while dealing selectively with the environment to regulate the internal concentrations. As a result, what determines the behaviour of the system is not “just” a network of self-organized chemical reactions under a set of system-independent (externally fixed) boundary conditions (the gradient of matter and energy across the boundaries). The capacity of the system to *actively produce and control part of its boundary conditions* via the membrane becomes a crucial new feature (Ruiz-Mirazo 2001, Ruiz-Mirazo & Moreno 2004). This capacity is made possible by a qualitatively new kind of (self)organization, that overcomes the limitations of auto-catalytic or chemical reaction networks. The presence of a selective membrane, produced and regulated by the system, brings about a form of organization controlling (constraining) its boundary conditions. The relationship between the system and its boundary conditions (i.e. the concentration gradient) is internally modulated, not just by an spontaneous osmotic effect, but by recruiting energy produced along the reaction network to *perform work* (e.g. when transporting molecules against the gradient). In Ruiz-Mirazo & Moreno's words:

Therefore, there is an important difference between the typical examples of “spontaneous” dissipative structures and real autonomous systems: in the former case, the flow of energy and/or matter that keeps the system away from equilibrium is not controlled by the organization of the system (the key boundary conditions are externally established, either by the scientist in the lab or by some natural phenomenon that is not causally dependent on the self-organizing one), whereas in the latter case, the constraints that actually guide energy/matter flows from the environment through the constitutive processes of the system are endogenously created and maintained. (Ruiz-Mirazo & Moreno 2004: 238).

But the membrane, by itself, is unable to perform any *action*. The thermodynamic approach of Ruiz-Mirazo and Moreno uncovers an additional requirement: the coupling of exergonic (spontaneous) and endergonic (energy requiring) reactions, that could lead to the capacity of the system to perform work; i.e. the system must be able to channel energy in order to run against its “natural” thermodynamic drive. In addition, energy currencies (high energy-bond molecules) need to be created to assure the coupling of processes, otherwise the system complexity would collapse (energy is dispersed and cannot be selectively applied to different processes).

Once this thermodynamic physico-chemical organization is in place we can distinguish two kinds of **constitutive processes** in the system (a distinction that will become of fundamental importance on what is to come):

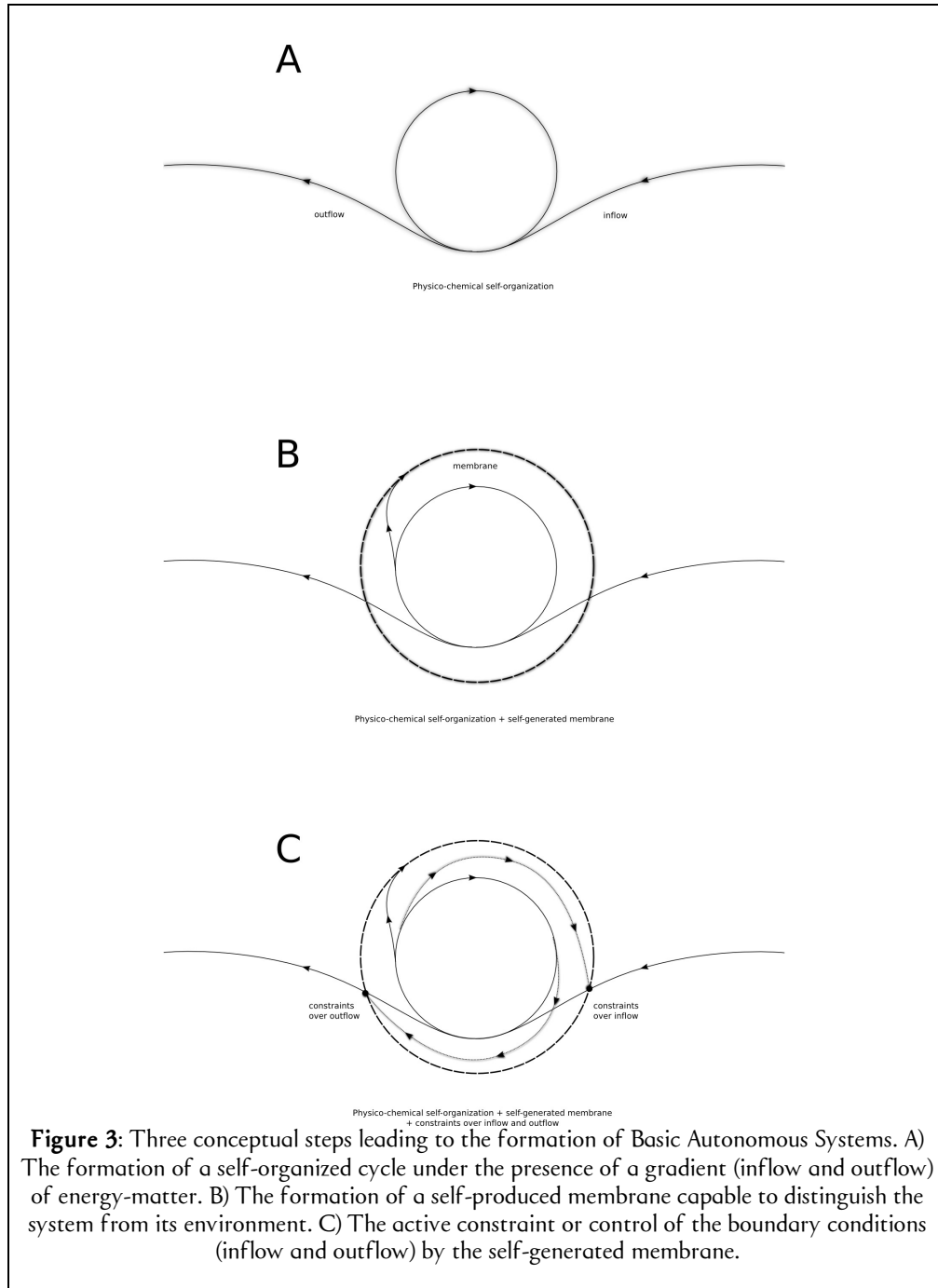
Constructive processes: constructive processes are those that participate in the continuous production of the system (e.g. chemical reactions) creating a networked organization. The **constructive cycle** is defined by the production of a set of constraints (e.g. concentration rates for certain molecular structures that will otherwise degrade) that recursively regenerate the conditions for their production. In other words, the appropriate rate of production of the components of the network is the result of the activity of the network. Because this network attains the construction and reconstruction of its components we call it constructive and we shall also use the generic term “metabolism” to refer to it. At a thermodynamic level this constructive cycle appears as a network of exergonic and endergonic couplings.

Interactive processes: Given that this network only exists as a thermodynamically dissipative organization, its maintenance requires a flow of matter and energy. Since external conditions might change or fluctuate, the system has to trigger interactive processes to sustain itself. Thus, interactive processes are those generated by the constraining action exerted by the constructive cycle to the flow of matter and energy between the system’s boundary and the environment so as to ensure the system’s maintenance. Active transport through the membrane is a characteristic example: the constructive cycle needs to channel energy to pump certain molecules out in order to maintain its organization.⁸

The constructive and interactive types of processes just described are, at this stage, ontologically interwoven so that both depend recursively and continuously on each other; instantiated in the same fundamental organization of component production systems. Both processes can, nevertheless, be conceptually separable and operationally distinguished: a) if you severely perturb the energy-matter flow (e.g. creating a very high concentration gradient), so

⁸ What is a “need” here can be equally understood as an “opportunity” to increase robustness; i.e. part of the space to be explored at the very beginning of prebiotic evolution.

that the constraining action of the membrane is not capable to compensate it,



the system will cease to exist (e.g. due to an osmotic crisis⁹); b) if you severely perturb the internal organization of the dissipative structure (e.g. by increasing or decreasing substantially the concentration of a reactant), the system will collapse. Thus, there are two kinds of perturbations that can affect this kind of systems, each of which distinguishes a different aspect or kind of processes on them. It needs to be noted that the constructive cycle has an ontological priority over the interactive one for two reasons: a) the constructive cycle *specifies* interactive functions (what *kind* of interactions are *required*, e.g. what kind of molecule and how much of it needs to be introduced or expelled) and b) the constructive cycle constructs, repairs and controls the necessary infrastructure for interactive processes to occur.

At this point we already have a minimal model of a naturalized system-environment distinction and of an *active* identity or individuality. The thermodynamically open dissipative structures that this model captures have not just an attributed identity based on a set of cohesive forces (as rocks, atoms or planetary systems might have) but a self-sustained FFE stability that is constructively *and* (inter)actively maintained. Unlike tornadoes and other physical self-organized phenomena, this new kind of system does modulate its boundary conditions (at least, in a minimal way, through the selective action of its membrane). Thus, there is a stronger sense in which we can speak of interactions properly. Tornadoes do have effects on their environments, but there is no sense in which we can speak of these effects as being controlled by the tornado itself. Component production networks endowed with a selective membrane, on the contrary, actively constrain the energetic and material flow necessary for their maintenance, so that the effect of their interactions are functional for the maintenance of their organization; *their cohesive identity is inter-actively maintained*. Figure 3 illustrates the three main aspects of minimal autonomous systems: A. a self-organized component production cycle, B. the system produces a membrane that encapsulates the network, C. the system is capable to constrain in a functional manner the inflow and outflow through the membrane.

9 An osmotic crisis happens when internal concentration is high, water comes in due to osmosis, volume grows over what the membrane surface can encapsulate and the system bursts.

What we have just described has been termed **autonomous systems**¹⁰ (Varela 1979, Kauffman 2000, Collier & Hooker 1999, Etxeberria, Moreno & Umerez 2000, Barandiaran & Ruiz-Mirazo 2008). For the sake of terminological clarification, following Ruiz-Mirazo and Moreno, we shall call this kind of self-maintained component production systems with a selective membrane, **minimal basic autonomous systems** (mBAS hereafter). Accordingly, we shall use the term **basic autonomy** (BA) to denote this mode of *material* self-construction and maintenance on top of which other forms of autonomy might appear (from neural to social, ethical or technological). **Basic autonomous systems** (BAS) will be the class of systems (minimal or more sophisticated) that instantiate a form or another of basic autonomy (independently of other types of biological functions that might accompany them: such as reproduction, immune defence, etc.).

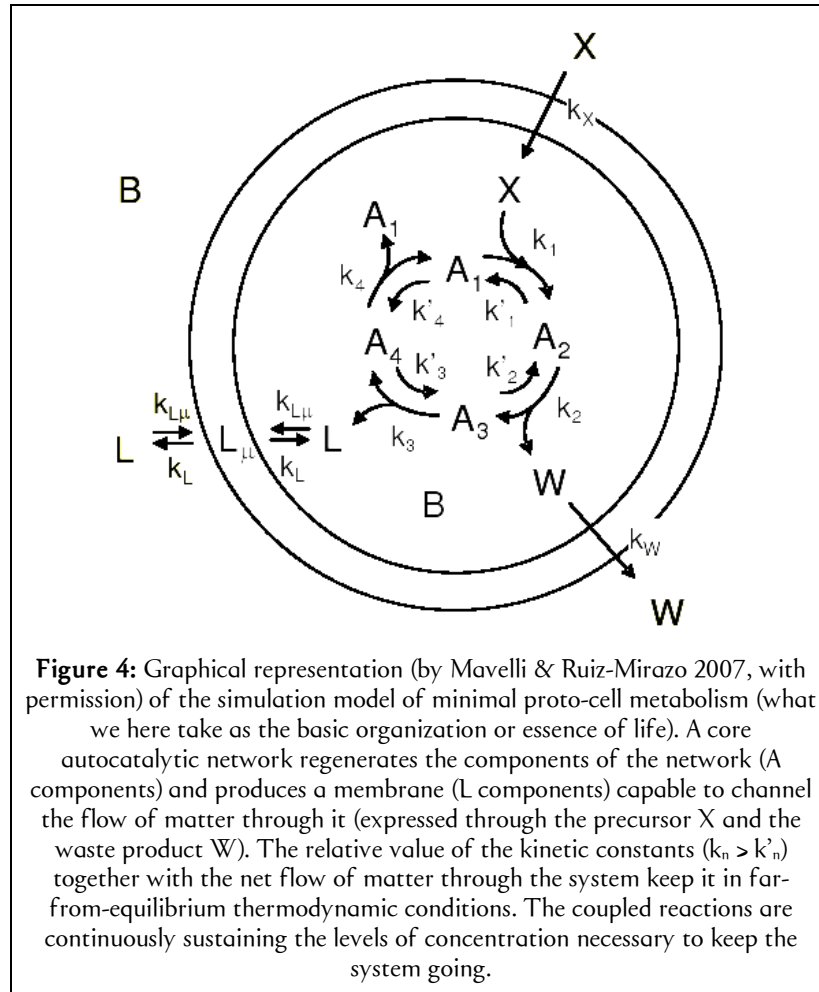
All living systems share this form of organization, the constructive cycle is realized through their metabolic organization (most of the organismic processes are directed towards continuous self-production and repair) and interactive processes take a full range of different forms (from ion-pumping through the membrane of a cell to motility in bacteria, from differential growth towards different chemical environments in plants to sophisticated cognitive strategies of predation or shopping in mammals). The working hypothesis is that cognition is to appear as a complexification of this fundamental basic interactive process and that this complexification will carry crucial transformations regarding the relationship between both, the constructive and the interactive cycles. Nevertheless we shall stop at this point to illustrate with a simulation model the features we have distinguished so far and latter extract and make explicit how the three conditions for agency (individuality normativity and causal-asymmetry) are satisfied by our minimal model of basic autonomous systems. The model will provide a first naturalist grounding for such conditions.

¹⁰ Actually, this concept of (basic) autonomy is very similar to the idea of autopoiesis developed by Maturana and Varela (1980). Our emphasis is, however, focused on the FFE and thermodynamically open nature of these systems, from which a crucial implication follows: interactive dynamics are constitutive of the system. Thus, interactions with the environment are essential for the existence and definition of the very system and not something that comes additionally to its constitution, as the original definition of Maturana and Varela leads to understand through the concept of structural coupling of autopoietic systems (understood as closed systems). Other authors, specially Jonas (1966, 1968) but also Rosen (1991) and Boden (1999), have used the generic term *metabolism* to refer to autonomy. However, we have favoured the term autonomy for two reasons: a) metabolism might be seen as constituting only one aspect of autonomy (the constructive cycle) and b) the common usage of the term metabolism (as the organic infrastructure in charge of generating energy for living processes) is not semantically as rich as that of autonomy with its implications of self generated normativity.

5. CASE STUDY: CHARACTERIZATION OF MINIMAL AUTONOMY THROUGH A SIMULATION MODEL

Of particular interest to us, in order to achieve a full naturalistic approach to the issue of minimal autonomous agency, is the set of models of *minimal living organization* that disciplines like Artificial Life, Systems Biology or, more recently, Synthetic Protocell Biology have started to develop, taking the cell as the basic unit and expression of life. Some of the early formulations of minimal models of life can be traced back to Maturana and Varela's autopoietic theory of life (1972/1980), Tibor Ganti's chemoton model (1971, 1984/1991, 2002), Stuart Kauffman's autocatalytic network theory (1971), or Robert Rosen's M-R systems (1958, see Rosen 1991 for an overview). The original formulation of these models was done under a conceptual or linguistic form, often accompanied by diagrammatic illustrations and/or by formalized descriptions. But since the early 70's, computer simulation models were used to illustrate the emergent order of the proposed organization, like in the case of Varela, Maturana and Uribe's pioneering work (1974). Also Kauffman explored the autocatalytic systems and other self-organizing biological processes making use of computers (1986, 1993) while Tibor Ganti's model's first simulation dates 1975 (Békés, 1975) followed by Csendes (1984). These simulation models played the crucial role of illustrating and conceptualizing the emergent patterns that result from the numerical integration of their mathematical models, often leading to unpredictable features that served as heuristics for further develop the models. As we saw in the previous chapter the need for simulation tools to explore the proposed models is highly tight with the non-linear nature of the reactions involved; a simulation is required to unfold the circular causality, the nested feed-back loop architecture, that is characteristic of such systems¹¹.

11 Rosen's approach is an exception to this rule of using simulation models due to his conclusion that the complexity living systems, and particularly metabolic circularity (which he shows to be closed to efficient causation—only material causation is necessary to keep the systemic dynamic order going) is not computable. Chemero and Turvey have applied hyperset theory to show how Rosen's model falls under non-well-founded logic which does not, however, make such system necessarily non-computable (Chemero & Turvey 2008).



Among the most recent approaches, we shall focus on Mavelli and Ruiz-Mirazo's simulation model of a minimal self-reproducing proto-cellular system (Mavelli and Ruiz-Mirazo 2007, Ruiz-Mirazo and Mavelli 2008). Their simulation model captures and integrates most of the essential features of the models of Maturana and Varela and Gánti and explicitly points towards a minimal form of autonomous agency. The simulation is based on a set of reaction equations that are simulated through a stochastic Monte-Carlo method where critical phenomena can be said to be, at least partially, emergent from the very simulation: systemic equilibrium, osmotic bursting and reproduction by fracture (physical division) coupled to chemical dynamical process. Precisely because of the closely realistic assumptions and relatively unconstrained dynamics of the simulation, it stands as a privileged model to illustrate the con-

cepts we have so far introduced. Figure 4 graphically represents the structure of their model¹².

5.1. The constructive cycle

At the nucleus of the model we have an autocatalytic cycle: this is a network of chemical reactions that reproduces the components of the network itself through a cyclic loop of metabolites (A components). The first core idea is that of self-organization at the chemical level: a huge amount of microscopic elements (the molecules) adopt a global, macroscopic pattern (a configuration of concentrations) in the presence of a specific flow of matter and energy. This flow is represented by the continuous inflow of X precursors into the system with the outflow of W waste products. A set of constraints on the equations, expressed as kinetic constants k_n , define the chemodynamic and thermodynamic relationship between reactions. Rules for diffusion/transport processes across the membrane are also included. Given the presence of precursor X the stochastic collision of A_1 molecules produces A_2 molecules that in turn produces A_3 molecules leading to A_4 which, closing the loop, generate A_1 type molecules. Given a set of initial conditions (the presence of X precursor molecules above a certain threshold) a set of macroscopic correlations appear (an interdependent set of concentrations of A types of molecules) as a result of recurrent local interactions (i.e. stochastic collisions). In turn the occurrence of this local non-linear interactions recursively depends on the macroscopic correlation: the higher the concentration of A_1 molecules the higher the probability of a collision between A_1 molecules to produce A_2 molecules; and the higher the number of collisions of A_1 molecules the higher will become its concentration (due to the circular set of reactions A_1 - A_2 - A_3 - A_4 - A_1) until this positive feed-back loop reaches a steady state. As we noted above the resulting pattern may be said to possess a form of *individuality* or *identity* in which, out of an undifferentiated chemical pool, a self-reinforcing order appears¹³.

12 The picture is inspired on the graphical representation of a simplified form of Gánti's chemoton model (what he called a "proliferating microsphere").

13 Ruiz-Mirazo, personal communication, raised to me some doubts about the interpretation of this simulation as an adequate model for genuine self-organization. The model is not built with the goal of addressing self-organized emergent properties. Other existing models like the *Bruselator* model (Nicolis & Prigogine 1977), the *Oregonator* model (Field & Noyes 1974) of the Belousov-Zhabotinsky reaction (Zaikin & Zhabotinsky 1970) or the *hypercycle* dynamics with negative interactions simulated by Boerlijst & Hogeweg (1995) are certainly better suited to study self-organization at the (bio-)chemical level. These models show the spontaneous formation of spatial and/or temporal structures (rhythmic oscillations, diffusion patterns, etc.) that are characteristically interpreted as self-organization. If there is no well distinguishable pattern that can be said to appear as a result of a distributed and networked process almost any chemical reaction could be trivially interpreted as emergent. In Mavelli and Ruiz-Mirazo's model, however, the circularity of the nested reactions and their capacity to produce and repair a self encapsulating membrane makes the model a non-trivial case. In any case, I shall opt for relaxing the interpretative framework and consider their simulation to model both self-organization at the chemical level and the production of a membrane with internally regulated selective capacities.

The very macroscopic pattern contributes to the maintenance of the dynamic cohesion at the microscopic level: the chemical cycle continuously regenerates its component processes. Thus, it is not only the local interactions that matter but the global patterns they generate: molecular properties are significant just in the context of massive stochastic collisions. In this context the effect of a particular molecule will depend on the reaction rates of other components whose concentrations are continuously maintained in far-from-equilibrium stability conditions by the network of reaction cycles that constitutes the system.

5.2. Interactive processes

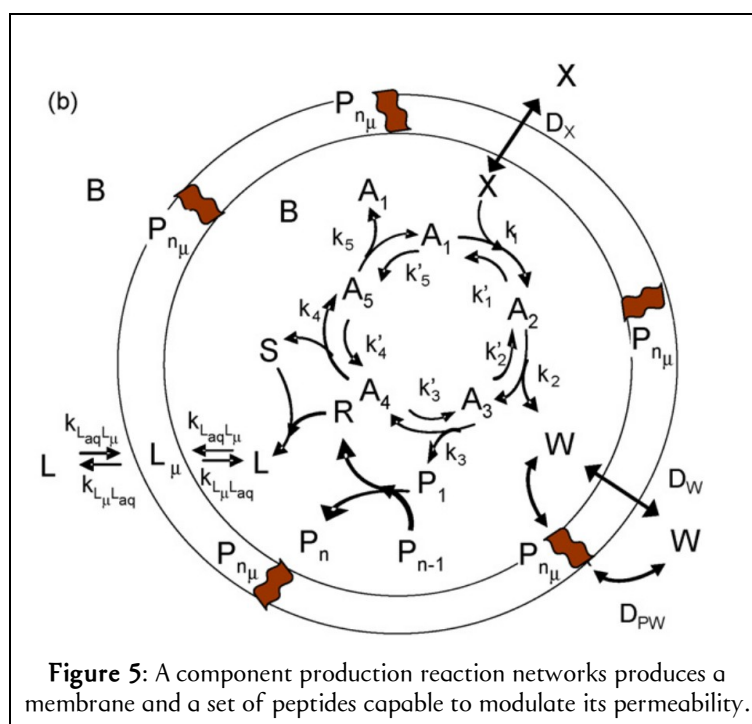
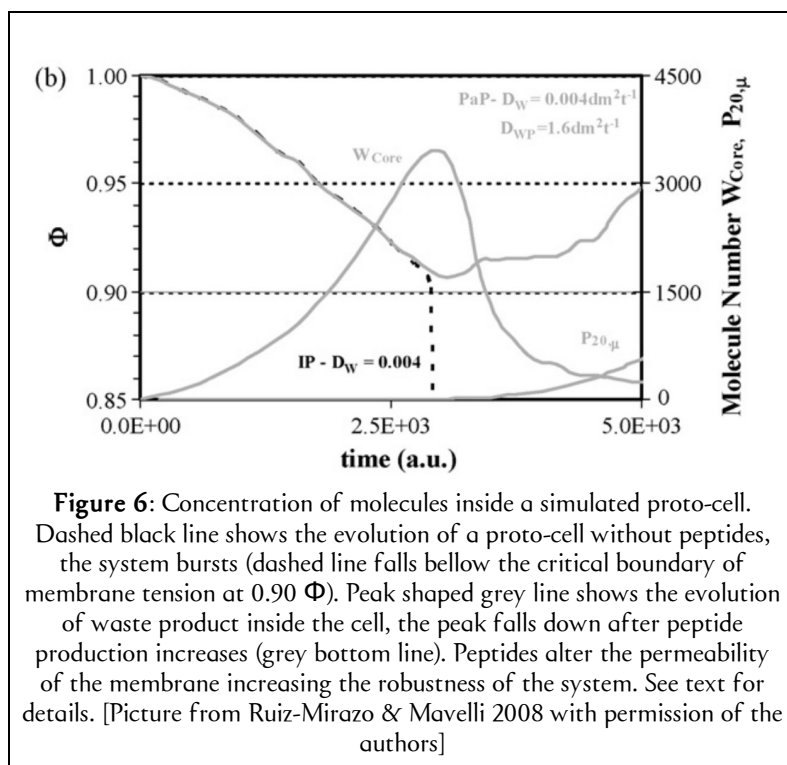


Figure 5: A component production reaction networks produces a membrane and a set of peptides capable to modulate its permeability.

The modelled system, as a self-organized and dissipative structure, needs a continuous flow of matter and/or energy in order to maintain its constitutive order. This flow is made possible by the precursor molecule X , which is not only integrated into the system as a reactant but as a high free-energy com-

This will permit us to use the model as a conceptual blender (Barandiaran & Feltre 2003) to merge the necessary and sufficient conditions for minimal autonomous agency (although the satisfaction of some of this conditions might not, strictly speaking, be completely reflected on the simulation but require a important dose of interpretation).

pound, whose energy is dissipated/distributed through the reaction network producing, among other components, a low-energy waste compound W . X diffuses into the system and W diffuses out of the system. So there is a net flow of matter-energy through the system as a consequence of the existence and permeability rules of a self-produced boundary: the membrane. This boundary requirement is represented in the model by L molecules (standing for Lipids). The network produces L molecules that ensemble with each other, becoming the L_μ molecules that, put together, produce the membrane encapsulating the reaction-network. In the absence of a membrane there would be no gradient or flow of energy as such. Some of the reactions would certainly happen but the observed FFE steady state would not be reached and the set of characteristic properties that it shows would not be achieved: a) the continuous stability of the constructive cycle, b) the regeneration of the membrane at the appropriate rate, so as to avoid osmotic crisis and burst and, c) the differential growth of the system's volume vs. membrane surface, leading to reproduction. Thus, the membrane, produced by the constructive reaction network encapsulates the system allowing both its cohesive integration and reproduction.



But the membrane should not just be an envelop for the autocatalytic network, it could also selectively control the diffusion of reactants between internal and external aqueous solutions. This is of fundamental importance, since changes in the core autocatalytic network could modulate membrane properties, in a way that the flow of matter and energy between the system and the environment is, somehow, channelled. In turn, this leads to a qualitative difference in organization with regard to that of an unbounded autocatalytic network (like Kauffman's —1986). This (minimal) proto-cell would be capable to start controlling its boundary conditions for self-maintenance: i.e., to regulate the input of matter and energy that ensures the ongoing regeneration of components, while avoiding osmotic crisis and other threats to their organization.

Ruiz-Mirazo and Mavelli's expansion of the above model (Ruiz-Mirazo and Mavelli 2008) addresses precisely the passage from a self-re-producing minimal proto-cell to a minimal autonomous *agent* by expanding on the interactive capacities of the membrane. Whether the system grows and divides or bursts due to an osmotic crisis has been shown to depend crucially on the rate at which the waste product is released (diffusion parameter k_w). Thus, the stability of the system critically depends on its capacity to get rid of the waste product (and not so much on increasing the rate of “food” inflow). This might anticipate that in mBAS (at least conceptually) the most simple form of agency would be that of active regulation of transport, through the membrane, of part of its components (in particular its waste products). Ruiz-Mirazo and Mavelli define the main viability or stability factor Φ of the system as the ratio between the proto-cell surface and the surface of an ideal sphere relative to the volume of the proto-cell (low values of Φ might be thought of as indicating high tension). Since the ideal sphere represents, topologically, the maximum volume/surface ratio, if the surface is smaller than the ideal the system will burst. This imposes a lower limit on the Φ factor ($\Phi=1$), which the system must not cross in order to maintain itself (a refinement of the model includes additional elasticity coefficients, so that Φ can also take values slightly below 1). In turn, if the stability factor crosses an upper value (approx.: $\Phi=2^{1/2}$) the protocell will divide, corresponding to the situation in which the surface-volume relationship is such that allows the co-existence of two identical ‘daughter’ spheres. As a result, the limits of the stability factor define the space of possibilities for any protocell in the model. And it is in relation to those boundary conditions (in particular, to the lower limit) that the system needs to regulate its growth and membrane properties (permeability elasticity...) if it is to remain stable (avoiding an eventual osmotic crisis).

In order to achieve such regulatory capacity Ruiz-Mirazo and Mavelli have expanded the model with the introduction of endogenously synthesized pep-

tide chains capable to ensemble into the membrane and change its properties. The organization of networked reactions is illustrated in figure 5. The peptides in the membrane permit to increase its elasticity and/or permeability so that the boundary conditions of the system are modulated by its very activity (i.e. by the production of peptides in the proto-metabolic network). The simulation model does not yet include the coupling of endergonic and exergonic processes that would lead to the production of work as required to fully characterize an *active* process (more on this soon). But the model sufficiently illustrates a naturalized conception of agency as the interactive contribution to self-maintenance of a FFE proto-metabolic organization capable to modulate or regulate its boundary conditions. Figure 6 shows the effect of the production of the peptides in the Φ factor. The dashed black line represents the state of Φ for a system in which the peptides do not increase the permeability of the membrane, it shows how Φ collapses when the low limit for viability is crossed (right after the middle of the plot). Grey lines represent the behaviour of the system with peptides capable to increase the permeability of the membrane. The bottom line shows the concentration of peptides. As this increases, the permeability also increases and the system avoids the osmotic crisis. The grey line with a peak represents the concentration of waste product, decreasing as soon as peptide density in the membrane is high enough. Although, admittedly, the simulation does not address sufficiently the requirements for minimal agency developed at the conceptual-theoretical level (Ruiz-Mirazo & Moreno 2004) it does, for our purposes, illustrate what is meant by minimal agency¹⁴.

It is time now to review the conditions stated previously and see whether the conceptual and simulation model of basic autonomous systems just described satisfies them and how.

6. INDIVIDUALITY AND ENVIRONMENT IN MINIMAL AUTONOMOUS AGENTS

We have seen how cohesion is a necessary condition for identity and how self-organized processes lead to the appearance of an early *self-maintained* identity. It is precisely component production chemical networks that would permit to speak of *self-produced* identity. And, finally, autonomous systems not

14 Further developments (Ruiz-Mirazo & Mavelli 2007b) have included in the model peptides that change their configuration (and, thus, the actual permeability of the membrane) as a more apparent self-regulatory mechanism of the protocell. These rudimentary channels would be operating just when the membrane's surface tension is high (i.e., when the system is under 'osmotic stress' but has some means to counteract it: the opening of channels for W release). In addition, the introduction of active ion-pumping and other energy constraints are under way (Ruiz-Mirazo personal communication).

only entail the above forms of proto-selfness but also a genuine *active distinction between system and environment*, achieved through the production of a membrane and the active control of boundary conditions that it brings forth. Whereas many of the systems around us (like chairs, walls, cars or roads) are identified from outside (i.e. they have an identity that is the result of an external agent attributing it, in virtue of their cohesive unity and spatio-temporal continuity), autonomous systems might properly deserve the term *individuality* to characterize their mode of identity. Hans Jonas anticipated this criterion for individuality taking metabolism (the continuous renewal of the material constituency of a living being) as the paradigmatic feature and minimal expression of those systems that do not depend on being identified to become singular entities. Metabolic systems are those whose individuality unfolds as a result of their own activity:

Only those entities are individuals whose being is their own doing (and thus, in a sense, their task): entities, in other words, that are delivered up to their being for their being, so that their being is committed to them, and they are committed to keeping up this being by ever renewed acts of it. Entities, therefore, which in their being are exposed to the alternative of not-being as potentially imminent, and achieve being in answer to this constant imminence; entities, therefore, that are temporal in their innermost nature, that have being only by everbecoming, with each moment posing a new issue in their history; whose over time is thus, not the inert one of a permanent substratum, but the self-created one of continuous performance; entities finally, whose *difference* from the *other*, from the rest of things, is not adventitious and indifferent to them, but a dynamic attribute of their being, in that the tension of this difference is the very medium of each one's maintaining itself in its selfhood by standing off the other and communing with it at the same time. (Jonas 1968: 233)

In a similar vein Varela expresses:

Autopoiesis addresses the issue of organism as a minimal living system by characterizing its basic mode of identity. This is, properly speaking, to address the issue at an ontological level: the accent is on the manner in which a living system becomes a distinguishable entity, and not on its specific molecular composition and contingent historical configurations. For as long as it exists, the autopoietic organization remains invariant. In other words, one way to spotlight the specificity of autopoiesis is to think of it self-referentially as that organization which maintains the very organization itself as an invariant. The entire physico-chemical constitution is in constant flux; the pattern remains, and only through its invariance can the flux of realizing components be ascertained. (Varela 1992:3)

In contrast to the metaphysical, at times poetic, formulation of Jonas' individuality and Varela's self-referential identity, we are now in the position to offer a more precise and naturalised grounding for the concept of individuality. We can unravel individuality as the combination of a set of properties: a) *cohesive unity*, achieved through the structural stability of the constructive cycle and its enclosure within a self-produced membrane, b) *singularity and histor-*

icity, resulting from the irreversibility of the way in which fluctuations are propagated along the system, c) *dynamical reflexivity*, as non-linearity and recurrence of interactions; and d) *active differentiation* made possible through the production and active control of boundary conditions by the membrane. All these properties have a definite physico-chemical reference that can be empirically explored and rigorously modelled, there is no soul or external agent required to justify them, nor a poetic connotative excess that hides, at times, what it is precisely meant to denote.

The organization able to generate these properties is not any more amenable to pure physical description. It is not any more the language of happenings, of passive particles subject to fields and forces, nor the lawful statistical character of Boltzmannian order, but it is the “language game of agents”, to use Kauffman’s (2003) expression, what becomes the proper language to describe and explain these “entities”. In addition, as we shall see next, BASs (even their minimal formulation) are subject to unavoidable normative descriptions.

The continuous process of individuation in autonomous systems has its reverse precisely in the definition of an environment. In Jonas’ words:

The challenge of ‘self-hood’ qualifies everything beyond the boundaries of the organism as foreign and somehow opposite: as ‘world’, within-which, by-which, and against-which it is committed to maintain itself. Without this universal counterpart of ‘other’, there would be no ‘self’ (Jonas 1968:243).

As we have repeatedly highlighted, the system (the emergent *self* understood as a circular macroscopic correlation) can only exist as situated within a material and energetic environment in order to persist. But what this environment is (in relation to the system) is co-determined by its organization, by its form of self-maintenance:

Now, in this dialogic coupling between the living unity and the physico-chemical environment, the balance is slightly weighted towards the living since it has the active role in this reciprocal coupling. In defining what it is as unity, in the very same movement it defines what remains exterior to it, that is to say, its surrounding environment. A closer examination also makes it evident that this exteriorization can only be understood, so to speak, from the “inside”: the autopoietic unity *creates a perspective* from which the exterior is one, which cannot be confused with the physical surroundings as they appear to us as observers, the land of physical and chemical laws *simpliciter*, devoid of such perspectivism. (Varela 1991:4)

In effect, an agent cuts out an environment out of what we, as observers, take to be surrounding it. There are two major senses in which an agent can be said to be defining or co-determining its environment. First of all, through the normative requirements of its constructive processes, an autonomous system co-determines those, among all the processes and entities surrounding it, that can have some causal effect on itself. The most trivial case is the spatial scale at which the constructive processes take place. Be it sub-atomic, molecu-

lar, social, economic, digital or otherwise the level of realization of its autonomy will already determine a particular type of environment, i.e. what kind of surroundings can have an effect on the system¹⁵. In addition, the specific mode in which the constructive cycle is achieved will also co-determine what an environment for that system is. In the case of mBASs, the molecular species that participate on the metabolic network will determine which among the molecules surrounding the system can have an effect (functional or dysfunctional) on its organization. For instance, the way in which molecule X in the environment becomes relevant for living organization is not determined exclusively on the basis of its objective molecular properties, but depends on how X becomes a precursor of the nested set of reactions. In this sense, out of an in principle undifferentiated physical surroundings, living organization selectively cuts out an environment setting up the sources of potentially destructive perturbations as well as boundary conditions necessary for self-maintenance. This internally selective cut defines precisely what Varela terms a “perspective” which becomes evident in contrast with what an external observer may define as the spatio-temporally external surroundings of the system.

From this “perspective” some molecular species might be inert or neutral in relation to the chemical reactions involved on the metabolic network and thus, the system may equally be “blind” to them, unless, and this brings us to the second mode of system-environment co-determination, this constructively neutral molecular species becomes correlated with other functionally relevant processes present in the environment. In such case, these functionally neutral molecules will become part of the environment of the systems if, and only if, the system has some means of interacting with them in order to exploit their correlation with functionally relevant molecular species. Thus, the interactive processes that the BAS sustains will also co-determine its environment. In fact, it is in this later sense that the environment is truly an environment for the system *as an agent*. A negative case might illustrate the issue. The constructive cycle of a BAS could be severely disrupted in the face of a source of radiation but if the system has no means to cope interactively with that source (by detecting it and moving away from it) it would remind “blind” and thus inactive in relation to it, until its organization is destroyed: the source of radiation does not pertain to the environment of the system *as an agent*. In the absence of any explicit sensory and effector mechanisms within our minimal model of autonomous agents we shall postpone the analysis of this second level of “perspectivism” and continue exploring how this minimal organization comes to satisfy the remaining conditions for agency.

¹⁵ There are important reasons to believe that the lowest level form of individuality must occur at the chemical level (and not at any level below it—e.g. atomic or subatomic). The reason is that only the chemical level permits the combination of dissipative and conservative order required to achieve an autonomous organization that can, subsequently, increase its complexity

7. NORMATIVITY AND FUNCTIONALITY IN MINIMAL AUTONOMOUS AGENTS

Etymologically, autonomous systems are those able to generate or provide themselves with their own norms (*auto*=self, *nomos*=law,norm). From the point of view of the MUN constraints (explained in the previous chapter and standing for Minimalism, Universality and Naturalism) mBAS are the most elementary form of organization in which it is possible (and necessary) to speak in terms of a norm or normativity, in a strong sense: i.e. where the nature of the norm (what is good or bad for the system) is not externally interpreted or derived from an adaptive history but intrinsically established by the very organization of the system.

For artificial systems, we, as designers or users, can attribute a certain goal to the system according to *our* purposes. Under the attributed goal or purpose, we claim that the machine works *properly* or that it is broken and *mal*-functions; even if the good or bad functioning of the system (the machine) is completely extraneous to the structure of the machine itself. Equally, what belongs to the system and what should be left out as irrelevant (what constitutes the identity of the system) is determined by the goal we project upon it; i.e. what is an essential part of the system (as distinct from its environment) is defined in relation to the desired functionality that we, as designers or users, expect to achieve within a set of contexts of use. So, for instance, we don't say that the car is broken because, let's assume, the surface paint or the bumper has been damaged. However, a much smaller change in the protective paint of the oil temperature thermometer or a tiny damage on a gear might break the assigned functionality of the car. What constitutes a disruption, what justifies the claim "shit! my car (that red coloured iron structure standing there) is broken" is the functionality we have assigned to it. The very same "red coloured iron structure standing there" would not be considered broken for having a tinny damage on a gear if it was exposed in a museum as a work of art. In such case a damage on the surface paint may have been dramatic, again for us, the artist or the security guard of the museum. The nature of the norm, dwells outside its structure and functioning.

In contrast, in autonomous systems, as Kauffman suggests:

"[T]he concept of 'doings', as opposed to mere happenings, takes its place in our conceptual system. Hume argued that we cannot deduce 'ought' from 'is'. But my definition [*an autonomous agent is something that can act on its own behalf in an environment*] jumps this gap definitionally. For once we admit that the autonomous agent is acting on its own behalf, we have a locus of value in a world of fact." (Kauffman 2003: 1090)

In fact, BAS are capable to define themselves (as we explained above) and, more specifically, to determine their own normative functionality: i.e. what is

good or bad (right or wrong) for them does not depend on an external observer, designer or user, but lies on their own organization. More specifically in autonomous systems a process (constructive or interactive) is **functional** if it contributes to its self-maintenance (Piaget 1967/1969, Umerez, Etxeberria & Moreno 1993, Bickhard 1993, Umerez & Moreno 1994, Collier 1999)¹⁶. A process, in turn, becomes **normative** if it is dynamically presupposed by other processes in their contribution to the overall self-maintenance of an autonomous system (Christensen & Bickhard 2002). This is so because a constructive or interactive component process is dynamically coupled with the rest of components so that the overall maintenance of the whole organization depends on this process interdependence. Normativity refers to the fact that a set of processes that constitute the system *must happen as they do* in order for the system itself to exist (assuming that each of the processes that together constitute the system can, in principle, function in some other way¹⁷). Due to this circularity in BAS, normative functionality is not observer dependent but intrinsically causal, since the whole network (the very system) will not exist in the absence or malfunctioning of the component processes (given its FFE nature and the circular dependency between processes). In other words, in BASs, unlike in most artificial machines, what-the-system-does (the way it functions) and what-the-system-is (its structure) are highly intertwined: they merge together in the context of its organization. In this sense, it is important to recover an idea that we have suggested earlier regarding the relationship that mechanisms presuppose between structure and function. We saw how mechanistic explanation proceed by establishing a, more or less, direct mapping between structural parts of a system and the function they perform within the context of the full system. What is characteristic of autonomous systems, however, is that the elements of this dichotomy appear intimately intertwined: the structure depends on the function as much as the function depends on the structure.

One could trivially expand this condition for conservative structures to say, for example, that the basement of a building must be there in order for the house to exist, that the functional normativity of the basement is to support the structure of the building. For the case of dynamic conservative structures we could also claim that in order for a planet to orbit around a star, the planet must have a certain momentum and the star a certain mass for otherwise the planet would escape from the gravitational field of the star. Therefore, momentum and mass may appear as functionally normative for the orbit to

16 This idea is already implicitly contained in Aristotle but, to my knowledge, it was first explicitly formulated by Piaget: "The biological function (...) implies the existence of a system, i.e. of a structure or cycle, that conserves itself and encompasses activities that concur in this maintenance" (Piaget 1969/1967: 51 §5.II, translated from the Spanish version).

17 This last clause leaves physical higher level regularities (expressed as first principles or laws) outside the realm of functions.

exist. These cases are trivially true but do not suffice for normative functionality. The building must not pre-exist in order for the basement to exist. The basement is not intertwined with the rest of the component structures of the building so that they mutually constrain and enable the existence of each other. Similarly, the mass of the star and the momentum of the planet are conditions for the latter's orbiting around the former. But if things had been otherwise (mass and momentum and spatial relationships were different, not giving rise to an orbit) both planet and star would have continued their peaceful existence through the universe. In addition, once the building is there or the planet orbits around its sun, the configuration will continue forever until it is entropically degraded. The system needs not (and cannot!) perform any work to maintain its organization. Nothing else must be "done" for the continuing existence of the building or the orbiting. Energy would be required to destroy that form of order but never to feed its continued process of self-maintenance.

In the case of a BAS the conditions of its existence are continuously being regenerated. The BAS is subject to a permanent *precariousness* (to use Jonas' term) that is actively compensated by its own active organization. This precariousness involves that whatever the BAS *is* doing (i.e. whatever its factual functioning is) there is something that it *ought* to do; not for an external observer but for *itself*, for the continuation of its very existence. In Jonas' words "[For metabolism] 'To be' is its intrinsic goal. *Teleology* comes in where the continuous identity of being is not assured by mere inertial persistence of a substance, but is continually executed by something *done*, and by something which *has* to be done in order to stay on at all: it is a matter of to-be-or-not-to-be whether what is to be done *is* done." (Jonas 1968:243).

A basic example of normative (proper, necessary) functionality is given by the active transportation through the membrane of cells. This process becomes normative because the level of chemical concentrations that the membrane's active transport keeps within the cell is necessary for some metabolic reactions to maintain the appropriate rate, that in turn is necessary to sustain the network of metabolic reactions, that in turn is required to regenerate the membrane, and so on in a circular and interdependent manner. For instance, on Mavelli and Ruiz-Mirazo's model the reaction $A_1 + X \rightarrow A_2$ (as an organizational part/process) has the function of producing A_2 molecules. This reaction is a *function* (mathematically speaking) that "happens", but it is also a *normative function* that "must happen", because of the way in which it is integrated in the overall organization (the circular network of constructive reactions). Other functions or reactions also happen within the system but are not normative. Say, for instance, that $A_2 \rightarrow A_3 + W$, then the normative function of the process (the reaction) will not be to produce W (although it does, in fact, produce it) but, as we noted above, to produce A_3 . As a higher-level example, the normative function of the kidney is to filter blood, because the dynamic-meta-

bolic organization of the rest of the organism relies on this blood filtering for its functioning and existence.

At this point it is convenient to make a distinction between normativity regulation and normative regulation. From an organizational perspective **normativity** is a somehow transcendental property defined by the internal *conditions of possibility* of a dissipative organization¹⁸. **Regulation**, on the other hand, refers to the control or active compensation of perturbations according to a given reference state or rule that is operationally fixed on the mechanism of control. For example a thermostat regulates the temperature in relation to a reference state that is fixed (by an external user or designer) on the workings of the very mechanisms. An organism may or may not be able to regulate itself according to its autonomously defined normativity. For instance it may be regulating its temperature to 42° C, maintaining this temperature invariant in the face of perturbations, although it would be harmful and could ultimately destroy the organization of the system. Equally, some membrane channels may regulate transport in relation to certain concentration difference fixed, for instance, by the width of the channel (which is, in turn, specified by a number of factors like the size of the proteins forming the channel). Some of the factors specifying the width of the channels may vary and this will, in turn, fix the concentration difference that the channel regulates. Therefore the regulation can occur that is dysfunctional for the organization of the system. The channels, however, are still regulating concentration flow, but to a dysfunctional rate.

Thus, it could happen (and it does occasionally happen) that a BAS is regulating itself wrongly, but still regulating itself in relation to a reference state. So, there must be a principle by which, independently of the current functioning of regulatory mechanisms within an organism, it is possible to determine that the mechanism is doing it wrong. This well/wrong, good/bad, value attribution must be naturalized in the normativity that the FFE organiz-

18 The term transcendental is appropriate here not in the sense of a metaphysically transcendent property that goes beyond what is at hand, but in the opposite direction, inaugurated by Kant, in the sense of what is logically previous and required for a phenomenon to appear. The Kantian turn was of an epistemological nature for he took some categories (e.g. space and time) to be transcendental in the sense of necessary and prior requirements for experience itself. But the issue of normativity enters the Kantian scene particularly in his second critique, in which pure practical reason was shown to be able to self-determine through principles grounded on the conditions of possibility of universalizable action. More recently Apel (1998) and Habermas (1981/1984) have defended a linguistic-pragmatic version of Kantian transcendentalism for ethical normativity. The core idea is that norms are derived not from facts (such as empirically derived principles for the attainment of a goal that in turn needs to be justified) but from the conditions of possibility of language acting as foundational for our social dimension. Thus, for instance, “not to lie” is a norm of language and an ethical norm since if lying would be generalized or systematically performed it would make communication impossible. The very act of communicating presupposes veracity. This is of course not to say that lies cannot occur, like temporary dysfunctions in an organism also happen, but it just implies that systematic lying is completely dysfunctional and that overall truthfulness is the condition of possibility of communication (it makes communication possible) and thereby becomes normative in a transcendental sense.

ation that an autonomous system brings forth. We call **normative regulation** the regulation that is carried out according-to or in correspondence-with the normativity defined by the basic autonomous organization of the system¹⁹.

The holistic, integrated and self-maintaining organization of autonomous systems has some important consequences on the way they are described. For instance, the use of teleological terms to characterize their normative regulation can be naturalized, unlike its use to describe some artefacts that perform a goal seeking behaviour, such as the thermostat or a target-seeking missile. These are artefacts that have been designed to correct their behaviour (usually by a negative feed-back mechanism) according to an *externally* defined “goal” state. Expressions such as “the purpose of the thermostat is to maintain the room temperature at 23° C” are used as metaphorical shortcuts to describe the behaviour of such systems. They serve to escape the details of their mechanisms and exploit our intentional description of purposeful human cognitive behaviour to interact with such mechanisms. But what-the-goal-state-is remains completely extraneous to the mechanism that achieves it, the system is independent of the goal state or set of parameters it controls (which are externally imposed). Autonomous systems are different. Their existence depends on the FFE stability they produce. The stability point or set of points through which the system can exist, and from which perturbations are compensated, constitute the goal states of the system. This goal state is not just a goal state because the system compensates deviations from it, but because the goal state is the *condition of possibility of the very system*. Thus, BASs do not just regulate a set of parameters but they normatively regulate their existence. In other words, in BAS the goal state of the system and the organization that instantiates it are one and the same thing. Autonomous systems have an **implicit teleology** since their internal causal circularity makes each process of the system contribute to its global self-maintenance (as envisioned by Kant in his third critique).

The basic kind of minimal autonomy that we have explored is the lower level, most fundamental kind of autonomy: that of material and thermodynamic self-construction and self-maintenance, constitutive of all living forms and upon which higher levels of autonomy may appear. However mBA already generates a cascade of emergent properties such as individuality, normative functionality and implicit teleology. By expanding the above ana-

¹⁹ Note that there are two major senses in which a BAS might regulate itself. Minimal BAS cannot exercise any regulation other than what its basic constitutive organization intrinsically permits (e.g. under some boundary conditions the system might fall into different, equally stable, global attractors) but there is nothing that allows the mBAS (as we have described it) to switch between different metabolic pathways under different environmental conditions. More sophisticated forms of BA will include specific regulatory mechanisms, under which normative regulation will acquire its full blown meaning with the concept of *adaptivity*. But this topic will constitute the target of the next section. For the sake of the argument we are developing at this point, the kind of rudimentary, holistic and integrated regulation that mBAS show will be sufficient.

lysis and properties to the interactive cycle of autonomous systems we will be able to achieve a full-fledged characterization of agency and expand minimal agency towards more elaborated forms, which are fundamental in our trip to cognition (such as adaptivity, motility and behaviour). But I shall first specify the last condition leading to a full account of agency: the causal asymmetry condition.

8. CAUSAL ASYMMETRY IN MINIMAL AUTONOMOUS AGENTS

From the point of view of modelling, the interactive processes of a BAS, we can abstract a set of parameters as representing the boundary conditions and stability regions necessary for the maintenance of a FFE organization. Following Ashby (1952) we call these parameters *essential variables*, and the range within which the system's organization can be maintained *viability boundaries*²⁰. Thus, for example, in a mBAS a certain concentration of a reactant might be necessary for the coupled metabolic reaction network to maintain a cohesive level of stability. The concentration of this reactant will be an essential variable, while its viability boundaries are given by the critical range of concentrations under which it effectively contributes to the overall metabolic network. Another example of a viability boundary is the critical value of the Φ factor (that measured the relationship between the surface and volume of the proto-cell in relation to that of a perfect sphere) so that values below 1 indicate an excess volume that leads to bursting.

The FFE nature of autonomous systems makes that at least one of their essential variables has an intrinsic inertia towards outside the viability boundaries, so that the activity or functioning of the system is essential for compensating this inertia. Some of the essential variables are also non-controlled variables, in the sense that no-change of internal variables of the system can directly control its state. Recovering our example the reactant in question (say X in the simulation model above), it may not be itself synthesised by the metabolic network so that its concentration (the essential variable) needs to be interactively controlled by getting it from outside the system. As a consequence, only the coupling between the system and the environment can maintain the essential variables within viability constraints. Hence the importance of interaction processes in autonomous systems.

²⁰ Ashby used the term "viability constraints". I consider however that it is much clearer to speak in terms of boundaries rather than constraints. The distinction might be captured in terms of the previous distinction between regulation and normativity. I shall use the term constraint in the sense of a real existing dynamical limitation that cannot be crossed or violated due to the way in which the system is organized. Similar to the notion of regulation, the notion of constraint, although more passive on its connotation, reflect something that is occurring or present in the system. The notion of boundary does not imply any impossibility or fact about the dynamics of the system as such, it is nothing more than a limit that "should" not be crossed and has a normative character.

Thus, among the system-environment relationships, and *from the point of view of the effect* of the process on the system, some processes are **functional** for the system when they contribute to maintain essential variables within their boundaries of viability: these processes function according to the norm that the autonomous system defines. Others can be **dysfunctional** when they “push” essential variables outside the boundaries of viability; these processes operate against the norm defined by the system. And many others are **neutral** (they have no effect on the state of the essential variables)²¹. This has been treated in the previous section. However, *from the point of view of the cause* of these processes we can also classify them as **active** (if the processes are triggered by the system as a whole) or **passive** (when the interaction is imposed from outside).

We can now properly use the term **agents** to name those systems whose coupling with the environment is *actively* controlled by them, resulting on changes that contribute to their self-maintenance. The fact that a coupled process is active, rather than passive, is of fundamental importance for the task of naturalizing agency since it constitutes the *causal asymmetry condition* stated at the beginning of this section. By active we mean, in negative terms, that the functional interaction is neither produced (i) by some external source nor (ii) by means of unconstrained physical laws (i.e. spontaneously and independently of the particular organization of the system). So, for instance, imagine we had a mBAS and we were in the position to introduce some functional reactant through its membrane to compensate some occurring instability. In such a case, our system would be able to create its own cohesive identity, to distinguish itself from its environment, and to define, by its own organization, a normative functionality. Then this action upon the system would be functional, in the sense that it would contribute to its self-maintenance, but the mBAS would still not be an agent with respect to this particular interaction, since it was externally imposed by another agent (ourselves)²². A second case of functional interaction that would fail to satisfy agency is that in which the interaction is the result of spontaneous, not internally constrained, processes. Osmosis is a characteristic example. Given a semi-permeable membrane, small molecules such as water, can easily pass through the membrane spontaneously, due to Brownian motion and osmotic pressure. Under some environmental conditions, internal concentrations may vary due to these spontaneous osmotic compensating processes. Yet, again in these cases the functional system-environment interaction fails to be agential: the process is caused by

21 Note that by introducing the notion of essential variables we come to redefine normative functionality in dynamical terms, now making the concept applicable to the model of an autonomous system, rather than to the FFE structure of its target system as explained in the previous section

22 A nurse taking care of a patient, or a cat providing bodyheat to its kittens, are higher level examples of genuinely functional but non agential system-environment “interactions”.

intrinsic properties of the membrane plus Brownian motion and physico-chemical laws or principles but independently of the organization of the BAS.

We have ruled out the cases in which the functional interaction is passive but... what does exactly make the difference so that we can say that a process is active and “caused” by the system? As a first approximation, an interaction will be active if the autonomous system can constrain its coupling with the environment in ways that secure its self-maintenance. Actions require work, and work requires suitable energy. Thus, at the minimal level, active interactions require that the system channels the thermodynamic flow of its FFE or organization into the creation of specific constraints. The term constraint is here used to mean that some physical or chemical processes are limited or shaped, channelled towards directions that do not follow from the effect of physical laws or principles without the presence of that limitation, that is in turn produced by the systemic organization (Pattee 1972, 1973—see also Umerez 2001). It is here where Kauffman’s conceptual model of autonomous systems, as those able to instantiate work-constraint cycles, provides an insightful approach (Kauffman 2000, 2003). To perform work (i.e. useful directed release of energy), constraints need to be built by the system and, conversely, to create constraints work is required. It is the characteristic coupling between *exergonic* (free-energy releasing and thermodynamically spontaneous) and *endergonic* (free-energy requiring and non-spontaneous) chemical reactions that provides the means to achieve a work-constraint closure in BAS. It is in this precise sense in which the system can be said to be the *active* source of a functional interaction, when these work-constraint cycles are recruited or mobilized to regulate or direct system-environment interactions.

Thus, in autonomous systems *only those coupled system-environment process that are functional (contribute to self-maintenance) and causally asymmetrically laden to the side of its organization can properly be said to be agential in nature.*

9. LIFE AND ITS INTRINSIC AGENTIAL NATURE: MINIMAL AGENCY NATURALIZED

We have specified a set of physico-chemical conditions leading to the appearance of the basic organization that characterizes minimal autonomous systems. When this type of individuality recruits energy to perform an active functional (normative) interaction within the environment, a minimal form of agency can be conceived that satisfies the conditions of individuality, normativity and causal asymmetry

This conception of agency, modelled as a circular, emergent, self-sustaining and far from thermodynamic equilibrium chemical organization, comes to sat-

isfy the MUN constraints stated above. It is a *naturalist* model since it stands grounded on the material and thermodynamic properties of the components and relationships that make up the system. No reference to *vitalist* forces, conscious awareness, rational thought or eidetic intentions is required to specify the “essence” of agency. The model is *universal* for it can generalize the basic organization of agency without reference to arbitrary or local contingent properties of life-as-we-know-it, while remaining coherent with its objective/material conditions of possibility. In addition, although the model focuses on a minimal cellular level, all its fundamental properties can be generalized to more complex living forms. Finally, it is out of question that it is a *minimal* model because it integrates all but only those component processes that are crucial to specify the most fundamental features of agency²³.

We have seen how a mBAS (as a component production network endowed with a self-produced selective membrane) provides a minimal model for agency. In particular, active pumping through the membrane provides an example of *a system doing something by itself according to a certain goal or norm within a specific environment*. It instantiates the three conditions stated at the beginning of this section: (i) the individuality condition, (ii) the normativity condition and (iii) the causal asymmetry condition. The concept of a basic autonomous organization became a pivotal notion that subsumed all three conditions. As a FFE dissipative form of cohesion, a mBAS is a system that recursively defines its own identity through a continuous process of self-production and interactive maintenance. The intertwined stability relationship between its component processes provided the means to ground normative functionality. And, finally, we distinguished a set of system-environment

23 In this sense Kauffman's previous attempts to define minimal autonomous agency deserve some explicit discussion. “I will call a system that can act on its own behalf in an environment an ‘autonomous agent’. (...) So my question became: what must a physical system be to constitute an autonomous agent? I will jump to the definition I found my way to after much consideration: an autonomous agent is a self-replicating system that is able to perform at least one thermodynamic work cycle” (Kauffman 2003). On the one hand, that the system be self-replicating is not necessary nor sufficient for agency. We can easily imagine even a sophisticated biological system that is incapable of self-replication (e.g. a mule) that is still a full-fledged agent. A self-(re)-producing organization is a much more adequate departure point than that of replication. In addition, in its minimal form, a self-producing system such as the proto-cell model explained above is capable of self-reproduction for free. If its reproductive rate is even infinitesimally greater than what mere self maintenance requires the system will grow and ultimately, under the laws of physics, will split in two. On the other hand, the work-cycle is necessary for the causal asymmetry in molecular agents but, in itself, remains insufficient for agency since, taken in isolation, it does not involve any necessary relation with anything like performing and action in its environment. For example the model of molecular agent described above does not include any reference to environmental interactions nor does the disappointing simulation model of an ‘autonomous agent’ that Kauffman and collaborators have recently published (Daley et al. 2002). This is not to say that self-replicating systems with at least a thermodynamic work cycle might be close or even constitute an agent if the appropriate environmental fluctuation and their necessary compensations are taken into account, but that the conditions of self-replication and work cycle alone do not capture appropriately what minimal agency requires. In this sense Ruiz-Mirazo and Mavelli's work (2008) seemed to me a much more precise and insightful model towards minimal agency; almost a unique example of a naturalized and bottom-up approach to the foundations of agency.

interactions that were actively generated by the basic autonomous organization thus producing a causal asymmetry characteristic of agency.

All these conditions for agency were satisfied through a circular and recursive mode of organization (that of BASs). That many scientists and philosophers have long ignored this form of reconstructing the notions of individuality and normativity is probably due to the difficulties that so many theoretical frameworks of western thought have shown with circular causality. Although the arguments have been linearly exposed (as there is no other way in which the inherently sequential linguistic discourse can be organized) it should be by now clear that individuality, normativity and causal asymmetry emerge together through BA organization. Within this form of organization, the often present dichotomies of local-global, part-wholes, microscopic-macroscopic, point towards some form of circularity. Rather than the notion of a *linear sequence*, what best captures this organization is the notion of a *non-linear cyclic network* and the simulation models that generate agential capacities through networked interactions. Properties that arise from networks considered as a whole are often elusive, if an aggregative analytic approach is used to study systemic properties. The aggregative decompositional method of enquiry into systems was characteristic of the biological and cognitive neurosciences that we had recently at hand to provide a naturalist conception of agency (Bechtel & Richardson 1993, Ruiz-Mirazo, Moreno & Barandiaran 2008). Under this received view, functional properties are usually local and transferred via some specific “communication” channel to other components of the system. For instance, the speed of a car is determined by the engine that transmits the kinetic energy to the wheels through a set of gears. Surely the motion of the car depends on a set of *ceteris paribus* clauses: there needs to be some friction between the wheels of the car and the floor, the car must be free to move, etc. But functions are local and not recursively dependent on one another, so one can isolate those functions and (de)compose the system linearly. This has been the model of an explanation in biology until the advent of complexity theory and the modelling tools it has provided. In a network, on the contrary, processes are highly interconnected, the activity of the system is determined by the global configuration of the processes, and systemic properties arise which depend on the global stability conditions of the network. Agency is one such property and its characterization requires to specify a model of the type of systemic organization capable to generate it.

Systems biology, using the modelling techniques for complex systems, is progressively delivering a picture of life that integrates its networked and circular organization in a scientifically tractable form. From its very origins, at least on the way these origins are hypothesized and modelled within the “metabolism first” school, life itself appears endowed with agential capacities. And it is by looking at minimal models that capture what would formerly had

been called the “essence of life” that the concept of agency can be modelled and related to existing biological phenomena. All known forms of life can be said to share this fundamental organization, from plants (growing and transforming their most immediate surroundings to absorb nutrients) to the sophisticated capacities of gorilla’s hunting, through the movement of a flagellated bacteria up the sugar gradient or the escape behaviour of a jellyfish. The struggle for survival is first and foremost a biochemical struggle for asserting and maintaining a dissipative existence that requires a continuous re-generation and negotiation of its boundary conditions.

Philosophers of mind and scientists have long attempted to ground the foundations of mind into the insights that our own creations have opened, from puppet automata to telephone systems and computers. But, as we advance on its understanding, it is life itself that is providing the metaphors and cues to solve the puzzle of reconciling the “language game of agents” with the inanimate picture of reality that we have so long hand over to physical mechanics.

Life found its way towards cognitive capacities and, ultimately, towards the generation of mindful organisms. This process requires that we include additional components and relationships to our minimal model. The mBAS we have taken as a model of a minimal agency reaches a dead end regarding its capacity to cope with anything other than a rudimentary and extremely simple environment composed of immediately functional molecular species and their concentrations. So, although we have succeeded in characterizing a set of essential conditions for agency, we are still far away from anything near cognition. In order to overcome this limitation we need to look for systems whose boundaries of complexity growth remain open and trace the trajectories that, through this complexity space, lead to the characteristic organization of cognitive agents. All three conditions of agency: normativity, individuality and causal-asymmetry will be transformed alike in the transitions that lead to cognition: environment and mode of organization will co-evolve together with additional regulatory needs and opportunities for more complex forms of agency. Our next steps will bring us to analyse the adaptive mechanisms that specifically regulate interactive processes. Through the appearance of a particularly important kind of adaptive interaction, motility, we shall see how minimal forms of adaptive BAS will reach an organizational bottleneck, and how a new form of organization will appear leading to neurally guided adaptive behaviour and getting, thus, closer to what is commonly accepted as genuinely cognitive. But first it must be analysed what precludes mBASs to expand their interactive capacities and how the mechanisms that overcome such limitations trigger a set of changes on their organization that opens up the path to be walked towards cognition.

Chapter 5: Adaptivity and sensorimotor coupling in chemodynamic autonomous agents

1. THE LIMITS OF PROTO-METABOLIC ROBUSTNESS WITHOUT EXPLICIT REGULATORY CONTROL

The cohesion of the minimal proto-metabolic agent, described in the previous section, depends mainly on the encapsulating feature of the membrane whose permeability permits the inflow- outflow of matter and energy and the retention of the appropriate concentrations inside the system. But cohesion also relies on the *stability* of the reaction dynamics maintained by the feed-back loops of the constructive cycle and on other factors (such as the buffering that molecule B facilitates reducing local fluctuations in Ruiz-Mirazo and Mavelli's model¹).

The notion of cohesion introduced at the beginning of the previous chapter, is closely related to that of stability and puts the emphasis on the *unity* of the systems in the face of internal and external perturbations. In par-

¹ The contribution to robustness of buffering molecules is common and well known in chemistry. Think on the effect of introducing padded columns on a stadium to avoid human avalanches. The notion of osmotic buffering can be illustrated by Ruiz-Mirazo and Mavelli's introduction of non-reactive B molecules into the system, assuming an equal concentration inside and outside the proto-cell. The problem that it comes to solve is the devastating effects that fluctuations can have on small chemical systems. These fluctuations are the result of Brownian forces and other sources of stochasticity in the system. When chemical systems are sufficiently large these local fluctuation can often be neglected due to the averaging-out effect on immense number of components. But small systems are prone to destabilizing local fluctuations. The introduction of non-reactive elements in this system serves to absorb the effect of these fluctuations and increases the stability of the system. (Mavelli & Ruiz-Mirazo 2007 provide different simulation experiments to test the effect of buffering molecules in the proto-cell, demonstrating that they do cut down the importance of fluctuations).

ticular mBAS, as specified above, can only compensate small perturbations recovering the steady state that maximally dissipates energy. They achieve robustness by negative feed-backs integrated and distributed over their constitutive organization (on the coupling between reactions). There are three outcomes that can arise from a perturbation of a self-organized and structurally stable system: a) the system cannot compensate the perturbation and is destroyed, b) it compensates the perturbation by falling into the same attractor (negative feed-back) or c) it changes to another attractor (if any) that is “equally” viable (this form of multi-stability is generally achieved through positive feedback, that brings the system to a new stable attractor). An example, in the case of our model of mBAS, can be shown by perturbing the concentration of, let's say, A_2 type of molecules through the injection of a high number of such molecules inside the system. As a result, because A_2 molecules react producing A_3 and W molecules, it can be the case that: a) the sudden increase in the concentration of A_2 produces too much A_3 and W and the system undergoes an osmotic burst, b) A_3 is quickly transformed in A_4 and the remaining W is tolerated until the system recovers the original proportion of concentrations after a short period of time (the perturbation moves the variables away from the point attractor but they remain within the basins of attraction and soon fall back into the the point attractor) or c) that the concentrations change and settle into another, equally stable, configuration (the system is moved to another attractor). We can see that in such minimal organizations the regulatory mechanisms are the very same constructive processes, so that regulatory and regulated processes cannot be explicitly distinguished. In fact, in such cases, properly speaking, one could not even argue in terms of regulatory mechanisms. *Structural stability* is a more appropriate term: reactions are nested in feed-back loops so that changes in component concentrations affect one another and define a dynamical system with one or more stable attractors². One can think of structural stability as a way of achieving robustness that typically involves the recovery of a stable state that was characteristic of the system before the perturbation took place. Jen proposes the following working definition:

Loosely speaking, a solution (meaning an equilibrium state) of a dynamical system is said to be stable if small perturbations to the solution result in a new solution that stays ‘close’ to the original solution for all time. Perturbations can be viewed as small differences effected in the actual state of the system: the crux of stability is that these differences remain small for all time. (...) A dynamical system is said to be structurally stable if small perturbations to the system itself result in a new dynamical system with qualitatively the same dynamics. (Jen 2002:2).

² Their stability corresponds with the idea of conservation of autopoiesis as originally formulated by Maturana and Varela in their original work (Maturana & Varela 1980).

Structural stability will typically involve a system configuration such that small changes in parameters leave the system behaviour almost unaffected and/or such that changes in variables are compensated through negative feed-back loops that bring the system to the original stable attractor (i.e. the spontaneous compensation of perturbations when these fall into the basins of attraction). But perturbations may be of a different nature requiring more sophisticated forms of compensation and different means to achieve robustness. One may threaten a system by perturbing some variable or parameter (e.g. by altering the concentration of a metabolite in the environment) or by radically changing the environmental conditions (e.g. by suppressing altogether the presence of a metabolite in the environment and substituting it by another one). Structural stability is only able to assure robustness regarding the first type of perturbations. Achieving robustness to radical changes of environmental (or internal) conditions is a different issue, it involves an operation that is equivalent to changing to a different attractor (but to the adequate attractor, not any one will do). I will try to argue that systems that possess just a highly integrated and structurally stable organization are very limited on their capacity to achieve robustness to radically changing environmental conditions and that overcoming bottleneck on the morphophylogeny of agency requires the appearance of adaptive systems with explicit regulatory mechanisms.

As suggested, in the absence of explicit regulatory mechanisms, the capacity that a given fully-integrated system has to maintain its identity against perturbations (its robustness) depends on its capacity to fall under the appropriate collective pattern or attractor³. Now, problems arise when the system needs sensitivity: i.e. specific organizational responses to specific environmental conditions (e.g. activating a new metabolic pathway). The pay-off between sensitivity and stability produces a “dilemma”. If the self-organization process is robust against variations in specific conditions the process will be reliable, but it will be difficult for the system to generate multiple finely differentiated global states. Alternatively if the dynamics are sensitive to specific conditions it will be easy for the system to generate multiple finely differentiated global states, but difficult to reliably reach a specific state (i.e. one that is equally viable and adequate to counteract the perturbation if it persists, e.g. if the concentration of a reactant in the environment is persistently

3 The following argument is partly inspired in Christensen (2007) where a similar (but not identical) claim is made regarding the need for hierarchical regulatory control in neural organization. Christensen's arguments does not hold for neural dynamics of complex brain functional networks but I will try to show that the argument is sound for minimal proto-metabolic networks. The difference lies on the connectivity structure that long axons permit in brain networks that renders functional network structure into very different regimes than those of other biological networks constrained by the limitation of diffusion processes (see Sporns et al. 2004 for details). I will relay on Hopfield networks to illustrate the problem of metabolic networks, but Hopfield architectures are not the type of scale-free networks that have been found to underlie neurodynamic processes.

high). At this point a crucial transition occurs due to a complexity bottleneck on the possible amount of functional responses that integrated self-organized networks can achieve to changing environmental conditions.

This limitation of self-organized and structurally stable proto-metabolic systems (without additional mechanisms) is of fundamental importance and requires a more precise analysis. The problem may be illustrated by recent work done by Molter, Bersini and colleagues (Molter 2004, Molter et. al. 2005). Although their work was carried out in the domain of neural systems and in terms of information storage, the underlying logic holds for a chemical network⁴. Their research shows that the dynamics of a self-organized network (unsupervised Hebbian learning in Hopfield networks) can achieve a high number of attractors, but only if those attractors are left to be assigned spontaneously. If they are to be functionally correlated with a specific response, only a low number of attractors can be built into the system. In addition, in this second situation, if the environmental condition is noisy (the structure of the input data does not fall under stereotyped training set) the system undergoes a chaotic regime (the boundaries of attractors are unstable). So, even if mBAS had “evolved” to accommodate a set of attractors which are adequate, given specific environmental conditions, most chances are that, if the current environmental condition does not fall under a specific type of condition, the system will be driven to a chaotic regime. One could argue that there is no need for the system to fall under specific attractors and that it could be left to fall under spontaneous ones even under chaotic regimes. After all, as far as the reaction cycle is completed, the organization will be maintained. However, we are not dealing with abstract computational relations but with real material systems and there is, even in our mBAS model, a set of viability boundaries (those of the essential variables) that the system must satisfy (and more constraints are to be expected in more complex metabolic like networks). For instance, the overall internal concentrations should be kept within certain viability margins, otherwise, an osmotic burst will happen due to an increase in volume that the membrane cannot enclose. Given the need to satisfy these constraints, all attractor states are not equally viable and there is no guarantee that spontaneous attractors will satisfy such viability constraints. Thus, since the system cannot rely on spontaneous attractors to achieve robustness it will suffer, as shown by Molter and Bersini, two major problems: a) the number of attractors will be much reduced and b) transitions

4 Neural networks of this kind are sufficiently abstract models to generalize to metabolic networks. First information storage is interpreted in terms of attractor states (irrespective of higher-level representational functions) so that it can be easily generalizable to other networks. The Hebbian learning algorithm that they use may be equally interpreted as a method for achieving network configurations that lead to a specific phase space, and take the resulting network configuration as a model (independently of the feasibility of the technique used to achieve it).

between attractors and noisy perturbations will be prone to chaotic regimes (with devastating effects).

Thus, in the absence of anything other than self-organized structural stability no possibility for robust complexity growth exists in chemical networks. This means that, given the proto-evolutionary capacities of mBAS⁵, these additional mechanisms would be retained, thus transforming their, now not so minimal, organization to encompass more sophisticated forms of robustness (other than mere structural stability). But... what kind of mechanism (other than positive and negative-feedback) may be found? And how would that new sort of mechanisms lead to new forms of agency? Do they introduce new organizational features leading to more “cognitive-like” systems?

2. ADAPTIVITY AND DECOUPLED CONTROL MECHANISMS

Kitano (2004) analyses alternative mechanisms used by biological systems to attain robustness, in addition to the structural stability permitted by feed-back loops. One of them is *redundancy*, so we could envision a further complexification of our mBAS model, where redundant loops within the constructive cycle were able to encompass stronger perturbations. For instance redundant metabolic routes could make the system able to metabolise different substrates and to compensate the lack of some substrates through alternative metabolic routes. There is, however, a price to be paid by redundancy in terms of thermodynamic efficiency and, although redundancy may provide a higher number of attractors or increase the basin of attraction, the problem explained above remains to be solved. Another mechanism that Kitano considers is that of *decoupling*, meaning isolation of “low-level variation from high-level functionalities” (Kitano 2004:828). In our case, decoupling will entail that a set of molecular processes which do not directly participate on the constructive cycle take over regulatory functions, thus “freeing” themselves from having to satisfy strong viability constraints, i.e. the inverse of Kitano's definition will hold for our notion of decoupling: *independence of low-level functionality (constructive processes) from high-level variation (state of regulatory mechanism)*. Thus, free from the lower-level constraints, higher level variation can be left to “spontaneous” dynamics, provided that a further coupling is established linking higher level states to lower level ones in a function-

5 Alvaro Moreno and Kepa Ruiz-Mirazo, have suggested to me not to use the term autonomous systems for those chemical component production networks with an encapsulating membrane that only show self-organized structural stability. Of course, it all depends on how many conditions one considers that a minimal BAS should need to satisfy and whether structural stability is sufficient or not. For clarity of exposition, I shall consider that mBAS are sufficiently well characterized by a chemical production networks and a selectively permeable membrane, as exposed on the previous section. And take this minimal form as a departure point to include additional mechanisms towards not-so-minimal forms of basic autonomy.

al way; so that this higher level variation (or freedom) is in turn functionally recruited to serve lower level functions. For instance, consider the case of a set reactants that are a product of the core metabolic cycle in a mBAS but do not contribute to it, i.e. they are side products. These reactants could in turn constitute additional reaction processes that may be tuned to specific viability conditions of the system (e.g. regulating membrane properties in correlation with metabolically relevant conditions).

Hence, in not-so-minimal BAS, robustness may be enhanced by decoupled active **control**, acting upon its interactive and constructive processes, both sensitive to different external conditions and internal global states so as to avoid or *prevent* dysfunctional situations by selecting and modulating internal processes according to external conditions⁶. In other words, in addition to performing a constructive or interactive process that contributes to self-maintenance, the system is also capable of switching between different alternatives, tuning them, etc. according to external (or internal) changes. And this is precisely the defining characteristic of **adaptivity** recently proposed by Ezequiel Di Paolo as:

A system's capacity, in some circumstances, to regulate its states and its relation to the environment with the result that, if the states are sufficiently close to the boundary of viability,

1. Tendencies are distinguished and acted upon depending on whether the states will approach or recede from the boundary and, as a consequence,
2. Tendencies of the first kind are moved closer to or transformed into tendencies of the second and so future states are prevented from reaching the boundary with an outward velocity.

6 The dialectics that is hereby established between complexity and robustness deserves at least a footnote. There seems to be two main alternatives to explain or justify the transition towards explicit regulatory mechanism: a) the more complex the more fragile a system is and therefore additional regulatory mechanisms are required to manage the self-maintenance of that complexity or b) higher levels of robustness (specially to strong changes in certain conditions) require more complexity, and regulatory control is the form that this complexity takes. If we think of robustness just as structural stability then the simpler a system it the more robust it may become so, it seems to follow, the more complex the more fragile. This however needs to be proven. It seems at first sight that "high complexity = high fragility or low robustness" is a reasonable assumption to make but complex systems often defeat intuitive assumptions (e.g. highly redundant metabolic networks are complex but redundancy offers stability). In turn it may all depend on the definitions of complexity and fragility that one uses to accept or reject the equation. However if one adopts a wider notion of robustness to accommodate radical changes of environmental conditions so that the system needs to be sensible and able to change its structure (not just to recover the previous stability) the second option seems more appropriate. We have seen how this is not a trivial issue and how, at least in Hopfield type networks, self-organized and holistic systems are very limited on this capacity. So I am inclined to say that the issue is not really that more complex systems require more sophisticated forms of achieving robustness *because* they are more complex and therefore structurally more fragile. After all, if this were the case, one would have to answer the question of why should complexity be selected if it downplays stability. It is rather that robustness to radical changes of environmental conditions or strong perturbations requires additional mechanisms; and it is the appearance of this new type of regulatory mechanisms what make the systems more complex

(Di Paolo 2005:468)

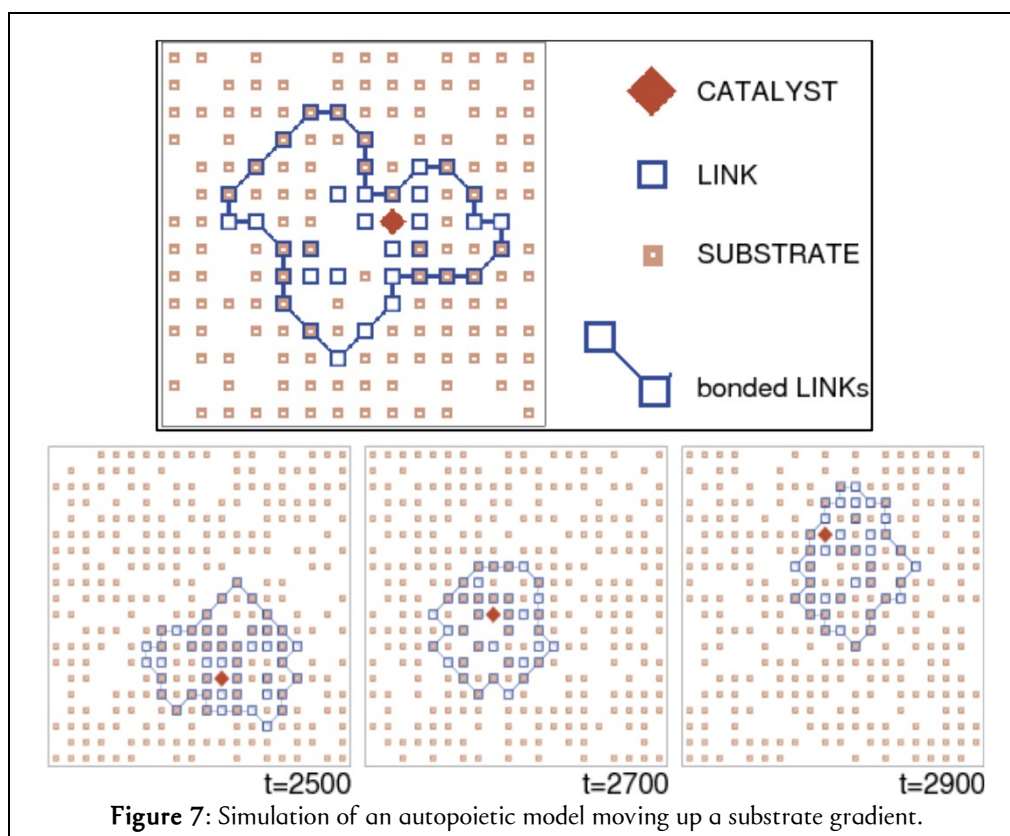
In addition, in most cases, adaptive regulation takes place not just transforming outward tendencies into inward trajectories (i.e. not just avoiding negative tendencies as suggested by Di Paolo) but actively seeking to improve the state of essential variables. Regulation takes place not just in reference to the boundary of viability but as graded and directed by a “sense of well-being” (which in non-anthropocentric terms amounts to the tendency to drive essential variables to their global optimum, if any).

However, the appearance of adaptive subsystems does not imply that the basic autonomous organization is not relevant any more. That something is adaptive implies that it is so in relation to a norm or goal and, as I have previously argued, there can be no sense of normativity or adaptation/maladaptation without an autonomous grounding. Thus, our sense of adaptivity does not depend on the selective history of a population of organisms (Mayr 1963, Millikan 1984), but on the internal organization and the conditions of self-maintenance that their organization defines.

Adaptivity is a capacity that all present-day living organisms have. The simplest mechanisms of adaptive regulation fall into two different kinds. One is exemplified in the “Operon” activation and deactivation of genes as a switch between metabolic pathways according to certain environmental conditions (or, to be precise, according to the effect of these environmental conditions inside the cell). The other one is constituted by a whole subsystem of biochemical pathways, not directly involved in the basic self-constructing metabolic network (as is the case of chemotactic agency in *E. coli*—that I will soon explore in more detail). But the common characteristic of both cases is that some degree of decoupling from the basic constitutive processes is required. In the first case metabolically off-line gene-strings act as instructive switches between different metabolic pathways. In the second case chemical pathways that are relatively independent of the basic metabolic-constructive cycle sustain the interactive loop.

These partially decoupled systems open the possibility to consistently speak in terms of an internally generated mechanism for **normative regulation**, as defined in the previous chapter. As Di Paolo sustains, in non-adaptive self-maintaining systems (as our mBAS model) the natural distinction between self-maintenance and disintegration is not yet “accessible” to the system unless it is also able to regulate itself with respect to the state of essential variables (Di Paolo 2005). Whereas in pre-adaptive systems robustness of self-maintenance depends on the attractor landscape defined by the constitutive processes, adaptive systems have the capacity to modulate (internally and interactively) the trajectories of the essential variables of its constitutive organization. I am here talking about adaptive regulation: the capacity to distinguish and compensate tendencies requires that whatever makes a distinction

and generates a compensation be differentiated from what it distinguishes and acts upon. This implies a kind of **control mechanism** (composed by receptors, effector and a transformation function between them) by which the system distinguishes and compensates tendencies that (if no compensation is carried out) would bring the system outside viability boundaries or far from an optimal state. Note that the very decoupling of the adaptive subsystem is what permits to speak in terms of detection and, ultimately, of control mechanisms. Whereas agency in the previous chapter was made dependent on the generation of constraints and modulation of boundary conditions (somehow centred on the notion of action) adaptivity permits to introduce a language of sensing or, at its most elementary level, of detection.



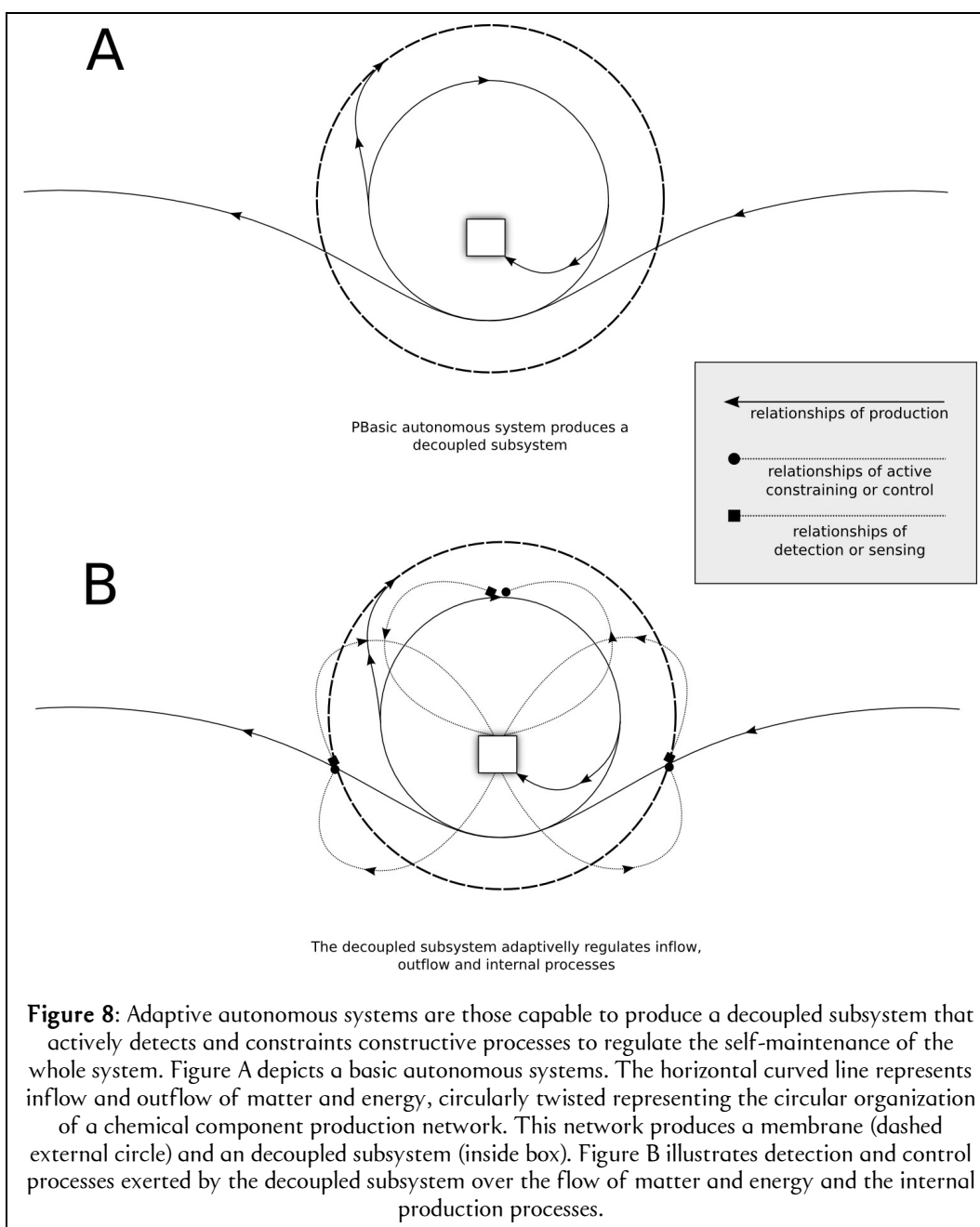
We can highlight the new situation that adaptive decoupling brings forth in contrast with a recent conceptual simulation model by Ikegami and Suzuki (2008). They extended Varela, Maturana and Uribe's pioneering model (1974) of an autopoietic automaton. The original model consists on substrate molecules that moves freely through space, a core catalyst that transforms substrate molecules into membrane molecules, and the membrane molecules

that that can link to each other and encapsulate the catalyst. Only local rules govern the system in a 2D arena and the resulting behaviour is meant to be a model of an autopoietic minimal system: the catalyst produces and repairs the membrane whose molecules degenerate progressively. Ikegami and Suzuki extended the model to include a new set of rules that specify the movement of the membrane molecules. As a result, given an appropriate degeneration rate for membrane components, the autopoietic system is able to move in space following up a substrate gradient in the environment (see figure 7). In this case, the system is in fact improving its boundary conditions (the available substrate molecules) but this movement is the result of the constructive cycle itself. It is the result of a differential growth and membrane transformation due to the presence of more substrates on one side of the protocell.

The central idea of decoupling is that the system generates a set of mechanism that do not participate directly on the metabolic cycle and become thus free to deal with the environment to change some internal or external conditions for the benefit of the system. Molecules (e.g. enzymes) may act outside the immediate logic of self-production but contribute indirectly to it (e.g. through the selective modulation of the intrinsic dynamics of metabolic processes). The language of signalling is often introduced when modelling adaptive regulatory processes. A molecule that binds to a decoupled mechanisms is considered a signal when it does not have a *direct* effect on the self-organized network of component production processes but can nevertheless trigger a cascade of processes that induces a subsequent functional constraint over the metabolic-constructive machinery (or back to the environment to produce additional changes that will in turn affect the metabolic core).

With this mechanisms at hand, we can now move beyond the implicit teleology of BAS and naturalize the claim that some interaction or process is detected as bad or good *by* and *for* the very system (and not only *by* and *for* the external observer); i.e. this well or bad functioning *for* the system is objective because it is detected and compensated *by* the system, in an effective, functionally integrated way. Thus, adaptive systems are an instance of **explicit teleology** since in addition to *having an intrinsic norm* (due to their basic autonomous organization) they also *act according to this norm* generating global constraints, over their minimal basic organization, so that a regulatory control emerges operating upon the basic self-maintaining and self-producing infrastructure.

Thus, I am talking about two dynamical “levels” in the system: the constitutive level, which ensures the ongoing self-construction of the system; and the (now decoupled) adaptive subsystem, which regulates the dynamical rate of the former by means of detectors, effectors and a transformation mechanism between both. This way, independence of low-level functionality (constructive processes) from high-level variation (state of regulatory mechanism)



is achieved. Thus, free from the lower-level constraints, higher level variation can be left to “spontaneous” dynamics, provided that a further coupling is established linking higher level states to lower level ones in a functional way so that this higher level variation (or freedom) is in turn functionally recruited to

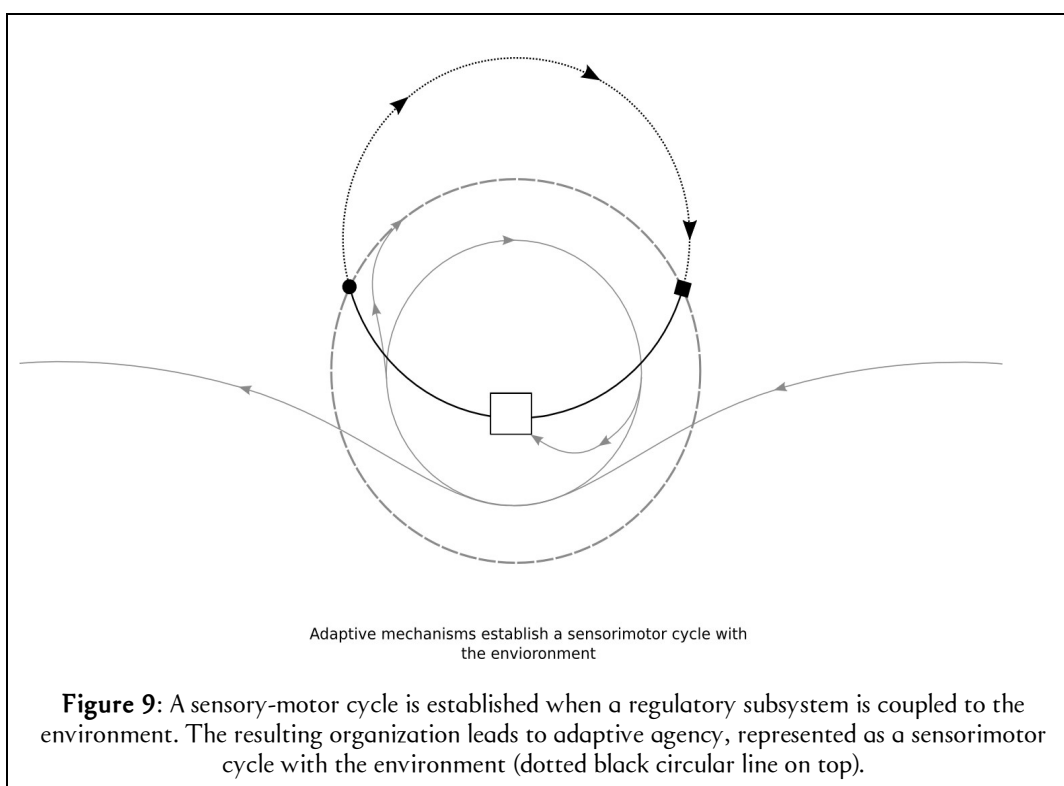
serve lower level functions. We can now move to interpret within this systemic framework some well known mechanisms of adaptive regulation.

For our purposes adaptivity takes two basic forms depending on whether the mechanisms of regulation take place at the constructive level or at the interactive one. In the first case, internal or external perturbations are compensated by adjusting or transforming *constructive* processes; such is the case of the Lac-Operon mechanism in *Escherichia coli* bacteria. The normal metabolism of *E. coli* depends on the availability of glucose. But when the levels of glucose become very low and another sugar (lactose) is abundant, a mechanism called Lac-Operon is activated: the detection of lactose triggers the expression of certain inactive genes that in turn instruct a new metabolic pathway which metabolises lactose⁷. This metabolic mechanism is adaptive because it implies a regulation of the internal constructive processes according to the detection of a certain environmental condition (the presence of lactose and the absence of glucose) which jeopardizes the self-maintenance of the system if no change of metabolic pathways is carried out (see figure 8-B). The decoupled mechanism is provided by the selective activation of the gene which carries off-line processes with respect to metabolic dynamics (in the sense that it does not directly participate on the network of reactions that transforms matter and energy for the system). However, in such cases the system-environment relationship remains unchanged, adaptivity is achieved transforming internal structure while preserving overall viability: i.e. switching to a different metabolic pathway (constructive cycle) involves no change of interactive processes. It is the second type of adaptivity, achieved by means of the adaptive control of interactive processes, what becomes of fundamental importance for our morphophylogeny of agency

3. RECURRENT ADAPTIVE INTERACTIONS: MINIMAL AGENCY REVISITED

The second form of adaptivity turns out to be of particular interest because it gives rise to **adaptive agency**: *adaptation is achieved through the regulation of interactions with the environment*. In such cases, interactions become functional in virtue of the changes induced *outside* the system or more specifically on the relationship between the system and its environment. For instance, secretion of poisonous chemical molecules under contact with an external force

⁷ This is of course a simplified version of the real mechanism. More than activated, the expression of lac operon gene complex (that encodes three enzymes required for metabolising lactose) is repressed. When lactose is present in the environment it binds to the repressor (inducing a conformational change that unlocks its repressive capacity) and transcription of lac operon starts (but at a very low rate). It is under the absence of glucose when an activator protein complex (CAP-CAMP) increases the transcription of an activator protein that the transcription of the lac operon increases.



(probably correlated with the presence of a predator) might be considered an instance of adaptive agency: decoupled from direct metabolic function the detection-secretion mechanism contributes to the self-maintenance of the system. We have here an instance of a functional action in virtue of the changes induced on the environment, i.e. *outside* the system.

Now we come across a problem with respect to the minimal model of agency I introduced on the previous chapter. Active waste product pumping through the membrane, strictly speaking, may not deserve the name of an action since the exchange becomes functional in virtue of the change achieved *inside* the system (the adequate decrease of internal concentrations). In this type of interactions there is no sense of environment for the agent, for whom the higher concentration induced into the environment as a result of active transport is completely “insignificant” (i.e. it has no effect on the system). In other words, the equation “boundary conditions = environment” does not really hold, hence the equation “active control of boundary condition = action in the environment” doesn’t work either. Otherwise, we would have to consider phenomena such as sweating, changes of skin in specific seasons, etc. as genuine examples of agency. In this sense, Ruiz-Mirazo and Moreno’s requirement for minimal agency falls probably too short to characterize full-

fledged agency (even in its minimal form). But, that the actively regulated process is functional in virtue of the changes induced *outside* the system may also fall too short. The example of the secretion of a poisonous substance described above is certainly an adaptive action. But a single isolated action may not be enough to qualify an agent: there is no concatenation of actions, no real *acting* within an environment but rather a single, local, *re-acting* to a specific condition which, once again, has no further effect detectable by the system⁸. I shall, however, introduce an additional requirement: that the transformations induced on the system-environment relationship become also functional inputs for achieving the adaptive regulation (Bourgine & Stewart 2005⁹). With this requirement interactive processes become a *cycle* controlled by an adaptive subsystem (see figure 9). We can think of chemical agents whose interaction cycles are based on secretion and detection of different substances as the simplest instances of this revised conception of minimal agency. Think, for instance, on the type of “chemical fight” that is established between plant cells and fungi. However, bacterial chemotaxis is, by far, the most studied case of minimal agency, where interactive processes become a cycle through motility.

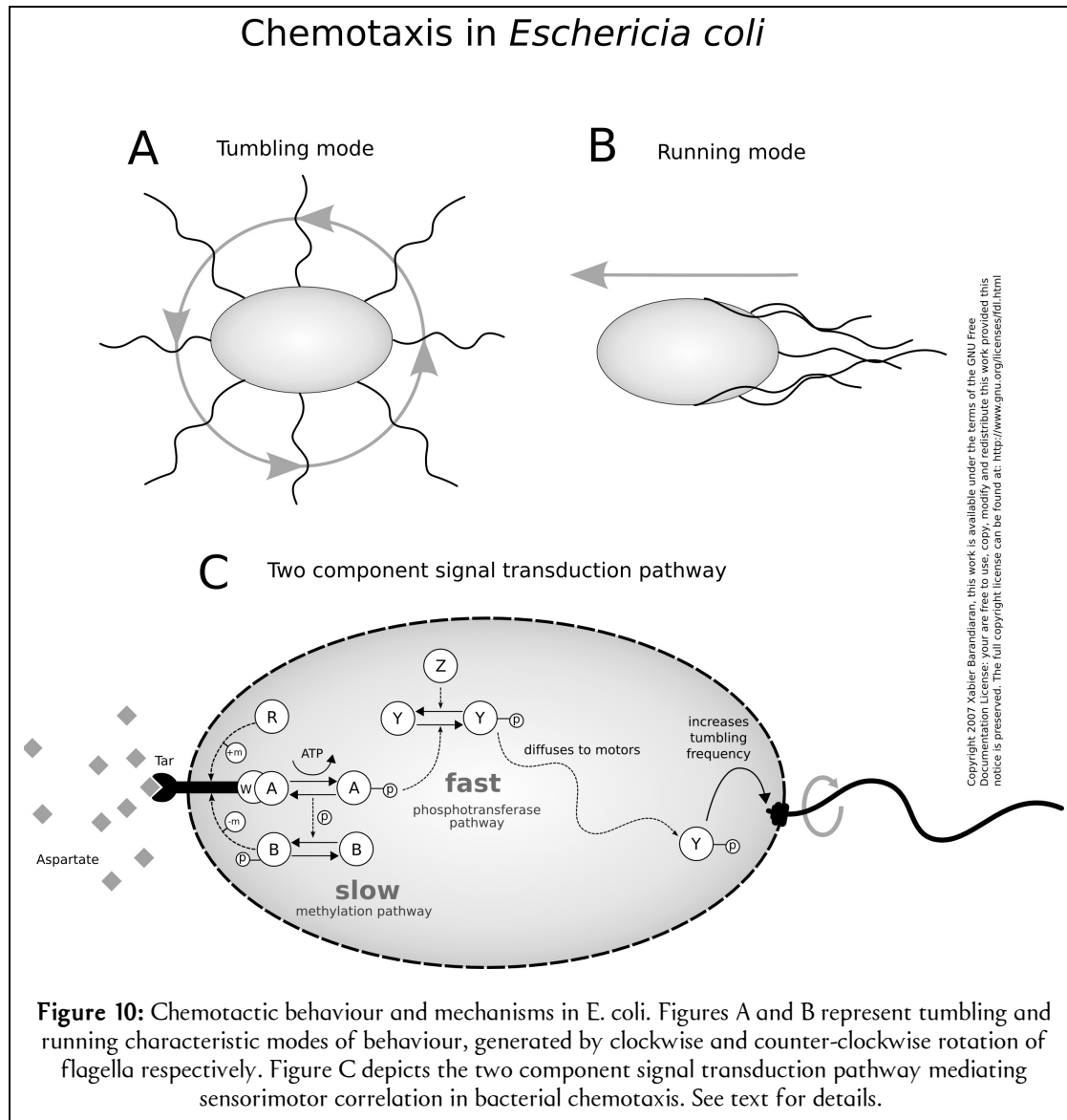
4. CASE STUDY: BACTERIAL CHEMOTAXIS

Bacterium *Escherichia coli* is a well studied biological model of chemotaxis and many other bacteria share similar mechanisms¹⁰. The behavioural analysis of bacterial motion has shown that up or down gradient directional movement is achieved through the differential combination of two basic types of movements: tumbling and running. When *tumbling*, flagella move clockwise and the directional movement changes to random turning (figure 10-A). When *running*, bacterial flagella move counter-clockwise generating a coordinated movement that propels the bacteria in straight line (figure 10-B). In a neutral medium, where no attractant or repellent gradient is present, tumbling and running modes alternate in a ratio of 0.1s of tumbling followed by a few seconds of running. Tumbling during 0.1s produces a mean change of direction of 60° so that the overall movement in a neutral environment is that of a random motion. However, when engaged in running motion if the gradient of attractant is positive (i.e. if the bacterium is swimming up the gradient) the

8 It is of course a question of degrees, the important point here is not to demarcate a clear-cut boundary for the appearance of genuine agency but rather the bottom-up inclusion of additional requirements to characterize agency in more detail.

9 Bourguin, P. & Stewart, J. (2004) Autopoiesis and Cognition. *Artificial Life* 10: 327–345.

10 Duijn et. al.'s reconstruction of the mechanisms underlying chemotaxis in *E. coli* (Duijn et. al. 2006) was a very valuable first approximation to this case study and part of this section was originally based on their explanation and reading of some of the references cited on their paper.



tumbling frequency decreases achieving a net behaviour of approach towards the higher concentration of attractants. On the contrary, if the gradient is sensed to be negative the tumbling frequency increases, thus leading to a higher probability of changing the direction of movement until the up-gradient direction is recovered. With repellents, the opposite behaviour is observed.

The mechanisms underlying the chemotactic adaptive motion in *E. coli* is a two-component signal transduction system (TCST hereafter). Through this

mechanism bacteria are able to “compare” current and past concentrations of different, attractant or repellent, chemicals so as to modulate the tumbling frequency with the result described above. Note that gradients cannot be directly detected, only concentrations. How does the gradient sensing work? As its name says the TCST consists of two transduction pathways (see figure 10-C). The first one, the *phosphotransferase* pathway, directly involves detection of attractants or repellents acting at the time-scale of milliseconds while the second, the *methylation* pathway, provides feed-back to the first pathway at a time-scale of seconds. The first pathway can be seen as a full sensorimotor connection chemodynamically linking transmembrane receptors to flagella. There appears to be a single signalling unit in most cells, each signalling unit being composed of several thousand receptor filaments of which 90% (Alexandre & Zhulin 2001) correspond to five receptor types (Tar, Tsr, Trg, and Tap). The structure of the chemoreceptors is that of a transmembrane protein filament to which attractants and repellers get attached at the extracellular end (Stock & Levit 2000) transforming, in turn, some properties of the chemoreceptor filament at its intracellular side. Inside the cell, chemoreceptors form complexes with CheW and other compounds together with Kinase CheA (pictured with just letters “W” and “A” in figure 10-C). CheA phosphorylates thanks to ATP and transfers phosphoryl to CheY. CheY-P (i.e. phosphorylated CheY) is then free to diffuse along the cell and binds to flagellar switch proteins increasing the frequency of tumbling behaviour (the mechanisms if pictured in figure 10-C). CheZ enhances dephosphorylation of CheY so it recovers its dephosphorylated state, ready to be phosphorylated again by CheA. The normal functioning of this pathway gives rise to the frequency of tumbling and running defined above for neutral environments; i.e. when no attractant or repellents is present the phosphorylation and de-phosphorylation rate of CheY is such that tumbling occurs every few seconds.

When an attractant gradient is present, binding of attractants to the receptor reduces phosphorylation of CheA thus reducing phosphorylation of CheY which in turn reduces the frequency of tumbling. Left alone the phosphotransferase pathway reduces the tumbling frequency in the presence of attractants thus increasing the probability that the bacterium maintains its straight motion (i.e. its running mode). However, as noted above, that the presence of attractants be high does not mean that the bacterium is approaching the higher concentration medium, it could equally be the case that, close to the higher concentration area, the bacterium be running down the gradient. To prevent this, under a high binding of attractants, receptors return to their original value (a process called “adaptation”). This resetting of the receptors is carried out by the methylation pathway. The sensitivity of the receptor is reduced if concentrations are very high, so that receptors become sensitive, not to the absolute value of the attractant concentration, but to the

relative value, making thus possible to sense the gradient. How is this achieved? The working of the methylation pathway involves an additional set of molecules (CheB, CheR, etc.) and their intricate interactions with the receptors and with the phosphotransferase pathway. However, the mechanism of “memory” or “comparison” can be explained in a simplified manner. I mentioned before how the methylation process is in charge of resetting or “adapting” the receptor. Now, since the methylation process is slower than the phosphorylation process if the concentration is progressively higher (i.e. if the bacteria is running up gradient) the rate of “adaptation” cannot compensate the decreasing rate of phosphorylation (remember that attractants reduce the phosphorylation). As a result, the phosphotransferase pathway becomes sensitive not to the total concentration of attractants but to its increase, thus provoking the tumbling frequency to decrease as the gradient is navigated upward. Simulation models have shown that the time-scale of the tumbling frequency must be smaller than that of adaptation (regulated by the methylation pathway) and higher than sensing (regulated by the phosphotransferase pathway) for this chemotaxis mechanism to work (Inoue & Kaneko 2006).

There are a number of interesting features of this mechanism that deserve a careful attention from our analysis of adaptive agency. First, the chemotactic system is decoupled from metabolism, it does not directly participate on any metabolic function (constructive cycle); only ATP energy is used to initiate phosphorylation and, obviously, to move the flagella. ATP is an energy currency produced by the system and required by the sensorimotor system thus satisfying the energy-based causal asymmetry condition for agency. However, and this brings us to the second interesting feature, the system is not always totally metabolism independent: which attractant gradients are followed and which are not is not something that happens independently of internal metabolic requirements. In a recent review on chemotactic sensing in bacteria, Alexandre and Zhulin (2001) champion the widespread assumption that bacterial chemotaxis is metabolism-independent, a view popularized by Julius Adler in 1969. The received view narrowed the research focus on metabolism-independent information flow, disregarding sensing mechanisms other than transmembrane chemoreceptors (i.e. reconstructing sensorimotor interactions independently of intra-metabolic regulation). There is evidence, however, for a number of astonishing phenomena linking chemotaxis to metabolic regulation and needs¹¹; some of them include

- Stronger chemotaxis is observed to the type of sugar present in the medium in which bacteria have grown.
- Bacteria are specially sensitive after starvation.

¹¹ Not all of them have been reported for all types of bacteria but, for instance, *A. brasilense* has been reported to show many of them (Alexandre *et al.* 2000).

- Mutants that cannot metabolize a specific attractant (due to metabolic effect of the mutation) do not show chemotaxis to that particular sugar (although the mutation had no direct effect over the receptors or the TCST).
- Presence of a metabolizable chemical prevents chemotaxis to all attractants.
- Inhibition of the metabolism of a chemical stops chemotaxis to that particular attractant (and only to it).
- Some bacteria show a direct correlation between the efficiency of a chemical as a growth substrate and sensitivity to it.

Which are the mechanisms involved in metabolic-dependent behaviour? Alexandre and Zhulin (2001) show how several signal transduction mechanisms may be linking chemotactic pathways and receptors with different metabolites and with electron transport pathways (which canalizes energy within the cell). As a result, metabolizable substrates and electron receptors become attractants and bacterial chemotaxis gets modulated by internal metabolic/energy requirements:

Measuring changes in the electron transport system is the most effective way to monitor energy metabolism in general and concentrations of available metabolizable substrates in particular. In energy taxis [movement regulated by internal energy requirements], the strength of the response depends on the presence of both a substrate and an appropriate electron acceptor. Should the optimal concentration of a substrate decrease, the cells will execute "chemotaxis", moving up the chemical gradient. Should the optimal concentration of an electron acceptor (for example, oxygen) decrease, the cells will execute "aerotaxis", moving up the oxygen gradient. Thus, via energy taxis the cells seek optimal balance between the oxidizable substrate and oxygen in order to generate maximal energy. (Alexandre & Zhulin 2001:4685)

Thus, the appearance of adaptive agency implies the emergence of two coupled cycles: (i) a sensorimotor cycle with the environment carried out by a decoupled mechanism that regulates interactive processes (the TCST) *and* (ii) the metabolic cycle defining the basic autonomous organization of the system. The coupling between these cycles is, as Alexandre and Zhulin have shown, *via* energy and metabolite signal transduction mechanisms that regulate the interactive cycle (selectively inhibiting and activating chemotactic processes) in relation to internal metabolic/constructive requirements and environmental opportunities (featured in terms of attractant/repellent binding).

5. MOTILITY: THE ENACTION OF AN ENVIRONMENT

Motility is the capacity of an autonomous system to adaptively *direct* the cycle of sensorimotor correlations through self-generated displacements in the environment. Detection-of and functional response-to environmental changes becomes, in the case of adaptive motility, a “sensorimotor” cycle, whose viability is strongly affected by size-time organizational constraints and affects, in turn, new kinds of the space-time relationships between system and environment. But more importantly for us, motility somehow implies the *enaction* (active generation) of an environment out of sensorimotor correlations. Note that until motility appears, there is no sense of environment for the system other than what is immediately continuous with its surface and externally given: i.e. from the point of view of the agent-environment relationships we cannot speak about spatial properties (nor objects, perception, distance, etc.) but just direct immediate detection of some surface conditions (which amounts to changes in boundary conditions). mBAS are affected only by what is directly functional or dysfunctional for themselves and, conversely, they can only affect what belongs, or is physico-chemically linked, to their constitutive cycle. On the model of mBAS described in the previous chapter, for instance, the permeability of the membrane is directly modulated by the amount of peptides, which is directly correlated with the concentration of certain metabolites, that in turn depends on the concentration of precursor molecules. In turn, the permeability of the membrane directly regulates the boundary conditions (surface tension and concentration gradient) for self-maintenance. In other words in minimal autonomous “agents” boundary conditions are modulated as “controlled variables”, there is no mediation of such control, no decoupled sensory and effector mechanisms.

Things are qualitatively different in adaptive systems, even more for the case of adaptive motile agents. In bacterial chemotaxis, for instance, the system is changing its boundary conditions (the sugar concentration gradient between system and environment) through non controlled variables (the concentration of sugar on its environment) and it is doing so by means of mechanisms that do not directly metabolize sugar nor do directly participate on the logic of metabolism. There is no direct possibility for the system to increase the sugar concentration on its immediate environment. Obviously it could pump sugar to the environment and change it, but this will not solve the metabolic need for sugar molecules, on the contrary it will increase it. We know that the solution to this problem is to move up the environmental sugar gradient but it is difficult for the observer to stop projecting her own phenomenology to the object of study. We see a bacterium moving on its environment and we immediately project our sense of environment to the phenomena under study. But from a bottom-up methodological perspective the chal-

lenge lies on reconstructing this system-environment relationship from within the organization of the system, revealing what the enaction of an environment consists of (something that, to a big extent, remains implicit and unconsciously constitutive of our own experience). The passage from the interaction processes exerted by minimal (only structurally stable) autonomous systems to adaptive motility is the passage from immediate modulation of boundary conditions to mediated regulation by means of decoupled sensorimotor correlations (decay of tumbling probability correlated with increase in detector binding). In turn, these correlations can only be achieved via the relative position within a spatial embeddedness. Achievement of adaptive regulation is open to spatial dimensions; explored and exploited by motility. What was formerly a relatively immediate (i.e. non-mediated) boundary condition becomes a temporarily extended, regulated and delayed “environmental” condition that is, in turn, partly a consequence of the motile activity of the system. A crucial requirement for the origin of an environment is the capacity of the system to break the simultaneity of its dependence on immediate boundary conditions, so that a (temporal) distance is opened between constructive needs and their satisfaction. This temporal distance, in turn, is spanned through movement in space. It is only when an autonomous system can detach from its immediate dependence on boundary conditions that an environment really starts to open up.

Yet, often, this detached space conforms an spatial environment in relation to the organisms only for us, external observers. All that is accessible to the system is nothing more than the temporal delay of some correlations that are actively maintained by its own activity, for which, we know, the spatial dimension of its embodiment is a necessary precondition. This temporal distance, as some readers will have probably noted, is already present in non-motile forms of adaptivity (like the Lac-operon mechanisms explained above). So... what is the real contribution of motility to agency (and more generally adaptivity) as a keystone towards cognition? The crucial difference can be grasped through the notion of *dealing* or *coping* within a domain of interactions. Whereas the internal adaptive processes like the Lac-operon involve the detection of boundary or external conditions (e.g. presence of lactose and absence of sucrose), agency brings forth the possibility of coping or dealing recurrently, circularly, with these conditions and spanning them on the form of an environment (a domain of interactions and flexible sensory-effector correlations). And, it is precisely on motility that we find the most clear and powerful example of the generation of a domain of coping. This is mostly due to the fact that, in motility, for any change of motor activity there is a corresponding change of both surface sensory states *and* metabolic boundary conditions. But the spacial environment that we take to be so characteristic of our own agential phenomenology and, furthermore what we inhabit as a “world”, will have

to wait its turn for a full-fledged unfolding. Importantly, the seeds for the emergence of this spatially structured environment *for the agent* are already in place.

One of the most important precursors for the emergence of a world of interactions is precisely the phenomenon of decoupling, which permits and interactive “liberation” from the instantaneous demands of the dissipative nature of BAS¹². I would like to highlight the crucial role played by the decoupling of interactive and regulatory processes. The form in which the sensorimotor architecture is embedded on the constructive cycle is not in virtue of the intrinsic reactive capacities of the molecules but by the mechanical shape of the sensors and motors (which is in turn produced or fabricated by the whole organization of the bacterium). When aspartate binds to tar receptors it is not being metabolised (i.e. it is not directly affecting the dynamics of the constructive cycle) but triggers a set of changes on the intrinsic phosphorylation chemical dynamics that in turn affects the probability of flagellar rotation. In turn, it is the dynamics of the two component signal transduction pathway, coupled to the environment as a unitary mechanism, that functionally contributes to the existence of the bacterium. The “sensing of the gradient” cannot be dissociated from the generated movement; if the bacterium was to stay immobile there would be no gradient for it, no difference of concentrations to navigate. But movement permits the bacterium to change the state of its receptors so that this change can be appropriately correlated with more movement (decrease of tumbling frequency when attractant receptor signals increases). Thus, it could be said that, in the case of bacterial chemotaxis, movement somehow transforms temporal properties (the difference between methylation decay and the phosphorylation in the TCST mechanism) into spatial properties of the environment¹³. We get a full sense in which effector changes (changes in flagellar rotation) become functional, not directly and immediately for the self-maintenance of the system, but for the sensorimotor correlations that permit motile adaptivity. It is the sensorimotor cycle itself that becomes functional, not a specific reception of attractants or a particular flagellar rotation. If one is to depict any functional relevance of a particular receptor or flagellar state, that could only be achieved along the tem-

12 Rosslénbroich (2005) has argued for a major evolutionary tendency towards autonomization from the environment as a trend that runs in parallel with an increasing complexification of life-forms.

13 This co-dependence between internal reaction rates and spatial properties becomes evident when one tries to model bacterial chemotaxis. Bray and colleagues’ make it explicit: “Biochemical changes in the chemotaxis pathway are context specific and make sense only when one knows the chemical environment of the bacterium. A complete description therefore requires one to relate temporal changes in the concentrations of signaling molecules to spatial changes in attractants and repellents. With this in mind, we embedded a detailed, quantitative simulation of the temporal changes in chemotaxis biochemical reactions within a graphical representation of swimming bacteria.” (Bray et. al 2007:12).

porarily and spatially extended context of reactions and correlations in which such transient states take place.

Two forms of coupling appear intertwined: the functional coupling by which sensorimotor interactions contribute to the basic autonomy of the system *and* the internal dynamical coupling established between metabolism and sensorimotor cycles in metabolism-dependent chemotaxis. This intertwined regulation (motility regulates boundary conditions for metabolism, metabolism regulates motility) brings forth a genuine kind of agency in which metabolic normativity is extended into environmental interactions, creating a first sense of “meaningful” environment for the organism. The full implications of this mode of adaptivity are still to be explained. Note that I have used quotation marks on “meaningful” where many have claimed that a full naturalized sense of meaning, sense-making or a cognitive language can already be used here or even in earlier forms of agency (Varela 1992, Weber and Varela 2002, Stewart 1996, Jonas 1968, etc.). As I shall argue later, one should be extremely cautious to use such mentalistic terminology at this point (even after some crucial transitions, that are still to come, are complete). However, we have already made a significant advance in our morphophylogeny of agency and this last example of bacterial chemotaxis has led us to the concept of motility and sensorimotor coupling. Motility is a characteristic example of the mediation of adaptive control that defines agency. Not only does it introduce a qualitative change on the system-environment distinction (with implications for individuality, normativity and causal asymmetry) but it will crucially define the path towards adaptive behaviour. We can expect the emergence of cognition to be found on some kind of expansion and organization of such mediation. However, this expansion and organization will only be possible through a set of transitions defined by organizational bottlenecks and consequent innovations that unicellular organisms need to undergo in order to accommodate it.

6. ORGANIZATIONAL BOTTLENECKS OF MOTILE AGENCY IN UNI AND MULTICELLULAR ORGANISMS

So one could think of bacterial chemotaxis as the origin of increasingly sophisticated agency, potentially leading to higher cognitive capacities. We know that this wasn't the case. Bacteria have been evolving for millions of years without achieving a much more complex agential motility than what the case study of chemotaxis has shown. And the long evolutionary history of bacteria suggests that there are important organizational reasons behind this limitation (Moreno & Lasa 2003, Moreno & Etxeberria 2005). In fact we can observe a huge evolutionary jump from this form of agency (and some similar

eukaryotic modes of motility) to the evolution of animals, where the path towards increasingly complex adaptive agency is, so to speak, recovered. The appearance of multicellular organization with certain characteristic bodyplans and, particularly, the invention of the NS are the most important transition towards cognitive capacities. However, an analysis of the details that made this jump necessary are important since they hide some necessary conditions that make cognitive capacities impossible throughout the intermediate steps, revealing some important aspects of cognition-as-it-could-not-have-been.

Prokaryotic potentiality to develop more complex forms of agency is severely limited. On the one hand the bottleneck appears because the level of complexity that the adaptive subsystem can achieve (within the biochemical medium), without severe interference with metabolic processes, is very limited. This is due to the limitations of decoupling that adaptive subsystems can achieve within the biochemical organization of prokaryotes. On the other hand, as the size of the organism increases, the fast and plastic correlation between sensor and effector surfaces becomes harder (or even impossible in multicellular organisms) due to the slow velocity of diffusion processes. This fundamental problem continued with unicellular eukaryotic organisms but they managed to increase in size (up to 10.000 times) without loss of agential capacities, revealing some interesting new solutions to the problem posed by motility at larger sizes (and also a number of additional limitations).

The most sophisticated unicellular adaptive mechanisms that appears in eukaryotic cells are made possible by a number of organizational features that became crucial: a) the presence of nucleated DNA with an increased regulatory power (as I shall analyse in more detail soon), b) membrane-bound compartments that become functional modules of metabolic organization and c) the appearance of mechanically organized microtubular fibres and filaments regulated by proteins. These three organizational mechanisms that are found in eukaryotes made possible the combination and channelling of molecular self-assembly and chemical dynamics in a much more sophisticated forms than what the previous prokaryotic forms permitted.

First, the nucleated DNA permits the development of a much more complex internal organization than what prokaryotes could ever achieve. This is probably due to the fact that the complexity of an organism is related to the amount of non-coding DNA. Introns are an important part of this non-coding DNA, which are spliced up from mRNA after transcription and before translation. As John Mattick notes:

Because bacteria lack a nucleus, transcription and translation occur together: RNA is translated into protein almost as fast as it is transcribed from DNA. There is no time for intronic RNA to splice itself out of the protein coding RNA in which it sits, so an intron would in most cases disable the gene it inhabits, with harmful consequences for the host bacterium. In eukaryotes, transcription occurs in the nucleus and translation in the cytoplasm, a separation that opens a win-

dow of opportunity for the intron RNA to excise itself. Introns can thus be more easily tolerated in eukaryotes. (Mattick 2004:63).

Now the role of introns is associated with genetic regulation which, in turn is required for developmental and organizational complexity to occur within the cell (see also Taft, Pheasant & Mattick 2007). Thus a further level of decoupling (between transcription and translation) permits a higher level of organizational complexity throughout the mediation of genetic regulatory networks (made possible by introns).

Second, internal membrane bounded compartments permit a higher level of decoupling by isolating some reactions from the internal milieu, so that interference between chemical pathways is avoided. This way some limitations of prokaryotic forms, due to potential interferences between chemical pathways, could be at least partially overcome. Mitochondria are a characteristic example, for which, Lynn Margulis has proposed a symbiotic origin and is now widely accepted that they most probably originated on ancient bacteria (Margulis 1992). Mitochondria, which are in turn internally compartmentalized, fulfil a number of key functions; among them one of the most important is that it provides a source of ATP for the rest of the cell.

Finally, the use of microtubules (whose appearance was probably only made possible by nucleated DNA) permitted a wide management of movement at much bigger sizes, due to the mechanical management of the internal structure that microtubules and similar fibres permit, forming what is called a cytoskeleton. Microtubules also permit a channelling of internal chemical pathways which permits a more complex spatial organization of chemical dynamics. As a consequence, eukaryotic cells can have a volume of 3 orders of magnitude bigger than any prokaryotic cell and, at the same time, show similar motile capacities. The increase in size permits a much more complex internal organization but finds also limitations of its own. The volume that needs to be moved becomes much higher, diffusion processes are slow and integration of sensorimotor pathways becomes increasingly complicated¹⁴.

Interestingly, the most efficient organization for a quick coordination of sensorimotor responses in big unicellular organisms occurs through the membrane itself, mediated by changes of electric potential that propagate and modulate the beating of cilia. This is the case of *Paramecium* and also of *Stentor roeseli* one of the biggest unicellular organisms (it can measure up to two millimetres) that shows a very sophisticated behaviour, including certain forms of learning. The motile and contractile capacity of *Stentor* is facilitated

¹⁴ Apart from ciliated and flagellar movement and body contraction by means of filaments and microtubules one of the major strategies displayed by unicellular organisms to coordinate movement is the reversible colloidal change between viscous and liquid states along the body. This is the strategy that amoeba use changing the viscosity of their internal medium in order to change their shape. But it can intuitively be seen how this strategy is also prone to severe limitations on its capacity for quick and versatile complexity growth.

by a set of longitudinal microtubular fibres and dense bundles of filaments called myonemes (Huang & Pitelka 1973). However, there is not much room for increasingly complex sensorimotor coordination even with this type of unicellular system. On the one hand, we find the problem of motile coherence: this is either quick and local, or when globally coordinated, it requires a full change of morphology and a “costly” reconfiguration of internal microtubular and filamentous structures. On the other hand we find the problem of fast integration and complex sensorimotor transformation, illustrated by the difficulties of achieving it through changes on ion permeability of an otherwise almost homogeneous membrane. But despite their own limitations, it is eukaryotic organization that will permit a further and crucial transition on the morphophylogeny of organismal agency. The genetic regulatory power that nuclearised DNA confers to eukaryotic cells, their internal compartmentalization and their microtubular management will permit the appearance of multicellular organization with an individuality of its own as integrated systems capable of establishing adaptive motile interactions with their environments.

The appearance of multicellular organization is a complex and controversial issue in itself. Which specific factors made it possible is far beyond the scope of this work. It will suffice however to extract some of the most important consequences of this transition:

- The extra cellular matrix, together with the formation of an outer cell layer (the epidermis), involves an additional separation of the organisms from immediate variations on external conditions, creating an internal environment for component cells; which are now “save” from strong adaptive requirements that the external environmental fluctuations require (Rosslénbroich 2005).
- Save from external fluctuating conditions, the extracellular matrix provides a milieu for collective regulation, signalling and individual differentiation: “Regulation and differentiation, and thus contingency of cellular functions, is mostly introduced by imposing inhibitions on conserved reactions, only allowing them to become activated under certain conditions and thereby integrating them into the regulated networks” (Rosslénbroich 2005: 250).¹⁵
- The processes of differentiation, specialization, migration and communication that the intracellular matrix permits, together with genetic regulation (that eukaryotic nucleated DNA permits), opens up a the possibility for a process of *construction* of the organism (development)

15 Unicellular organisms can also coordinate their activity through chemical signalling. The Slime mould's capacity to coordinate unicellular lifeforms into an organismic whole is a fascinating example. But the reason of this fascination is the mark of its nature as a limit case. Signalling pathways need to be “secured” and stabilized for inter-cellular coordination and the coordination shown by Slim moulds is probably the maximum that can be achieved in terms of collective pattern formation at the size of unicellularity.

as a different and differentiated process from reproduction (which is left for specialized germ cells).

- In turn, development permits that organization be the result of the temporal unfolding of the collective and interactive (system-environment) activity of multiple autonomous systems.

The potential of multicellular organisms to achieve more sophisticated forms of adaptive agency is apparent, but these forms will need to find their way. The above features do not but open a possibility space that needs to be traversed. At the same time the appearance of bigger and multicellular forms of life, as it did with the eukaryotic unicellulars, defined a new ecological space to be explored at the higher scale of size and different modes of organization. But, within this space of possibilities, the scope of motility based living forms was, again, severely limited, at least if the previous mechanisms for motility were to provide the core architecture to support agency. This is due to the increase in size that multicellularity and differentiation permitted (Bonner 2003) and the exponential growth of the problems encountered by earlier mechanisms when transferred to bigger sizes.

Three main bottlenecks appear as defining the difficulties of achieving more sophisticated forms of agency in uni and early multicellular systems: (i) the increase on the distance between effector and motor surfaces (that need to be efficiently connected), (ii) the need for plastic and selective modulation of this connection and (iii) the need for an additional mechanical infrastructure to coordinated a type of body movement capable of unitary displacements. As the size and metabolic sophistication of organisms increases if these demands cannot be satisfied by the biological organization underlying agency only local responses are possible. No complex motile agency can be supported and multicellular motility cannot take off from the grounds of its most limited manifestation. Eukaryotic cells included an additional tool kit of organizational solutions to the size-motility problem but could only partially overcome the limitations of bacterial agency. Yet this innovations made possible the appearance of multicellular bodyplans and functional organization and with it the opening of new avenues for agential complexity. *Paramecium* and *Stentor*'s use of changes of electric potential across the membrane to coordinate body movement somehow anticipated the "electric solution" that was to come (the invention of the nervous system) but this had to be accompanied a by host of developmental tools to achieve a successful functional articulation and integration of sensorimotor dynamics. It is not just a problem of "processing speed or velocity" what further innovations came to solve but, as I shall explore on the next chapter, a problem of biological organization of agency in multiple dimensions.

Chapter 6: The Nervous System: origins, evolution and organization of Behavioural Agency

Individualized systems, capable of defining their own normativity and to co-specify an environment out of a previously undifferentiated surrounding, saw their interactive motile capacities limited by a number of organizational factor. The morphophylogeny of agency was somehow truncated by the material and energetic embodiment of autonomous systems whose interactive processes, at the unicellular and early multicellular forms, appear in need of a new reorganization if agency is to find its way towards cognition. An essential part of this reorganization takes the form of the nervous system (NS hereafter), a highly specialized adaptive subsystem able to support unprecedented levels of complex mediation of interactive processes. The appearance of the NS implies not just an increase on internal sensorimotor coordination in comparison with previous motile organisms but a full change of biological organization, including the relationship between metabolism and agency, internal regulation and organismal identity. Yet, this changes should not be considered as driven exclusively by the appearance of a NS but emerge as a result of new developmental tools and bodyplan transitions. As a consequence, we shall see the emergence of *adaptive behaviour* as a specific mode of agency. Yet, the very definition of a neuron, the specification of the NS as a system, its modeling assumptions, its specific mode of contribution to biological organization or the very definition of the concept of behaviour is far from trivial. Although the study of neural tissues, electric conduction patterns and molecular processes is at the core of current scientific (neurobiological) attempts to reconstruct the mind, little is said about why the NS is so important and how it was originated¹.

¹ Neurobiology and neuroscience textbooks (e.g. Kandel 2001, Kolb & Whishaw 2000, Purves et al. 2004, Squire et al. 2003) are scarce on their treatment of the early origins of the NS and remain

Throughout this chapter I shall first look at the origins of the transformations that made possible the appearance of the NS. I will next analyse the architecture and dynamics of some of the most simple forms of neural based agency found in nature (that of Cnidarians). I shall then extract a number of consequences for agency that such forms of organization permit, together with some additional bottlenecks that Cnidarian developmental bodyplans suffer. We will see how it is along the Bilaterian bodyplan that the full potentiality of the NS will be exploited, making finally possible to specify the general properties that make the activity of an embodied NS appropriate to model cognitive processes. The main goal is to characterize the nature of the NS along specific bodyplans and within the context of biological organization in order to make explicit the set assumptions and abstractions it may involve for modelling behaviour.

1. THE ORIGINS OF THE NS

Most of the molecular components (called anions) in cells are negatively charged and the hydrophobic nature of the lipid bilayer that constitutes the membrane of all cells makes it impermeable to ions. As a result, a strong osmotic pressure is created due to the electric potential gradient between the interior of the cell and its surrounding. In order to avoid osmotic bursting, the membrane needs to manage ion transport through a set of channels which possess “gates” that open in response to specific stimulus of either mechanical, voltage or chemical (ligand) nature. In addition, due to the need for keeping balanced electric equilibrium, exchange of ions (of K^+ in particular but also, and most importantly for electric potential propagation, of Na^+) is ubiquitous in all cells (even *Escherichia coli* possesses a type of voltage gated ion channel). Yet, an electric potential appears due to an imbalance between electric and chemical concentration equilibrium. Thus, every cell has an electric resting potential. Due to the properties of the ion channel gates (specially of voltage-triggered ones) the membrane of a cell can quickly polarize and depolarize, giving rise to a propagating change of electric potential along its surface. Polarization and depolarization of the membrane is a passive result of the positive feedback that Na^+ voltage gates undergo. The conducting cell needs to expend no energy on these process. Action potentials are propagated without the cell “having” to do anything in particular (other than activating ATP consuming ion pumps to recover its equilibrium after the action potential occurs).

particularly silent on addressing why is neuroscience a privileged science for the study of behaviour in the context of biological organization: why, for instance, are plant synapses (Baluska et al. 2005) not a case study for neurobiology or why the prefix “neuro-” is so privileged if behaviour extends along body an world as well as neural, glial and muscular cells.

Some of these mechanisms were already present on unicellular organisms and even served for agential purposes. This is the case of the *Paramecium* that responds to different stimuli by changing its membrane potential triggering, in turn, an action potential wave that coordinates beating of cilia along its membrane. However, it is along the internal functional specialization and developmental organization that eukaryotic multicellularity permitted (including specific detection and effector cells and the need for coordinating multicellular movement), that electric charge variability and propagation became a fundamental and increasingly complex mechanism for agency.

Epithelial conduction is the term used to refer to the propagation of a change of electric potential along the membrane of epithelial tissue cells (Anderson 1980). It constitutes one of the most simple and primitive mechanisms whereby electric changes mediate adaptive agency in multicellular organisms, with or without nervous tissues and even among plants (Simons 1981). Epithelial cells can be sensitive to local chemical or tactile stimuli triggering a change of electric potential that spreads radially and symmetrically and then decays (both temporarily and spatially). The evolutionary appearance of such mechanisms was facilitated by the fact that cells need not be specialized on sensorimotor control to begin with. Due to the complexity and importance of active ion transport across the membrane all cells, epithelial cells already had the capacity for electrically mediating between specialized sensor and effector cells that were already present on the surface of the organism (such is the case of the most primitive animals, the sponges). Thus, as suggested by Albert (1999), while keeping their previous function, such as protecting the interior of the organism through the formation of one or more epithelial tissues, epithelial cells came to be progressively recruited for sensorimotor control (without the need for creating specific and highly organized control mechanisms *de novo*).

A number of characteristics of epithelial conduction play a crucial role in the amplification of agential capacities:

1. Epithelial conduction can both provoke the contraction of muscle cells and be affected by different stimuli.
2. Electrical impulses can propagate on an all-or-none basis: allowing for a non-linear selective activation and the amplification and integration of signals.
3. Electric action potentials (spikes) can propagate from cell to cell via gap junctions throughout the epithelial tissue so that effector responses can be produced at points that are distant from where the impulse originated.

These properties already permit some advantages over chemical diffusion mechanisms (the other major alternative for mediating intercellular interactions). Epithelial conduction permits both a high speed and distance for

propagating signals and the integration and amplification of signals, without causing interference with the more fundamental molecular metabolic mechanisms (an interference that risks provoking catastrophic outcomes). More importantly, epithelial conduction comes to satisfy a fundamental requirement regarding motility since, in multicellular organisms, *coordination and synchronization of effector cells becomes critical if full body response is to be achieved*. Although some agential strategies may rely on local stimulus-response systems (as it happens in sponges) motility and other adaptive strategies, require a global coordination of effectors. And, at the scale of multicellularity, an integrated and synchronized coordination is very difficult to achieve by molecular signalling, single flagellar rotation or other unicellular strategies. Note that, at this point, the fundamental problem is not really that of sensory “processing” or plastic sensory-effector linkage, but that of *coordinated* effector activity (e.g. mouvement, chemical, etc.). It is precisely due to the property of epithelial conduction to spread rapidly, radially and symmetrically all along the body surface that some early organisms could develop strategies that involved full body contraction (such as contraction of the swimming bell of Siphonophores) or effector responses that spread along the full body surface (such as luminescent or secretory responses—Mackie 20044). As reviewed by Horridge (1969) epithelial conduction may also coordinate some specific organ contraction, such as the tentacles of the polyp of *Tubularia* or the “mouth” of the fly trap *Dionaea*².

Now... which are the factors that facilitate and require the appearance of specifically nervous cells? An insightful limit case of multicellular agency mediated by epithelial conduction is provided by an *Hydra* without nerve cells as simulated by Albert (1999) and as shown by two real cases produced experimentally. One of such studies is due to Campbell et al. (1976) that removed the stem cell lineage that gives rise to neurons in *Hydra*. The second one, by Sugiyama and Fujisawa (1978), used mutants that did not develop nerve nets. Both approaches (empirical and simulated) show that *Hydra* can perform significant adaptive behaviour by means of epithelial conduction mechanisms alone, i.e. when the action of nerve cells is suppressed³. Albert’s experiment with a simulation model⁴ of epithelial conduction and primitive nerve

2 The importance of action potential activity over cell membranes (before the appearance or specialization of nerve cells) is not only crucial to explain the origin of the NS but also for its development. In this sense it is remarkable the important role that spontaneous and rhythmic muscle contraction (which appears before the NS has innervated the muscles) bears for the formation of the nervous system (Llinás 2001). It appears that ontogeny recapitulates phylogeny: rhythmic patterns of epithelial conduction along muscle cells come to solve a coordination problem coherent muscle movement (which is extremely important for other developmental processes to occur) and only latter in-forms neural tissues to endow them with more “abstract” sensorimotor capacities.

3 Some other full body interactions with the environment have also been reported to rely exclusively on epithelial conduction in Cnidarians (Mackie 2004).

4 It needs to be said that Albert’s paper does not show sufficient technical details of the simulation model

nets of *Hydra* suggests that the main crucial factors enhancing the adaptive responses are the density of conducting cells and the distance that electric impulses can traverse along a cell. This distance is obviously limited by the size of the cell in epithelial conduction and favours the formation of specific *nerve* cells that can grow branches that propagate the signals to further distances. His simulation model beautifully illustrates this transition.

Albert simulated a simplified *Hydra* consisting in a tubular body (an empty three dimensional cylinder) with three cells types (muscles, epithelial cells and nerve cells) encoded on a “genome”. The encoding was “populational” on that no specific connections or individual neural or cellular properties were specified but only cell types, defined by their function (muscular, epithelial or neural), and a number of properties of such cells types (such as number of cells of each type, number of connections per cell, time constant and space constant of the electric potential propagation, activation threshold for muscles and neurons, etc.). Changing several parameters of the genome (and their subsequent changes on the phenotype) Albert was able to test the limits and contribution of epithelial conduction on an adaptive agential task. This task consisted on the artificial *Hydra* having to move its body so as to catch a falling piece of food. The falling of the food served also as a chemical stimulus to the epithelial cells.

In some of the first experiments Albert tested artificial *Hydra* without neural cells on the aforementioned task showing that “[N]erve-free conducting system can have a similarly successful performance only if the space and time-constant have an extremely high value.” (Albert 1999:85). But, in nature, this parameters (space and time-constant of conduction decay) cannot take an “extremely high value” for epithelial conduction. This suggests that epithelial conduction alone is not able to propagate electric action potential in sufficiently accurate and efficient manner and that nerve cells need to appear to overcome this limitation. In a second set of experiments Albert tested the effect of a number of nerve cell parameters. The result what an improvement of behavioural performance correlated with an increase on the number of nerve cells and, secondarily, on their space constants (i.e. how far could they grow their axons/dendrites—there is nodistinction of both in early neural cells).

Although the results may, at a first sight, seem trivial, the simulation model permits to throw some light into a set of question that could not be so apparent if we only had at hand the limited empirical evidence that experiments with *Hydra* (or other early neural and pre-neural organism) can provide. On the one hand, the first set of experiments show that “the increase of the effectiveness of conducting stimuli alone is able to increase the adaptivity of the organism, which suggests a possible direction for evolution” (Albert 1999:84).

and the presentation of some results seems far from consistent. However, the model serves as a method for posing the right questions on a workable minimal-framework.

On the other hand, the second set of experiments dealing with number and space constant of nerve cells shows that, at least for the earliest forms of nerve-based response to environmental stimuli, a quantitative increase on the very basic parameters (such as number of cells) might be sufficient to produce significant adaptive improvement. “At this level of the evolution” Albert concludes “there is no need for developing complex centers in the nervous system to increase its effectiveness, so there is no need to store organizational information in the genome” (p.85). Thus, it seems plausible to hypothesize that there exists a relatively smooth evolutionary pathway that leads from simple multicellularity (where cells already possess bioelectronic capacities and epithelial tissues) to the appearance of early NSs. This might have occurred via epithelial conduction mechanisms and a progressive specialization of certain cells on amplifying signals and increasing their conductance.

Illuminating as it may be, however, Albert’s model falls short to provide us with an overall understanding of the factors that contributed to the appearance of the NS and the amplification of agential capacities that it accompanies. For a long time Parker’s *The elementary nervous system* (Parker 1919) stand as one the main (nearly unique) major references for the question of the origins of the nervous system. However, Parker, at the time, could not know about epithelial conduction nor about the light that molecular biology was about to throw over the fine grained process involved on cell membrane electric transmission. Today, his hypothesis of the origin of nervous tissues as mediating a two cell reflex arc (connecting single sensor and an independent effector cell) is widely rejected and new hypothesis have been proposed instead (Horridge 1969, Mackie 1990).

At this point it becomes important to distinguish between two aspects, which appear often mixed in the literature, regarding the origin of the NS. One of them concerns the material origin of nerve cells, i.e. what kind of cell might have been the precursor of current neurones. The other aspect refers to the functional properties and requirements that became the selective pressure for the increasing specialization of such cells. As for the first question there seems to be no current agreement, but most theories (as reviewed in Horridge 1969 and Mackie 1990) converge on two main alternative hypothesis: a) neurons arose from increasingly specialized and internalized epithelial cells performing some kind of pacemaker function coordinating effector cells and/or mediating between sensory and effector cells, or b) neurons first originated as neurosecretory cells performing internal chemical coordination and developmental growth regulatory roles and only latter came to mediate behavioural interactions.

Even if the second hypothesis came out to be the most accurate one (which might never be fully empirically demonstrated) we can assume that bioelectronic conduction mechanisms (like epithelial conduction) were already in

place mediating sensory and effector cells and coordinating synchronized effector responses. What requires the conceptual model that we are constructing, is to make explicit the set of limitations that non-neural electrical conductivity possesses. And also to address how certain cells came to overcome these limitations (independently of their origins) leading to a set of properties that specify neurons and made them crucial for the morphophylogeny of agency. I shall first summarize some of the limitations of epithelial conduction for mediating sensorimotor interactions in multicellular animals: a) the lack of directional and selective propagation (impulses spread almost homogeneously in 2 dimensions), b) the quick decay of action potentials in amplitude and conduction velocity, c) connections between epithelial conducting cells are not known to connect between them other than through gap junctions (where one of the layers of membrane lipids is shared among cells), as a result no means for modulation or regeneration of the signal exists. We know that these limitations were overcome by a new type of cell, the neuron⁵. Its elongated axon can project to specific targets *selectively* propagating action potentials to specific effectors or to other neurons via synaptic connections. In turn, different chemical components (neurotransmitters) are capable of regenerating, inhibiting or amplifying the signal. In addition, neurons can also receive connections from any other neurons so that signals can converge *and* diverge (unlike epithelial homogeneous propagation). Montalcini (2001) concludes that it is precisely the selective pressure for integration (convergence) and differentiation (divergence) that became the key factor defining the appearance of nerve cells.

5 Interestingly it is the very definition of a neuron that is really at stake here and both anatomical and functional criteria, together with some not less important aspects of the history of neuroscience (and to its technologies of visualization), seem to be highly intertwined on specifying what counts as nerve cells. The very definition of neural cells goes necessarily back to the neuron doctrine, first established by Ramón y Cajal at the end of the 19th century, establishing that a specific kind of elongated and excitable cell, the neuron, was the principal component and functional unit of the activity of the NS. The very idea of the nervous system (like any other aspect of the sciences of mind) somehow departs, one way or the other, from the top-down, i.e. from the study of human cognitive capacities to the extension and application of such knowledge further down the phylogenetic scale. However, the uniqueness of neural cell types in higher mammals and some of the distinctive features that permit their recognition (from chemical to anatomical) might be of little help concerning their most elementary and early stages. And as Horridge (1969) reminds us, it is the acceptance rather than the recognition of a cell as a neuron what is critical. And this acceptance “ultimately depends on the demonstration of their *function* as elongated conducting elements which control other cells by signals involving changes in permeability of the membrane to particular ions (...) Synapses, with definite cleft against which is a row of vesicles, prove to be the best indicators of nerves. They are critical indicators not only because they resemble the synapses of higher animals but also because they indicate an interaction of some kind between cells. The electron microscopic demonstration of a typical synapse tends, specially when the postsynaptic cell is known to be an effector, to indicate that the presynaptic element *functions* exactly as a neuron is supposed to do by definition. Doubt concerning the function of structures seen in the electron microscope is then a spur to further tests. Discovery of synapses, or of transmitters, as in sponges, does not settle the issue, which is a *question of the nature of the interaction*” (pp. 14–15, italics added).

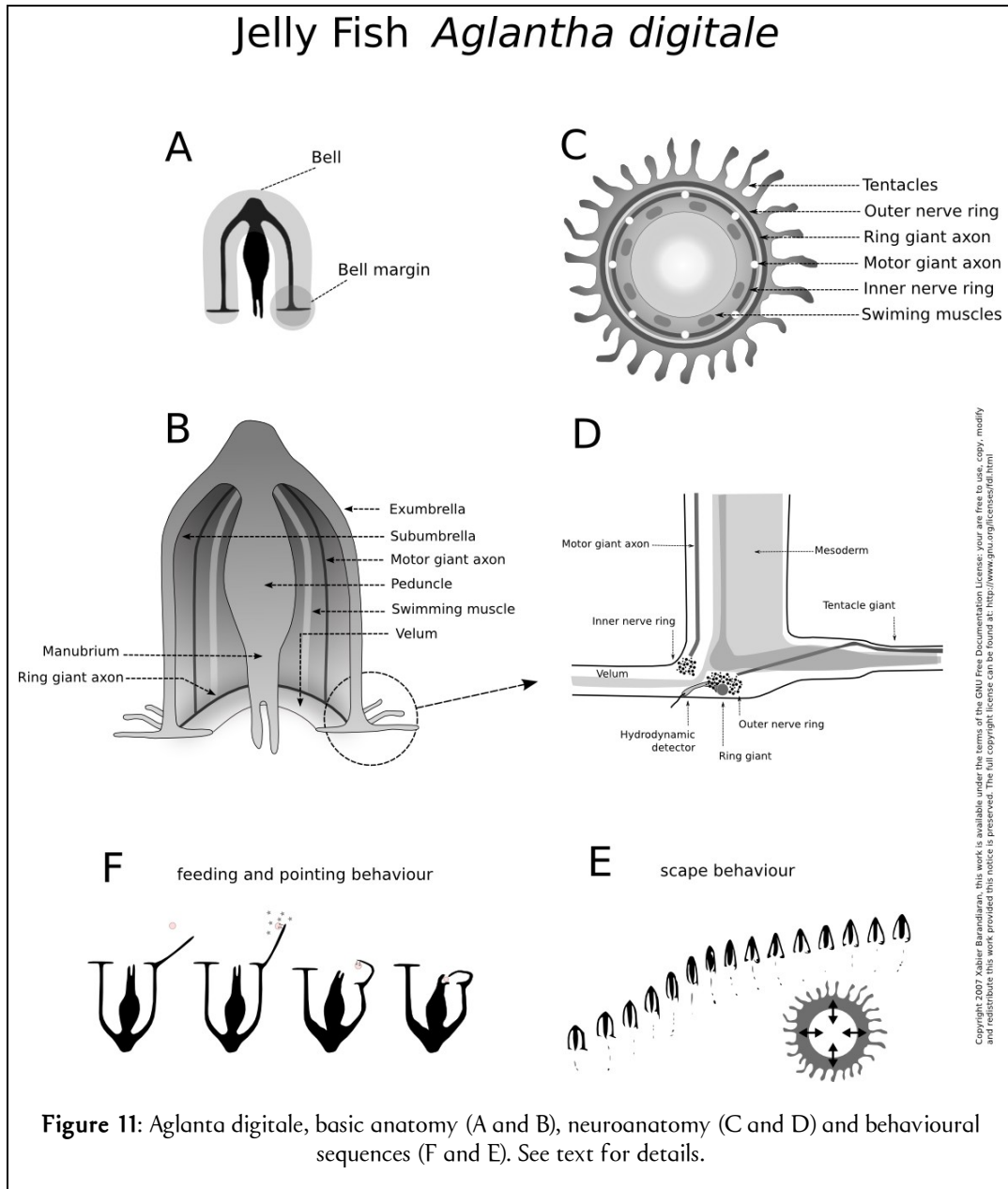
We are now in place to develop what are the consequences of these new mechanisms for sensorimotor control (along with the bodyplan of Cnidaria) for motile agency, through the case study of one of the earliest (yet not that simple) examples of neural based agency: that of the Jelly Fish *Aglantha digitale*.

2. CASE STUDY: A SCIENCE FUSION⁶ JOURNEY WITH JELLY FISH *AGLANTHA DIGITALE*

We merge down the sea near Vancouver Island. Prof. George O. Mackie accompanies us during the journey, he is a specialist on Cnidarian nervous physiology and will be able to provide us with some of the mechanistic details of what we are to observe today (Mackie 2004, Mackie 1990). Accompanying us we also have Prof. Satterlie (2002)⁷. We dive into the sea endowed with two special machines required for our scientific-fusion journey: a zoom-in spatial visualizer and a slow motion time machine. One hundred and fifty meters down the ocean we find our special guest today: *Aglantix*. It is a 1.5cm transparent multicellular organism, belonging to the species *Aglantha digitale*, class Hydromedusae, phylum Cnidaria. Its body, as most Cnidarians, has a bell-like cavity shape, with a falling elongated body inside, the manubrium, and several tentacles extending along the margin (see figure 11 A-B). Eight longitudinal muscles fall from top to the bottom margin of the bell together with eight giant motor axons. The most important physiological details required to understand its behaviour, however, Mackie (2004) reminds us, are found at the bell margin; two strands of nerves (the inner and the outer) extend circularly making two coupled rings along the margin, the projections of eight motor giant axons into the inner nerve ring are also found there, together with the giant ring axon and the tentacles (as depicted in figure 11 C). The inner and outer nerve rings are connected between each other acting as a single nerve ring. Yet, far from a uniform “primitive” physiology (as often simplified) the nerve rings contain several functionally distinct pathways, interacting with each other in complex ways. Along the inner ring there are also several concentrations of neurons, the pacemakers.

6 In the absence of an adequate term for this narrative-descriptive-explanatory literary gender, I have decided to call it science fusion. The new term allegedly resembles that of the much more popular *Science Fiction*. However, there are differences and similarities that make the analogy possible but not identical. It is certainly a form of fiction insofar the situations in which some of the characters found by the scientist are not real i.e. they never happened and could even be impossible. Yet the described scientific facts, the explanations provided, and the mechanisms and situations of the main character (the organism under study) are as accurate as possible, merging references from different scientific sources and sometimes even disciplines (hence the term *fusion*).

7 I need to thank Prescott (2007) for introducing me to the work of these two senior neurophysiologists.



Right now Aglantix is peacefully swimming. Small contractions of his bell produce the upward motion thorough a jet like effect: the bell compresses, water is pulled out from the bell through the bottom cavity and propels Aglantix up. The observed contractions of the bell, Prof. Mackie explains, are

the result of the muscular rhythmic contractions. The cells of the muscles are electrically coupled, but the electric signal only propagates locally, so that full body coordination requires innervation, controlled by the pacemaker system that extends along the nerve ring. The pacemakers generate endogenous rhythms which are conducted circularly and propagate up the subumbrella through the giant axons to the muscles producing the regular slow contractions that we observe. Although the sensory system has not yet been completely worked out, it is considered that slow swimming is modulated by sensors in the tentacles and the outer nerve ring. This happens for instance when Nitric Oxide (NO) is present in the environment, detected through sensory cells located at the margin, but could also be the result of photosensitive cells although “these aspects have been little investigated” confesses Mackie (2004b: 9).

I want to move to the other side of *Aglantix* so that I can see better where Prof. Mackie and Satterlie are pointing throughout the explanation. As I swim over it, *Aglantix*, swimming upwards, collides with my chest and gently stops till I have passed through. *Aglantix* has stopped, Prof. Mackie explains, due to contact with the excitable exumbrellar epithelium which is sensitive to touch. Upon collision, epithelial conduction spikes travel to the nerve ring and inhibit the pacemaker system. While our Cnidarian neurophysiologist keeps on into the details of the inhibitory process, *Aglantix* has turned upside-down and extends its tentacles. That is its “fishing” position, Prof. Satterlie notes. After a while a brine shrimp larvae comes across *Aglantix* and I really had to zoom in and slow the motion (thanks to Kass-Simon and Scappaticci 2002) to see what was going on at that point. Upon contact with the tentacle local nerves activate a very special type of cells, the nematocytes that contain amazing MadMax style powerful weapons (called nematocysts) which are immediately fired with great power. In less than 10 μ s, with an acceleration of 40000g, the nematocyst has impacted the prey. As an effect of the micromechanical attack, the tissue of the prey larvae is damaged, its molecules and cells spread over and are detected by the epithelial sensory cells on *Aglantha*’s tentacle. As a result, it discharges more nematocyst, which in turn destroy more tissue of the cell and the process enters into a positive feed-back loop effect until most nematocysts are discharge against the prey. The larvae is now immobilized, and probably dead, due to the poisonous neurotoxines that some of the nematocyst carry. The prey is also attached to the tentacle by some other harpoon-like nematocysts that have penetrated it. The tentacle starts to bend towards the margin of the bell moved by a muscle that is activated and controlled by a tentacle specific nerve net. This nerve net is connected to the inner nerve ring projecting to the muscles of the manubrium and triggering what is called the pointing behaviour: the manubrium bends towards the flexioned tentacle while the lips of the mouth opens, activated by the epithelial

pathways of the manubrium, preparing to attach the prey and engulf it (figure 11-F).

Due to the bending of the manubrium and the flexion of the tentacle *Aglantix* has lost balance and comes back to its original position. However, it does not swim any more. The pacemakers have been inhibited by the endodermal epithelial pathway activated during digestion. After a while, Prof. Mackie turns around pointing to a small approaching fish, yet bigger than our friend *Aglantix*. “The big fish eats the small fish... or medusae” says Prof. Satterlie. “If it can!” claims Prof. Mackie, while smiling stealthily, perhaps excited about the mechanism that is just about to be activated at *Aglantix*'s nervous system, a mechanism that has attracted his attention and research effort for a while (not without recognition). The predator fish, now quickly approaching *Aglantix*, generates specific vibrations on the water; these vibrations must certainly be different from ours since we have been close to *Aglantix* for a while now, without provoking what we are just about to see: *Aglantix* has displaced upwards about five times its body length milliseconds before the predator bites into the empty ocean! (see figure 11 E). “[The capacity to escape so quickly] is ultimately due to the ability of the motor neurons to conduct two different sorts of action potential with differing effects on the muscles” explains Satterlie (1990:916) “a phenomenon still without any known parallel in other organisms” (Mackie 2004b:8) concludes, while he remembers the *Nature* paper that followed his discovery (Mackie and Meech 1985).

The predator does not give up so easily so we get the chance to zoom in, slow once again the motion and get ready to observe the process with some more detail. One of the main problems to be solved here, Mackie explains, is that “swimming by jet propulsion requires synchronous, symmetrical contraction of all parts of the muscle field. Otherwise, the bell could not contract as a unit, and swimming efficiently would be reduced” (Mackie 1990:916). This may not be a problem if the swimming is slow, since differences on propagating pacemaker signals may be minute specially if synapses to the muscles and contraction of muscles are comparatively slow. But in a fast escape behaviour differences on propagating signals along a nerve ring or through the bell may be deadly and the pacemakers of the nerve ring could not do the job. This is where a single giant ring axon comes into play. Small hair-like sensory cells located around the vellum and the margin (these are the hydrodynamic receptors activated when the predator approaches) connect directly to the ring giant. In about 1.6ms the ring giant axon has propagated a spike in both directions all along the ring activating the eight motor giants. But the “normal” spikes of the motor giant axons are slow. The rhythms they propagate during slow swimming will not do the job required to escape the predator. The new arriving signal from the right giant axon, however, triggers a new spike through sodium channels which is much more powerful than the previous

ones (conducted through calcium ion channels). The peak of calcium spikes does never reaches the activation threshold of the sodium spike, so that escape behaviour is never activated when normal swimming occurs. As the predator approaches again the sodium spike activates the neuromuscular systems of the bell with a very short junction (of about 0.7 ms), the muscles contract and Aglantix is safe again.

“Response latency” Mackie comments, “measured from stimulus to first detectable movement, is about 10 ms, which compares favourably with fast start response latencies of many fishes [potential predators]” (Mackie 2004:12). “Not that good” says Satterlie, while pointing to a missing tentacle on Aglantix. After the third attack the mouth of the fish is open (and probably in hard pain due to the action of the defensive nematocysts) while a small tentacle falls into the deep ocean. This time the escape behaviour must have originated on the biting of a tentacle (not through hydrodynamic receptors). Upon sharp touch the tentacle signals the ring giant and detaches itself from the autotomy joint located at the base of the tentacle (near the margin).

The predator seems to have abandoned his attempts and I decide that we should equally make our journey to an end. “It looks like it has learned the lesson: do not try to eat an *Aglantha digitale*” I conclude, “even if you are driven by the hate of Aglantix having eaten your daughter larvae”. “Not so fast, young man” says Prof. Mackie “this is not *Nemo* or any other of the silly Disney animal cartoon films. We do not know the physiology of that predator fish, but certainly ‘hate’ is not the physiological mechanism that produced the behaviour we have just seen, you will still have to carve deep before being able to use that vocabulary”.

3. CNIDARIAN AGENCY

We have moved from unicellular motile agency (controlled by different chemical pathways within the cell) to a multicellular mode made possible by a number of specialized nerve cells (excited or modulated by sensory cells) forming networks that generate and transmit spiky action potentials innervating contractile cell complexes (the muscles) distributed along the body to produce coordinated movement. This mode of agency brings about crucial changes on the characteristic features that constitute agency (namely: individuality and system-environment relationship, normative functionality and causal-asymmetry) .

First of all, the appearance of the NS in multicellulars permits a new mode of decoupling of adaptive and regulatory processes. As a development of the requirements that epithelial conduction came to satisfy, neurons could produce an integrated and differentiated control system, decoupled from other

metabolic or constructive roles, dynamically rich and flexible, efficient and fast, and capable of extending all along the body to coordinate different processes. In Bonner's words:

This example [nerve nets in Cnidarians] is particularly significant because it shows a kind of cooperation between differentiated cells with which the adult is endowed for life. In more primitive forms all cooperation between cells types occurred during a transitory period of development. The very basis of all the extraordinary power of motility and coordination which are so characteristic of the entire animal kingdom stems from this kind of permanent interaction between cell types. (Bonner 1988:164).

It is thus at the level of individuality where the possibilities of the NS for multicellularity seem more apparent. Specially at the agential level⁸, but certainly not exclusively, since neurons in Cnidarians are also considered to play a significant role “in the regulation of growth and in the production of chemical morphogenetic gradients [in development]” (Mackie 1990:913). As I have previously mentioned it is at the level of movement coordination where the significance of electric conduction is made most apparent. Different cells acting locally (even muscular ensembles of contractile cells) have little chances to achieve a synchronized fast movement without the mediation of a subsystem like the nervous one. Only when this form of global coordination is in place can the multicellular organism move as a whole, sensitive different environmental or internal conditions, integrated, once again, through the NS. Thus, unlike the individuality of uni and proto-cellular systems, individuality at the multicellular scale, and particularly in the agential domain, is achieved though a system of coordination of component individualities. Yet this NS is in turn composed of cells and their function (within the whole) critically depends on their capacity for individual adaptive self-maintenance: i.e. on their ability for recovering a homeostatic balance of ionic concentrations, letting in the process an action potential pass through their elongated surfaces. This action potential in turn can coordinate its effector cells to achieve proper integrated actions. As Keijzer rightly points out:

[W]hen compared to the cellular level, *Aequorea victoria* is a huge organization which involves different and complex new forms of coordination compared to those on the cellular level. The animality present in *Aequorea* is thus not its simple agentive functionality—deciding to swim either fast or slow—but, rather, the kind of

8 Fred Keijzer (2006) has defended that *Aequorea victoria* could be taken as a case study for the notion of animality. He borrows the term animality from the biophilosophy of Hans Jonas and argues that the biological grounding of animal behaviour (in particular the coordination problem that the NS comes to solve on its early evolution) provides a better departure point to address cognitive issues that does the notion of agency. Yet, his approach seems to take a rather abstract and unproblematic conception of agency (unlike the bottom-up approach that I have favoured). It is precisely in contrast with such conception of agency that Keijzer finds in “animality” a biologically grounded departure point for cognition. From our morphophylogeny of agency, however, the contrast between agency and animality does not really appear. Animality is rather a specific kind of agency that comes to solve the problems that multicellularity brings to biological organization and motility.

problems that must be overcome to produce such large-scale functionality, given the initially microscopic cellular building blocks. (Keijzer 2006:1596).

Moreover, with neural coordination different neural pathways can elicit sequences of movements giving rise to proper actions, with their preparatory phase, their initiation, duration and extinction. Recall, for instance, the “fishing-behaviour” of our friend *Aglantix* with a preparatory phase (in which the organism stands the bell margin at the top and the tentacles extended), then an initiation phase (upon contact with a prey and local response of nematocytes) a duration or execution (tentacle bending, flexion of the manubrium, lip opening, etc.) and finally an extinction phase (turning back to the normal position of swimming which is now inhibited due to the digestive muscle contraction). This is all made possible by the coordination of different local (tentacular) and global neural pathways, each activating and inhibiting (deactivating) different sensorimotor couplings between the system and the environment. The body (its shape and muscular distribution), together with some crucial environmental features (like the hydrodynamic properties of the water), is not less important than the NS. For instance the “righting” manoeuvre (the turning to the “normal” position after feeding is called) is brought about by the asymmetric deformation of the velum and the differential flow of water into the bell. More evidently the slow and fast swimming behaviour critically depends on the water propulsion that the bell shaped body permits.

Yet... how does the type of agency that the NS permits fit into the broader living and metabolic organization in which it is embedded? Now capable of executing actions, the organism can control its interactive processes in a much more sophisticated and effective way than ever before. The decoupling of adaptive agency from the metabolic infrastructure expands the relationship between system and environment in a number of ways. First, as already mentioned in the case of chemotaxis in *E. coli*, motility permits to control the boundary conditions of the system in an indirect, yet much more powerful way. For instance, if the conditions for self-maintenance (temperature, nutrients, etc.) are not favourable the system can make use of internal energy to produce detector-guided displacements to migrate to more grateful surface conditions. What is distinctive of this transition is probably the way in which the selective activation of different sensorimotor pathways permits to switch between alternative modes of engagement with the surrounding. More importantly from the point of view of agency, the organization of the NS in different layers permits to *execute nested sequences* of interactive couplings (such as the fishing behaviour in *Aglantha*) to strategically achieve a particular boundary condition (i.e. to get nutrient into the walls of the digestive system).

It is interesting to study *Aglantha digitale* because it is taken to be a limit case of the degree of complexity of the nervous system that a Cnidarian body-

plan can encompass. As Mackie points out: “[*Aglantha digitale*] has been extremely instructive to work on, not because it typifies hydromedusae as such, but because it shows what the Cnidarian body plan is capable in terms of nervous organization”. (Mackie 2004b: 18). This is due in the first place to the nervous organization of Hydromedusae (to which *Aglantha* belongs) which, unlike scyphomedusae (whose nervous system is distributed all along the bell), appears centralized or integrated in two nerve rings around the bell margin. This is considered a first step towards centralization and has made some researchers affirm that “The term ‘central nervous system’ can legitimately be applied to hydromedusan nervous systems as these animals have concentrations of hundreds of axons running in parallel forming ‘nerve rings’ in the margin” (Mackie 2004b:5).

However, behavioural diversity in *Aglantha digitale* appears limited grossly by the number of pathways (either organized in rings or in different types of networks), in the absence of more complexly organized NS (i.e. in the absence of a more selective connectivity between neurons). Interestingly, the circuit diagrams that show the neurophysiology of *Aglantha* are drawn as a series of horizontal lines representing different pathways of the nerve rings, together with vertical lines representing inhibitory or excitatory synaptic connections between them and different connections coming from different sensory inputs. Behaviour can thus be pictured out as the interaction between the different pathways affected by incoming sensory signals. The behavioural repertoire of Cnidarians, as Prescott (2007) and Keijzer (2006) rightly point out, very much resembles a layered behavioural architecture (so characteristic of behaviour based robotics—Brooks 1991) with each layer in charge of a specific sensorimotor coupling or stereotyped behaviour. No complex internal dynamics involving differentiated states within a pathway are found, only weak (inhibitory or excitatory) interactions between layers or pathways.

It thus seems like we have reached another bottleneck at this point. This bottleneck does not seem to be related to any intrinsic of Cnidarian neuronal cells. As reviewed by Mackie (1990) Cnidarians (even the *Hydra*) present different neural cell types, some of them enveloped on glial-like epithelial sheaths, they present a wide range of different junctions, including sophisticated synaptic connections (although this is admittedly the less developed area of research), the action potentials are similar to those found in higher mammals and *Aglantha* has even developed a two spike giant axon with no known parallel in nature. In Mackie's words “In no way can nerve cells in these animals be regarded as primitive, if ‘primitive’ implies that they have gone only a small distance toward evolving a repertoire of functions comparable to those found in higher animals”. (p.913).

Thus, rather than to the limitation of its component cells, the bottleneck might well be related to the symmetrical radial bodyplan imposed on Cnidari-

ans. A radial order is difficult to break to produce anything other than parallel neural pathways and the behavioural layers they embody. It would, however, be too precipitous to reach this conclusion so quickly. As reviewed by Satterlie (2002) Cubomedusae are capable of very flexible behaviour including change of directionality of their swimming (unlike other medusae) to follow and catch a prey. The radial symmetry of its behavioural organization is broken by semi-coupled rhopalia pacemakers each of which contains two compound eyes and four pigment-ocelli. Whereas the rhopalia in Hydromedusae are completely coordinated⁹ in Cubomedusae rhopalia appear semi-coupled; each being able to control local muscles almost independently of other rhopalia. Yet this rudimentary form of symmetry breaking driven by the differential sensory excitation of the loosely coupled rhopalia might well be all that a Cnidarian bodyplan can achieve in terms of asymmetric behavioural complexity.

Unfortunately, extremely little is known about the interaction between metabolism and NS in Cnidarians (Mackie 2007, personal communication). This is certainly an important area of research to fully understand the biological embodiment of Cnidarian agency, the way in which the constructive cycle of early multicellular animals is coupled to the interactive cycles that the NS brings forth. Thus, we have no other choice than to ignore, at least temporarily, this crucial aspect of the morphophylogeny of agency for this particular transition. (The metabolic implications of the Cnidarian bodyplan, and particularly the study of the internalization and specialization of a digestive system, might well permit to infer some useful information. But I will not pursue these questions here.)

If, together with some Cnidarian neurophysiologists, we take the ring-shaped centralization not to be a limitation *per se* but an adaptation to radial coordination (Mackie 1990), then, it is the radial bodyplan what imposes an important limitation. A radial organization can certainly generate little diversity and directionality of behaviour other than what may be produced by local differential stimuli (i.e. the contraction of a tentacle acting in isolation or or Cubomedusae's behaviour made possible by the differential activation of loosely coupled regulation). It is thus at the level of the functional organization of the NS and its embodiment that we are to found a bottleneck, not at the level its components¹⁰. Now the disposition of this radial organization is

9 In Hydromedusae all rhopalia are connected to each other and the fastest firing pacemakers resets the others imposing a uniform rhythm for coordinated swimming.

10 It rests, however, to be seen to which extent this radial symmetry may not be broken by experience itself. A higher plasticity of Cnidarian neurodevelopmental plasticity (possibly instantiated by more sophisticated synaptic mechanisms) could probably give rise to a symmetry breaking of its neural organization. However, how could then this architecture coordinate a radially symmetric body? Certainly comparative studies of artificial evolution of radially embodied neural architectures with different forms of plasticity could throw some light onto this issue, thus providing an interesting research avenue for simulation models of embodied agents.

severely dependent on the developmental processes that bring it about. And these developmental processes, incapable of building but on the basis of a radial symmetry, might well be touching a ceiling of agential complexity that can not be overcome other than by the introduction of the new developmental mechanisms. These mechanisms gave rise to Bilaterian bodyplans and the huge explosion on behavioural diversity and sophistication that it made possible.

4. THE EVOLUTIONARY DEVELOPMENT OF THE EARLY NS AND BILATERIAN AGENCY

Becoming agential is a two way process for an organism: so do new agential capacities make some forms of biological organization possible so do these forms of organization (type of metabolism, required temperature, etc.) make agency necessary. An equivalent transformation happens in ecological terms. The very same process that makes you able to spend long times searching for food, without immediate nutrient income, also makes you a high quality source of nutrients for your ecological partners. Thus, acquiring increasingly powerful agential capacities is inextricably linked with increasingly complex requirements (avoiding predators among them). And as Sterelny (2001) points out the hostile world that agency produces in terms of competition and predator-prey dynamics also becomes a strong selective drive for increasingly complex agential capacities (“representation of the world” he calls it). One could also argue that, in general terms, the faster and more free to move you become, the more agential and sophisticated, the more embedded on an increasingly complex and fast changing world¹¹. But, although the logic of evolution seems to carve this inclined path into the valley of increasingly complex adaptive agency, a bodyplan is required that can make the journey down. Gravity alone did not make the mole delve. Nor can an abstractly and externally defined evolutionary selective pressure make an organisms become increasingly complex in agency.

Evolution does not happen in the abstract space of all possible phenotypes, including Universal Turing Machines and representational architectures. Evolution does not even operate designing organisms whose biological autonomous roots can be metaphorically downplayed as mere “raw materials” (Godfrey-Smith 2001) to be engineered by natural selection. The metaphor of biological organization as just a material constraint, like iron imposing strength but lack of flexibility for the sculpture maker (Dennett 1995), is a misfortuned conceptual trap. On the contrary, evolution can only happen on the dirty and

¹¹ Obviously this heuristic is not meant to state that intelligence is directly and uniquely correlated with motile freedom and speed. But it may serve to understand the early evolution of the NS.

limited realm of biological developmental organization, with the opportunities and limitations that it brings for the self-construction and self-maintenance of stable forms and functional compositions (defining the very functional space to be “navigated” by evolutionary trajectories). It is thus on the framework of evolutionary developmental biology where we are to find the path to follow on this quest. We have seen how Cnidarians, which are considered to have been evolving for over 700my now, present a bottleneck on their capacity for increasing agency. This limitation is most probably due to constraints on the development of their bodyplan (including the developmental pathways leading to their neurophysiological organization and body shape). Yet Cnidarians could survive in a hostile agential motile environment even if limited on their agential capacities and plasticity. This could probably have been impossible without their nematocysts. In Kass-Simon and Scappaticci's words: “That the Cnidarians have persisted in marine and freshwater niches since the Cambrian is due not least to the effectiveness with which their nematocyst-bearing tentacles act as lethal weapons of defence and predation” (Kass-Simon and Scappaticci 2002:1773). However limited, nevertheless, the path is open to explore motility at the multicellular scale and new developmental mechanisms soon became available.

The next major evolutionary transition comes with the appearance of Bilaterians (bilateral symmetric organisms with a “new” bodyplan). The appearance of this new multicellular organization is attributed to the evolution of the HOX gene cluster capable of much more sophisticated developmental regulatory processes than the previous genetic equipment permitted. And it is precisely together with the appearance of this new genetic regulatory mechanism that a huge explosion in behavioural performance and NS evolution occurred. In a period of time of about 5 million years the basic bodyplan for the subsequent evolution of the NS appeared: “complex neuromuscular systems arose very rapidly at the base of the Cambrian and this rapidity of construction has important ramifications for where any given nervous system ‘sits’ on an anatomically defined landscape of constructional domains and the genomic costs and time constraints of moving it elsewhere during evolutionary time periods.” (Gabor Miklos et al. 1995:275). According to these authors the “potential universe of neuronal architectural space” has been very little explored since the Cambrian explosion. This may be due to two main factors: there has been not much time for exploration since the Cambrian and there is not much room for variation along the well stabilized bodyplan of Bilaterians, whose genes specifying NS architecture are highly pleiotropic (they have multiple phenotypic effects). As a result all the subsequent motile agents on earth share a common bodyplan whose genetic and phylogenetic structure has recently attracted the attention of researchers.

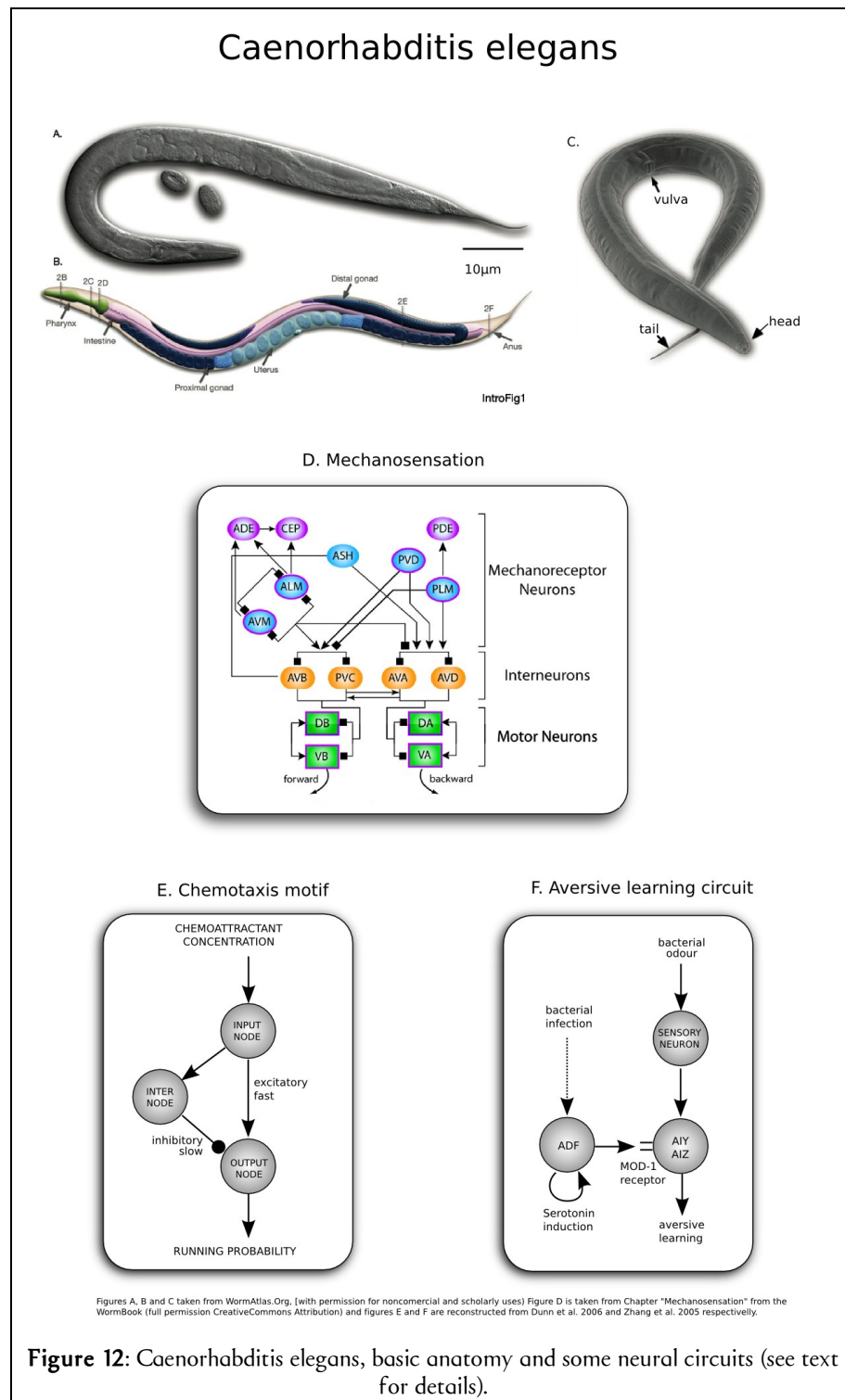
The common ancestor of Bilaterians that gave rise to complex embrained animals has been named *Urbilateria*. It is recognized that *Urbilateria* already shows a sophisticated nervous system (Ghysen 2003). It needs to be said that *Urbilateria* is nothing more than a theoretical model animal that integrates the basic bodyplans of Bilaterians into a common ancestor. The structure of the model is built using phylogenetic, morphologic (homologies), developmental and genetic comparative biology. According to Ghysen, the Urbilaterian nervous system, defining the basic bodyplan shared by all (except Cnidarian) organisms with neural systems, presents a set of essential features:

1. *Orthogonality*. The NS is basically a three-dimensional structure that develops almost-independently along three axes with their own specialized cells: antero-posterior, dorso-ventral, and apico-basal.
2. *Inheritable connectivity*. Connectivity of brain circuits is mainly established by axon growth guidance mechanisms (relative to orthogonal axes) during development facilitated by fibre scaffolding. Genetic instruction and switching of axon guidance make it seem plausible (although not generally demonstrated) that “elementary functional circuits were already genetically designed within this orthogonal net, and could have been inherited by all derived animals” (p. 557).
3. *Cell diversity and specialization*. Mechanisms to permit and channel neuronal cell diversity and specialization were most probably already present on *Urbilateria*, depending on “the differential expression of a large number of transcriptional regulators which are common to all triploblasts” (p.557). The requirement for neural specialization may come to satisfy a universal requirement if we follow Ghysen on that “[g]iven the major importance of the capability to generate defined types of neurons in a defined succession, it may be that no complex nervous system could evolve without the ability to program and control asymmetric divisions” (p.557).
4. *Longitudinal subdivision along the antero-posterior axis*: brain-boundary-cord. There are specific genetic and developmental constraints that participate in the differentiation and morphology of each of the parts.
 - a) *Brain*: Unlike Cnidarian nerve nets, Bilaterian's NS is centralized in a brain or protobrain also called cerebral ganglion in arthropods. Apart from this centralization of neural cells another common characteristic of the Urbilaterian brain is the existence of convergence zones for sensory axons. Two kinds of convergence zones appear: a) somatotopic organization “where the position of each axon terminal reflects the position of the sense organ”, and b) single convergence zones (glomerulus) for specific odorant receptors.

- b) *Hindbrain*, posterior to the boundary region: what is called hind-brain in chordates and sub-oesophageal ganglion in insects. “This region differs from the brain in that it is clearly segmented, and its segmental diversity is patterned by the Hox genes (which are excluded from the protobrain)” (p.558). This region can be considered a cephalized part of the cord and its main function is the control of vegetative functions: blood circulation, breathing, and some reflexes such as coughing and vomiting¹². The hindbrain also receives inputs from facial touch perception and olfaction. Interestingly, the hindbrain provides the basic infrastructure for the neural coordination and regulation of internal organismic functions.
- c) *Cord*: Innervating the body along the anterior-posterior axis and providing an architectural axis for the coordination of the articulated body

Ghysen argues that only after a process has become developmentally stabilized can it undergo large variations without compromising its function; what he calls the “rule of conservative changes”. If we apply this rule to the evo devo of agency it becomes apparent that the Urbilaterian bodyplan and NS development is not just an ancient vestigial caprice for the palaeontologist delight but a core architectural plan from which the infrastructure of cognition is one specific type of variation. One of the characteristics that has called most attention in this bodyplan (and that has consistently been related with increasingly complex agential capacities) is that of centralization (i.e. the formation of a brain). Following Prescott (2007), although centralization is generally attributed to the process of cephalization (the concentration of sensors in the anterior end of the antero-posterior axis) Koopowitz and Kennan (1982) have proposed an alternative view: it is the change of Bilaterian bodyplan that forced centralization under the requirement to coordinate different “peripherally-based reflexes” and integrate directional and coherent behaviour. Although Prescott takes this hypothesis to support the idea that the brain may have evolved “as a centralized substrate for action selection” what remains uncontroversial is that the Bilaterian bodyplan permitted (and equally required) the centralization of adaptive agential control. Thus, unlike the Cnidarian bodyplan which, despite its not-so-elementary nature, suffer a number of important bottlenecks, the Bilaterian bodyplan permitted a three dimensional centralization (recall the orthogonality principle of Urbilaterian NS) with a convergence zone for sensory projections, a selective growth of axons

¹² Hindbrain in insects (the sub-oesophageal ganglion) may also be involved on vegetative functions. Although previously it was considered not to have neural respiratory function (due to the existence of a mainly diffusive respiratory system), recent studies on insect respiratory system and function might seriously question this assumption while ascribing relevant regulatory functions to the hindbrain region.



that defined a differentiated circuitry (unlike the Cnidarian nerve nets), with a high cellular specialization and differentiation and an anterior-posterior innervation of the body permitting directional articulation of movement.

Caenorhabditis elegans is a flatworm (phylum *Platyhelminthes*) and one of the best known and more simple model animal and can be taken as a living model of the common ancestor of the Bilaterian family¹³.

5. CASE STUDY: A SCIENCE FUSION JOURNEY WITH *CAENORHABDITIS ELEGANS*

This time we shall start our travel through the electrically conducting wires attached to my Ethernet card. First stop is the *wormatlas.org* from where we shall recruit a number of neuro-anatomist for our trip. Second stop is *wormbook.org*, another free and open content repository of data and literature around our new friend, the object of our case study: *Elegantix*, an hermaphrodite exemplar of the nematode species *Caenorhabditis elegans*. It is a very elegant specimen indeed on its nearly aesthetic biological minimalism and it is one of the smallest and better scrutinized Bilaterian organism on the surface of this, our home, planet. It has been a model animal for genetics and development since the 60s and 70s (Brenner 1974) and more recently, since 1990, for learning and plasticity in behavioural neuroscience (Rankin et al. 1990). Although natively a soil-dwelling creature, its population in *Petri-Dishland* might well be by now bigger, given huge amount of experiments that are done with *C. elegans* every day. As we travel around the *wormatlas.org* and the *wormbook.org* we find a huge number of molecular and developmental biologists, neuroscientists, even philosophers, joyfully discussing about many of the details that shall become a crucial part of today's trip. This is the *C. elegans* research community that has decided to overcome the complexity of this organism by sharing their knowledge and organizing it on the basis of open access and participatory and collaborative research, reviewing and databasing. We gently recruit a multidisciplinary team of researchers for our trip.

As we are getting ready for our journey, I learn that most of the neurophysiological studies of *C. elegans* are carried out using two main techniques: genetic manipulation and laser ablation. The full genome of *C. elegans* has been sequenced, and detailed descriptions of the developmental pathways are known, often including which gene transcription is carried out at which mo-

¹³ It must be noted however, following Prescott, that although originally thought that *platyhelminthes* was a stem phylum from where all other Bilaterians might have evolved, new research results have classified most flatworms in the protostomes group and may not be accurate representatives of the first Bilaterians. However, for our enterprise, it is enough for us to take it as a minimal example of a Bilaterian agent, although it may be not evolutionarily representative as a common ancestor.

ment of development and throughout adult life. This knowledge permits to create specific mutants whose genomes have selective molecular effects on certain cells (including neurons) under certain conditions (see Mori 1999 for a review). Genetics becomes thus a generalized research methodology to infer the physiological function of certain molecules on the behaviour of an individual (e.g. a detector binding molecule of the membrane of a sensory cell or the role of serotonin for learning¹⁴). Combined with genetic studies, immunoreactivity techniques are also used: staining antibodies are created that match certain molecules (like serotonin, see Zhang et al. 2005) and the production of such molecules can be monitored through the microscope during behaviour. Laser ablation studies (often used in conjunction with genetic studies) are used to isolate the contribution of a certain neural cell to specific characteristic behaviours which permits to isolate the minimal circuit that is required to achieve it (see Rose and Rankin 2001 for a good review on ablation techniques for the study of habituation). Finally, computer simulation models are starting to be used to disentangle the complex dynamics that the neural architecture of *C. elegans* supports. Due to the small size of *C. elegans* neurons and the high hydrostatic pressure of the internal milieu, however, there was, until very recently, no possibility for electrophysiological recordings to be made. But new techniques are starting to be used (Faumont & Lockery 2005) and a much better understanding of the nematode's neural dynamics is expected to be reached soon. Obviously the above methods rely on very well quantified behavioural studies and the full neuroanatomy of *C. elegans*, a neuronatomical architecture that is extraordinarily similar (almost identical) between individuals of the same species and whose full mapping is available since pioneering work of Prof. White and colleagues (1986).

Equipped with our zooming and slow motion machine we travel to Algeria. We are pointed by Kiontke and Sudhaus (2006) to find our friend Elegantix in a local farmer's compost and we belittled to about 1mm, the average size of *C. elegans*. We are now ready to analyse what nematode agency looks like. We find Elegantix twisting its body while producing forward motion in the biochemically rich environment of the compost. I ask Prof. Schaffner to introduce us *C. elegans*. He gently explains the well known facts that *C. elegans* is an elongated worm (whose aspect is similar in shape to the huge worms of planet Arrakis in David Lynch's 1984's film *Dune*—but 6 orders of magnitude

14 For instance, on a now classical paper by Zhang et al. (2005) the role of serotonin on learning is detected using mutant *tph-1* and other variants. In particular *tph-1* mutant is known to be deficient (due to the set of mutated gene sequences) on the production of the molecule tryptophan hydroxylase necessary for the production of serotonin, but not required for the production of other molecules of the same family. Other mutants lacking other neurotransmitters (such as dopamine) were also tested under the same conditions without recognizable effect on the learning assay. Finally other conditions such as normal detection, innate immunity or behaviour were intact despite *tph-1* mutation. These kind of test permit to infer an essential role for serotonin on that particular learning task through systematic behavioural test and comparison with other genetic mutants.

smaller) composed of 959 cells (on its hermaphrodite form) of which 302 are neurons (282 are somatic and 20 are pharyngeal, and both systems can function independently), 56 cells are glial or associated support cells, and there are 95 muscle cells. Thus, almost half of the organisms component cells are specialized in agency. Neurons form approximately 5,000 synapses, 600 gap junctions, and 2,000 neuromuscular junctions. Thanks to Prof. White (White et al. 1986) we are able to zoom-in and see the full picture of its nervous cells and their connections. Compared with that of to a close friend, *Elegantix*, the anatomy of both is remarkably similar, almost identical. In fact, Kaufman points out (2006:1564) “within the limitations of the currently existing large-scale data we successfully identify statistically significant information characterizing the fundamental relation between the expression [of genes] and synaptic properties of neurons”. Yet, despite the fact that the neuroanatomy (both at the level of neurons and their connections) is extremely accurate “[i]t has turned out, however, to be very difficult to tie the behavior in any simple way to the neuroanatomy” (Schaffner 2000). Rankin is explicit on this, ablation studies on habituation training have shown that “the activity/plasticity in a single cell or cell type may not correspond directly with behavior and it is important to understand how all of the circuit elements contribute to the behavior” (Rose and Rankin 2001:66). This is why, Dunn notes, researchers are currently using computer simulation models to study dynamic network motifs (dynamically equivalent connectivity structures) to figure out how sensorimotor correlations are produced by its NS (Dunn et. al 2004, 2006). But it is not only the dynamics of the circuit that needs to be understood, Rankin points out, “[t]aken together the cell lineage data and the nervous system reconstruction gave the impression that the 302 neuron nervous system of *C. elegans* is ‘hard-wired’” (Rankin 2004:R617). But far from ‘hard-wired’, she explains, the NS of *C. elegans* shows a huge plasticity Prof. Hobert confirms that the expression profile of genetic developmental processes does in fact “change under defined environmental conditions and upon changes in neuronal activity” (Hobert 2006:9). Prof. Mori is not less assertive: “Despite the prevalent notion that the nervous system in *C. elegans* has rigidly connected neurons, sensory axon morphology is subject to modification even at the adult stage, according to the change in sensory activity” (Mori 1999: 412). And these changes of connectivity and molecular modulation of NS activity are crucial to understand *C. elegans*’ behaviour as we will soon see.

A closer look at the surface of *Elegantix* shows a wide variety of sensory cells concentrated on the anterior part of the “head”: mechanoreceptors (some of which have a proprioception function and are distributed covering almost the full surface of the body—Goodman 2006), thermoreceptors, chemoreceptors, and odour detectors, among others. Sensory cells detect different surrounding conditions (chemical concentrations, pressure, etc.) and

transmit an electric signal along their axon to interneuron cells, that in turn connect to the motor neurons (innervating the muscles distributed along the body) thus producing a wide range of sensorimotor couplings with the environment: chemotactic behaviour, aerotaxis, thermotaxis, egg laying and even a sophisticated mating behaviour (Barr & Garcia 2006) for which males are endowed with 52 additional mechanoreceptors and an additional number of muscles on their tail, where their spicule is located (Goodman 2006).

Intrigued and excited to start provoking some reaction I gently push Elegantix on its tail. The reaction is quick, the nematode accelerates the forward movement. I keep playing the game with some delight and Elegantix moves even faster. After a minute, however, it seems to be 'bored' and does not react any more. This is a well known response, Rankin explains, initially, when tapped on the tail, the nematode sensitizes (i.e. its response to the stimulus is stronger than usual), however, as you keep tapping, it gets habituated. The circuit involved on mechanosensation is quite complex (see figure 12-D) and we lack detailed knowledge of its functioning. However, activity dependent extrasynaptic release of dopamine (Sanyal et al. 2004) might well be related to this process of sensitization and habituation and "the most likely sites of plasticity for habituation to tap are the sensory neurons and/or their synapses onto the interneurons" (Rose & Rankin 2001:67).

Amazed by Elegantix's ability I go back onto tapping its tail again but, as I repeat the process, our friend gets habituated even faster than it did before. Rankin tells me how that is called "context-sensitive habituation", the worm is probably detecting my smell or some chemical compound that my skin might be releasing and has associated it with the previous habituation process accelerating the new habituation response. However, as we are still to see, this is not the only form of learning that Elegantix is capable of achieving.

Elegantix has now started to make bizarre bouts of turning and changes its direction of movement. Those bouts have been called "pirouettes", explains Prof. Lockery, and are followed by around two minutes of forward movement. "Pirouette probability" Dunn follows "is modulated by the rate of change of chemical attractant concentration ($dC(t)/dt$). When $dC(t)/dt < 0$, pirouette probability is increased whereas when $dC(t)/dt > 0$, pirouette probability is decreased. Thus, runs down the gradient are truncated and runs up the gradient are extended, resulting in net movement towards the gradient peak". (Dunn et. al. 2004: 138). I wonder around the similarity between *E. coli*'s and *C. elegans*' chemotaxis. Meanwhile Prof. Lockery has taken a multivariable chemical detector to measure the surrounding chemical gradient of potential chemoattractants here and there. As he comes back he shares with us the conclusion that our friend might well be engaged on a chemotactic behaviour moving up a chemoattractant gradient that is a combination of different compounds. That is the distinctive odour of the product of *Serratia*

marcescens bacteria's metabolism, explains Prof. Bargmann, while she looks at the chemoattractant reader. In fact, she continues "[what we are seen is most probably an] olfactory chemotaxis towards food-associated odours, an innate behaviour that is highly reproducible among animals" (Zhang et al. 2005:179). I zoom in to get a grasp on the neurophysiological details of how this "innate" behaviour might be instantiated on a sensorimotor circuit, but I can get little understanding of what is going on. Dunn points to me the relevant circuitry that has been shown to be at least necessary and sufficient for chemotaxis (using laser ablation experiments) but there are about 16 interneurons mediating between chemoreceptor neurons and motor command neurons with a non negligible number of feed-back loops involved, not to speak of the huge number of synaptic and other molecular mechanisms in between. Little understanding can be gained from this circuit diagram. Dunn smiles with abatement, he has been working with simulation models in Lockery's lab, optimizing different types of dynamic recurrent neural networks to match the sensorimotor correlation data previously available. The optimized networks show a small number of stereotyped solutions, he explains, that are called "motifs". "By discovering the simplest motifs capable of producing observed dynamic behavior, we gain functional knowledge of the system's probable internal dynamics" (Dunn et al. 2006:553). One of the most recurrent motifs that results from the optimization experiments, and that most probably considered the one used by *Elegantix* for chemotaxis, is shown in figure 12-E. Very much like our previous case study with *E. coli*, our agent needs to enact the gradient, because a single detection of the attractant concentration does not suffice to determine a movement up or down the gradient. For doing so the motif performs what is called a differentiation function. The slow inhibitory pathway maintains the running probability constant if the concentration detected some time ago is smaller than the presently detected one. Otherwise, if the previous concentration were to be higher than the presently detected one, the slow inhibitory connection through the inter-node will inhibit the output-node and the running probability would decrease (leaving room for a pirouette, and a subsequent change of direction, to occur). Yet this model is a higher abstract dynamical level of modelling, it is not a direct mechanistic mapping between observable parts and behavioural functions. "Although we have modelled the differentiator motif using neural networks" Dunn explains "it is also possible that these motifs explain processes that function at the intracellular level" (Dunn et al. 2006:552).

Elegantix is now surrounded by *Serratia marcescens* bacteria. It slows down engaging in what is called a "dwelling" behaviour, moving slowly around the area while ingesting some of the bacteria. While *Elegantix* is feeding, Cornelia Bargmann takes the chance to remind us that the neural chemosensory system of *C. elegans* also accomplishes a number of internal regulat-

ory functions: from the detection of appropriated conditions for certain developmental processes to occur, to the regulation of the lifespan of the animal or the regulation fat accumulation (Bargmann 2006). In fact, the NS is not only in charge of mediating interactive agential processes but also many other “internal” functions. In particular, and specially relevant for what is going on inside Elegantix at this moment, we find a neuromuscular “mechanical” control of the pharynx, where a sophisticated neuromuscular system is in charge of filtering water to trap and introduce bacteria into the intestine (Avery & Thomas 1997). In the pharynx, the muscles are connected by gap junctions and their synchronized and asynchronized action potentials (at a scale of milliseconds) produce “swallowing” patterns moving water back and forth, trapping bacteria in the process and spelling the waste water out. Although the muscles can perform their regular synchronized contractions on their own, they appear surrounded by a ring of 20 neurons that regulate the timing, initiation and frequency of contractions and the relaxation of the muscles while loosely couple the pharyngeal behaviour with the somatic NS. “The rate of pharyngeal pumping is regulated by the presence of food, the nutritional state of the worm, and neurotransmitters such as serotonin. Most of this regulation requires MC neuron” and, they explain, another neuron labelled NMS probably releases serotonin under the presence of food, communicating with the rest of the organism. That could be responsible for the slowing down of motion that occurred when Elegantix reached the bacterial colony. But before I can confirm this thoughts with my colleagues, Prof. Bargmann interrupts. Something is going wrong: *Serratia marcescens* bacteria are rapidly infecting Elegantix. Although other variants of this same bacteria are not infectious, this variant seems to be capable of killing our friend in less than a few hours. Horrified under the possibility of loosing our friend, I tap Elegantix on its head to make it move backwards. The nematode makes what is called an omega turn, changing 180 degrees direction and I can see how the pharyngeal system inhibits pumping when touched. This halting of the ingestion process is due to the activity across the gap junctions between the extrapharyngeal RIP neurons and the pharyngeal I1 neurons (Avery & Thomas 1997). Reinforced by my initial success on saving Elegantix from a major infection I now start tapping it on the tail to make it move forward and escape. But it soon habituates. I look desperately around to our colleagues, powerless on my attempts to make Elegantix move away from the infectious bacterial colony.. Caroline Bargmann smiles, I should not worry, she explains, “*C. elegans* modifies its olfactory preferences after exposure to pathogenic bacteria, avoiding odours from the pathogen and increasing its attraction to odours from familiar nonpathogenic bacteria” (Zhang et al. 2005: 179). And she proceeds to show us the result of her experiments carried out together with Yun Zhang and Hang Lu. We all gather around neurons ADF, AIY and AIZ (this last one

was already familiar to me, I remember it from the neuronal knot mediating chemotaxis). The bacterial infection is soon detected by *Elegantix* body. We can see how serotonin production increases on ADF sensory neuron through the activation of gene THP-1 (among others). In turn, serotonin affects the activity pattern of interneurons AIY and AIZ through Cl⁻ serotonin gated channels called MOD-1 (see figure 12-F) changing the sensitivity of the chemotactic circuit to the attractant odour of bacteria. As a result *Elegantix* is now changing its behavioural preference and moving away from what turned out to be a poisonous banquet. It seems time to bring our journey to an end, now that *Elegantix*, save from pathogenic bacteria, moves away looking for additional food sources, preferred temperatures and/or appropriate places for egg laying; ready to face the situations in might well encounter in the near future.

It is now time to generalize a little further than what our case studies and the history of agential evolution on planet earth might permit. There where, no doubt, important advances on the mechanisms and the dynamic organization of agency observed on the case study of *C. elegans* but I should make them explicit on a wider framework: that of behavioural agency made possible by the NS, embedded and embodied, as we know, on a bodyplan able to exploit the *possibilities* it brings forth. It is precisely these possibilities that deserve a careful analysis since they contain necessary requirements for cognition to arise. The next section is devoted to abstract some important properties of the neuromuscular system, within the general biological organization, that make it amenable to a particular kind of modelling: dynamical system modelling. The case study of *C. elegans* will help us illustrate this domain and generalize its most relevant features.

6. THE HIERARCHICAL DECOUPLING OF THE NERVOUS SYSTEM

Since the very beginning of its evolution, as we have seen, the electric conductivity that cells are able to sustain made possible an extended network of dynamic variability capable of coordinating distant cells in multicellular organism, generating endogenous rhythms that could further be affected by sensory signal transduction mechanisms producing different sensorimotor couplings between organism and environment. However, epithelial conduction alone was not able to exploit the implicit potential that electric conductivity contains for intercellular coordination. Cnidarian agency, sustained on elongated cells specialized on signal transmission (the neurons), made possible a more specific and agentially relevant use of electric conduction. Yet, Cnidarian agency appeared limited by a radially symmetric bodyplan and its

primitive developmental mechanisms. It was the Bilaterian bodyplan, together with the appearance of the developmental regulatory capacity of the Hox gene family, what permitted a much wider exploration of what neurons (and bodies) could do for agency, as it became apparent on our journey with *Elegantix*.

But... what makes neurons so special? Why is a NS such an important aspect of behaviour and cognition? In short, its significance relies on the fact that the NS can generate a high dimensional dynamical organization of behaviour which is plastic and stable at the same time and free from a high number of limiting constraints that other modes of intercellular coordination present. This is due to the capacity of the NS to generate highly plastic and reconfigurable electrochemical patterns. Not less important is the role of neurodynamic patterns on the control and regulation of internal metabolic and organismic processes. We can make this more explicit if we abstract a set of properties of nervous cells and their interaction:

- **Fidelity of signal:** neurons can connect to farther distances through their prolonged axons without losing effectiveness on the signal transmission due to the active ion pumping properties that allow currents to propagate. Although we know that neurons in *C. elegans* do not spike, in bigger animals action potentials change in spikes thus increasing the fidelity of signal for distances longer than 1mm (above which the amplitude of the voltage potential decays to a third of its original value). In addition, many neurons in higher animals appear wrapped with myelinated Schwann cells electrically isolating their axons. Through the myelinated part of the axon electric action potentials are conducted at high speed and the spike is regenerated at the nodes of Ranvier (gaps between myelinated portions of the axon).
- **Integration and differentiation of signals:** neurons *can* innervate specific effectors, connected from different sources (sensors) and connect to each other in highly directional and selective manner (provided that an adequate developmental process exists capable of situating populations of neurons in an appropriate distribution).
- **Modulation of signals:** neuronal interconnections are mediated by chemical synapses, where molecular interventions and modulation of the propagating signals can occur, thus providing a selective transformation of the electric impulses making inhibitory connections possible (unlike epithelial conduction or neural connections mediated by gap junctions) and permitting signal dependent long-term conservation of variability

These three properties are necessary to produce a high dimensional dynamic domain: If signals were not generated and regenerated as spikes then the domain would be severely limited by spacial constraints, rich internal dynamics

being confined to local circuits (possibly not being able to coordinate full body movements in big organisms). If no integration and differentiation (i.e. selective connectivity) were possible then signals would diffuse spatially (like in the case of epithelial conduction) or would be confined to specific nerve rings (like in Cnidarian NSs) and, as a consequence, the effective dimensionality of the system will be severely reduced. Finally, if no modulation was possible, then long term variability within the system would not be conserved. Just to give a hitch over the dynamic possibilities that these three properties permit it is worth noting that CTRNNs (Continuous Time Recurrent Neural Networks), mathematically similar to those used on Dunn et al.'s simulation (2004, 2006) and much less powerful than biological neural networks (in terms of the dynamic complexity of each node) have been demonstrated to be able to smoothly approximate (given an open number of nodes) any possible dynamical system (Funahashi and Nakamura 1993). For this to happen units with non-linear transfer functions need to be in place, together with a high and non-homogeneous interconnectivity between such units. And this is, grossly speaking, what a NS is: a potentially universal approximator of any possible dynamical system.

Thus, what makes neural interconnections so special is that they create an incredibly rich and plastic internal world of patterns of fast connections, a high dimensional dynamic domain, as characterized before, that is *hierarchically* decoupled¹⁵ from the metabolic processes (Moreno & Lasa 2003, Barandiaran 2004, Barandiaran and Moreno 2006a). By the term *hierarchical decoupling of the NS from metabolism* I mean that the metabolic constructive cycle of the basic autonomous organization (its basic autonomy or metabolic individuality) generates and sustains a dynamical system (the NS) minimizing its local

15 A phenomenon of decoupling appears when a system gets organized in such a way that 1) a part of this system constitutes a new level of interactions which operates according to a set of rules independent of the dynamics of the low level (the remaining system) and 2) both levels become causally connected in such a way that they depend on each other. Thus, each level appears as a relatively autonomous system, though, as I have stressed, in the end both are mutually dependent. The very same phenomenon took place at the genesis of adaptivity: the appearance of the nervous system does not but follow this trend, albeit with a much powerful capacity and decoupling. Whereas Ruiz-Mirazo, Umerez and Moreno (2008) have preferred the term “dynamical decoupling” I have favoured the previous term “hierarchical decoupling” (Moreno and Lasa 2003, Barandiaran 2004) for a number of reasons. First and foremost the term “dynamical decoupling” seems at odds with the additional “dynamical coupling” that is established between neural activity, the body and the environment. The terminological choice is crucially determined by the departure point one opts for: ontological or epistemological. Departing from an ontological viewpoint (following Pattee's approach) Moreno and Umerez (personal communication) take the term “dynamical” to mean rate-dependent intrinsic temporal properties of a system. Hence the decoupling phenomena appears as dynamic in nature: two systems obey different intrinsic rates, they are dynamically inconsistent. From a modelling perspective the ontological approach needs to face the problem of explaining how is it that when making a predictive and explanatory model for the system both (inconsistent) sub-systems appear dynamically coupled (in the mathematical sense). Therefore I have favoured the term “hierarchical decoupling” meaning that two, distinct, levels of description and variability co-exist within a system but remain, however, dynamically coupled to one another.

interference with it (or maximizing its dynamic variability across time and space). The term “hierarchical” refers to the fact that metabolism produces and maintains the architecture of the NS, specifically, it provides energy and matter for its constituent cells—the neurons. On the other hand, the term *decoupling* means both: a) that neurons act as minimizing the interference of their local metabolic processes with their ion-channelling capacities (which is another way of saying that neurons are specialized on the maintenance of electrochemical dynamics) and b) that the metabolic-constructive organization of the organism under-determines the activity of the NS, which depends on its internal dynamics, coupled through signal transduction mechanisms with dynamic regimes that are external to the NS (in particular with the environment of the organism and its internal processes). Operationally speaking if we are to predict the state of the NS, hierarchical decoupling means that neither local states of cell metabolism (other than severe disruptions) nor the state of metabolic organs (digestive, circulatory, breathing, etc.) alone are going to be very useful. On the contrary the electrochemical states of other neurons and their embodied sensorimotor coupling with the environment may provide a much better model for prediction. For instance, *C. elegans* fat accumulation, digestive molecular processes, growth, oxidation, etc. should be taken as conditions of possibility for the activity of the NS but not as determinants of its dynamics.

Thus, the biophysical specificity, high connectivity, embodiment and situatedness of neural electrochemical dynamics make it “irreducible” to the metabolic substrate of its constituent components (the neurons) and the organismic context of processes of self-construction and repair (state of other body organs and processes involved). Thus, the NS becomes a relational domain in itself (Maturana 1970) whose relation with anything external to its electrochemical dynamics needs to be established through specific transduction mechanisms. This is not to say that the neural domain is independent of its underlying (biological) supporting architecture. On the contrary, the activity of the NS depends on its participation in the logic of the global maintenance of the animal (its metabolism requires an adequate sensorimotor activity) and, on the other hand, metabolic organization supports NS’s construction, functioning and maintenance. Thus, NS and metabolism are connected in such a way that their respective maintenance (and therefore, existence) depend upon one another. Yet their processes appear locally decoupled and “globally” coupled only through specific channels¹⁶.

16 This is not surprising if we look at how different systems of organs are specialized in multicellular organisms, specially in animals and particularly in the Bilaterian bodyplan. The *C. elegans* anatomy can be divided on three major systems: a) the “metabolic-digestive” system in charge of transforming external nutrients into metabolizable components for the rest of the organism and expelling waste products, it is mainly composed of the mouth, pharynx, intestine and anus; b) a reproductive system composed of the uterus and gonads; and c) the agential or sensorimotor system composed of a variety

It is within this domain, made possible and necessary by multicellular adaptive motility, that cognition is to be found. The organism is now “free” to generate within the domain of the NS a highly dimensional and loosely constrained dynamics of fast and slow variability, with its own endogenous rhythms capable of transforming, in a coordinated manner, almost any kind of sensory variability into, potentially, any kind of effector variability. Earlier adaptive subsystems were highly constrained on their complexity growth by the material constraints (molecular-shape and chemical reactivity) and the interference with the rest of (more fundamental) biological processes (those of metabolic re-generation, i.e. the constructive cycle of their basic autonomous organization). Thus, the NS permits the creation of a genuinely *open-ended* adaptive sub-system. The processes leading to bacterial chemotaxis have been simulated using artificial neural networks (Di Paola, et al. 2004) but it is evident that the number, connectivity and parameters of the artificial neurons used in the simulation (if the variables of the model were anyhow to map into observables of the system) could not be optimized or modified but under severe constraints. This limitation also appeared evident on Albert’s simulation experiments of *Hydra*’s epithelial conduction where the parameters that the simulation had to take to adequately satisfy adaptive demands were over the limits of what biological constraints of epithelial conduction could accommodate (Albert 1999). This is not to say that biochemical pathways or epithelial conduction possess no plasticity but that, *comparatively*, the appearance of the NS implies a qualitative change of possibilities. The consequences for the evolvability of agency are evident if we take a look to the enormous flexibility and variety of the interactive motile organization of those organisms endowed with NSs.

Briefly subsumed, on the above characterization, under the terms “signal transduction”, “embodiment” and “situatedness” the environment and the body are called to play a crucial role in this story. However, it is in relation to the dynamic domain that it constitutes that NS, body and environment are to become entangled to give rise to a particular type of agency: adaptive behaviour. It is thus of fundamental importance to specify with more precision how the neurodynamic domain can be abstracted and modelled so it can later be embodied and situated to give rise to adaptive behaviour. This requires to determine the variables incorporated in NS modelling together with the observables and operations that shall permit to relate the model to natural systems.

of sensory receptors, neurons and muscles.

7. DYNAMICAL MODELLING OF THE NEURODYNAMIC DOMAIN

The emergence of the NS as a decoupled dynamic domain carries with it the appearance of a set of variables, functions and parameters on which its dynamics depend. Yet, the notion of variable pertains to the domain of the model, understood as a formal construction, and not the postulated entities or processes we are modelling (Kampis 1991). Thus, a precise specification of a modelling domain requires to make explicit how the variables of the modelling domain relate to observables¹⁷ of the natural system and how the choice of such observables is co-determined by a set of theoretical assumptions regarding the causal organization of the phenomenon under study.

As it is by now clear, the normative functionality of the NS lies on the interactive regulation of boundary conditions of the autonomy of the organism through motility and other regulatory functions (internal to the organism) that need to be coordinated through it. Thus, specifying which observables are required to construct the domain of modelling requires that we pay attention to the function of the NS and to its characteristic decoupled nature. In other words, the relevant variables of any model of the NS are those that permit to predict how the NS satisfies its function within the autonomy of the organism. This is a problem that (implicitly or explicitly) even the most ruthless reductionist neurobiologist needs to address; she has to take decisions over the immense (almost infinite) combinations of operations she could potentially perform within the anatomical limits of the NS (even escaping the question of what constitutes such limits). The question can be very clearly expressed in the following ways: What are the specific operations of measurement and manipulation that need to be carried to do neurobiology? For instance, where should the intracellular recordings be done? why there and not somewhere else? Why is it relevant to do a laser ablation on a particular cell (the neuron) and not in another (a glial cell)? Which staining antibodies should be used, to which molecule should they bind? Which is the gene expression that should be monitored? And more importantly.. which is the criteria that is used to make these decisions?

Although it is rarely made explicit, the generally assumed operational criteria is the one that responds adequately to a further question: Which are the observables (and their specific relationships and properties) that best allow us to build a model that predicts how the NS connects sensors and effectors so as to satisfy its function within the organisms (i.e. to coordinate behaviour and internal regulatory processes)? Thus, the neurobiologist, does not observe or

¹⁷ By observable I do not mean that something need to be directly observed but rather measured or detected by different mechanisms. In neurobiology in particular, as we saw in the case of *C. elegans*, a variety of techniques is used to measure or detect different molecules and precesses that participate on the activity of the NS, only a small portion of which can be observed through the electron-microscope.

operate upon cellular/molecular/genetic changes of neurons (e.g. mitochondrial processes) *unless* they are thought to have a relevant effect upon the way in which the NS operates as a network of *signals*¹⁸ (either by altering its membrane properties or its synaptic efficacy). In more abstract terms, within the almost infinite space of possible observables (cells, molecules, changes on ionic concentrations, etc.) that are held inside the limits of the NS, the functional criteria isolates a dynamically (causally) relevant hyperplane where neural processes occur.

According to these criteria it is commonly accepted that the primary operational primitives are the change of *membrane voltage potentials of neurons* over time which conserve and transform dynamic variability in terms of absolute value, spike frequencies, time distance between spikes or other higher level patterns. *Dendrite-axon connections*, on the other hand, mediated either via gap junctions or synaptic gaps, specify a connectivity matrix (the transformation functions between primary operational primitives). In turn, *neural modulators* (local and global synaptic modulators and action potential threshold modulators) become secondary operational primitives (since they become operational primitives in virtue of their effect on the spikes) and can be considered as different variables or parameters of the transformation functions of membrane action potentials (or voltage).

The molecular nature of synaptic connections and the molecular structure of ion channel gates along the membrane make action potential dynamics depend on the activation of specific genes that are used as templates to create specific proteins. In turn, certain neuromodulators trigger a cascade of internal reactions inside neural cells activating, inhibiting or promoting certain genes. These relationships give rise to a deep causal entanglement (bottom-up and top-down interactions) from genes to protein synthesis, to changes on permeability, to spike frequency and ultimately to changes in the behavioural

18 The term signal often slides into otherwise chemical or physical descriptions without explicit justification. It is certainly a key transitional term between the language of chemistry and physiological mechanisms and the language of communication and cognition. Its usage (and particularly scientific modelling in terms of signalling) generally hides a number of crucial implicit assumptions and deserves a few words. First and foremost, signals imply a sort of decoupling from the causal efficacy of the most basic physical properties of the interacting components of a system (such as mass, force, energy, etc.). It is often helpful to think in terms of formal causality as distinct from material or physical causality, involving processes that are causally connected in terms of topological (shape), combinatorial (order, composition) or undulatory (e.g. frequency) properties. At some point this formalcausal chain (e.g. the diffusion of proteins with a specific shape or the electric action potentials) is transformed back into physical/material causality by means of a transducer (e.g. a muscle generating mechanical work, a binding protein that triggers a cascade of chemical reactions, etc.). In-between we speak of signals. Note that in order to introduce the vocabulary of signals we need not necessarily appeal to cognition, although some “interpretation” or “transduction” mechanisms need ultimately be in place so that one type of causality be translated into another. The terms signal can be appropriately be used when there is a domain of processes (electric or chemical) whose mode of causal integration on a wider systems is not determined by the intrinsic gross physical properties of such processes but in virtue of their dynamic variability and a set of transduction mechanisms.

response. Molecular neurobiology is delivering an increasingly accurate functional mapping relating under which activity dependent conditions certain genes are activated and how the molecules that are thereby produced affect the action potential activity. Once this mapping is available the molecular details become transparent, even unnecessary in order to build a predictive model of the action of the NS. What matters, ultimately, is the change of state that can propagate along the network, and that change will, sooner or latter take the form of an action potential. Crudely put:

From a dynamical systems point of view, neurons are excitable because they are near a transition, called bifurcation, from resting to sustained spiking activity. While there is a huge number of possible ionic mechanisms of excitability and spike-generation, there are *only four different bifurcation mechanisms* that can result in such a transition (...) A biologist would say that the response of a neuron depends on many factors, such as the type of voltage and Ca^{2+} gated channels expressed by the neuron, the morphology of its dendritic tree, the location of the input, etc. These factors are indeed important, but they do not determine the neuronal response per se. They rather determine the rules that govern the dynamics of the neuron. *Different conductances and currents can result in the same rules and hence in the same responses, and conversely, similar currents can result in different rules and in different responses.* The currents define what kind of a dynamical system the neuron is. (Izhikevich 2007:1—6, italics added).

Thus, we can say that the molecular mechanisms underlying the functioning of the NS can be translated into a set of dynamic rules that govern the electrochemical behaviour of neurons. And after these rules are in place the molecular mechanisms underlying them become negligible if we are just to model the causal organization of behaviour.

Yet, isolated neural spikes alone do not make agential organization possible, because they operate in relation to the action potentials of the neurons they are connected with; and the properties of their connections can be modulated by different neurons and molecular processes. Even if we had a full dynamical model of each neuron, together with the connectivity matrix and the dynamic rules governing the modulatory process, we would need to determine which features of spikes (and their modulation) are determinant for the organization of behaviour. Thus, higher level variables might appear as relevant: the interspike intervals, spike train coincidence, the frequency of spikes, their synchrony and rhythms, etc. The search of these higher level variables constitutes the search for a neural code: i.e. what is dynamically conserved or modified at the global scale so as to generate functional changes on behavioural organization. This use of the term code is, in principle, devoid of any semantic or computational load. As Friston puts it: “a neuronal code is a metric that reveals interactions among neuronal systems by enabling some prediction of the activity in one population given the activity in another” (Friston 2000: 219). Hence, once a lower level model of neural activity is in place neuroscience can proceed to build higher level models that capture the causally rel-

evant organization. So for instance at the level of a single neuron a single spike may turn out not to be relevant but only the frequency of the spikes it generates. If that were the case, then the level of modelling may include a higher level of abstraction and focus exclusively on the rate of spikes, so that certain small modulatory changes be neglected or simplified accordingly. Similarly at the collective level the activity of a neural ensemble can be reduced to a simpler higher level dynamic model if the sub-network is redundant or its full specification comprises more information than what is required to predict its effects on the rest of the network.

Thus, the dynamical modelling of the NS requires a previous process of analysis which abstracts away a number of levels of mechanistic granularity, particularly the anatomical and molecular levels. The anatomical organization is translated into a connectivity matrix and, when distances are sufficiently important so as to have relevant temporal effects, these can be translated into temporal delays on the model. On the other hand, observable molecular details can be abstracted away in terms of additional functions, parameters and variables of the system that represent the collective effect of molecular changes on the action potential dynamics (including the growth of new dendritic branches as changes on the functions connecting different neurons). This is not to say that molecular mechanisms or anatomical details do not matter. They do, but the relevant issue is to elucidate how and why they matter. In this sense *mechanisms are important to manipulate the system, to know what (electrochemical, macro-molecular or genetic processes) and where (anatomical structures) needs to be changed so as to affect localized processes in order to achieve a certain global dynamic change.*

To illustrate these crucial issue (which ramifies along reductionist and methodological debates) I will make use of our journey with *Elegantix*. We can take the example of, *mod-1* mutant, a *C. elegans* that fails to retain the association of the odour of infectious bacteria with an avoidance behaviour. If we carry on behavioural experiments we discover that other modes of association are intact and function appropriately so that the nematode can regulate its behaviour according to other chemosensory associative correlations. Thus, according to its dynamic organization of behaviour, the agent lacks the capacity to conserve changes in internal variables for only a given a certain type of sensorimotor coupling. This could be named a “selective memory problem”. If we want to solve this problem we need to know the mechanisms that sustain the behavioural organization of the worm. Yet “the problem” is defined in terms of behavioural dynamics and, accordingly, its solution must have a dynamical effect on behaviour. Thus, given that the problem is not generalized to all types of memory (which could imply a generalized molecular deficit), the solution may involve the change of the relationship between some specific action potentials in the appropriate areas of the NS. We can look at the vari-

ables (neurons) and relationships (pathways) mediating chemotactic preference and find the relationships that could be critical. We know from Zhang et al. (2005) that serotonin receptor MOD-1 in neurons AIY and AIZ are the most likely loci of aversive learning of odour preferences related to bacterial infection after ingestion. Thus, if the current worm is lacking MOD-1 due, for instance, to a genetic deficit (like that shown by *mod-1* mutants) it could be “fixed” with the “expression of *mod-1* cDNA from a *mod-1* promoter” (Zhang et al. 2005: 182). Molecular details are thus relevant to discover how to solve the “memory problem”; but this does not, however, imply that MOD-1 reception of serotonin *is or equals-to* “memory of aversive odour preference learning”¹⁹. Other mechanisms could do and, in fact, do perform similar functions. Unfortunately, we admittedly lack a detailed knowledge of the voltage dynamics of *C. elegans* NS and how local synaptic modulation affects them at the global behavioural level. This knowledge is crucial to reconstruct global behavioural organization from molecular details²⁰ and to discover what exactly is memory of aversive preference learning, not in terms of its correlation with particular molecular details, but within the full context of the neurodynamics organization of behaviour. As we gain an increasingly detailed knowledge of this intermediate level, other mechanistic “solutions” to the problem, not involving MOD-1, could equally be feasible: a rewiring of the network, a potentiation of other types of modulatory signals, etc.

Another example may come to clarify the point I am trying to make. Chang et al. (1998) induced in *Aplysia Californica* the expression of octamine receptors endowing *Aplysia* with a completely new response capacity to octamine (previously not a neuromodulatory compound in *Aplysia*’s sensory neurons). This new excitability response acted as if it were serotonin, which is the main neuromodulatory compound for different forms of learning on the well studied withdrawal reflex in *Aplysia*. Induced expression of a receptor mechanisms for a previously non-functionally integrated molecule makes this molecule (octamine) have the same neurodynamic and behavioural modulatory effect as did what was previously thought to be *essentially* linked with such neurodynamic effect. Thus, the contribution of a particular molecular mechanism is depicted in the context of the dynamics of the rest of the network

19 The same could be said of *mod-1* gene. It cannot be said to be *the* cause of this particular type of learning within a context and even a hierarchy of contexts from genetic, to synaptic, to cellular, to network, to behavioural. This is so because the set of contexts are continuously changing so that no *ceteris paribus* clause can be invoked so as to render the gene the unique variable cause of behaviour or the chief among the multiple but subordinate causes. It is but one among the requisites for transforming behavioural organization in a particular direction. The fact that there are a few specific mutations that are correlated with relatively isolated behavioural effects does not suffice to support a full reductionist program.

20 At the extreme case a model of neural dynamics could be reconstructed from a full knowledge of all the molecular and electrochemical mechanisms involved. But we do also lack this knowledge and we can’t expect to achieve it any time soon (if ever) to such a degree of accuracy.

and other mechanisms could equally produce similar or functionally equivalent effects. This fact makes molecular implementation details relatively independent of the neurodynamic organization of behaviour and its causally relevant modelling.

In addition, functional mechanisms are often distributed, involving an often unknown number of relevant processes that altogether contribute to the generation of a particular form of behaviour. The way in which knowledge is achieved about the effective action of serotonin release of neuron ADF and MOD-1 reception in neurons AIY and AIZ for aversive learning is illustrative. Exhaustive behavioural tests are made upon different mutant variations and conditions. Selection of mutants is guided by the heuristics of known anatomical and molecular details but equally limited to those mutations that have only selective effects on certain neurons and under certain conditions (most of mutations are pleiotropic and their effects propagate affecting different developmental and organizational processes of the organism rendering it inviable). When significant correlations appear (such as learning impairment correlated with *mod-1* mutations) additional tests are carried out (like staining immunoreactivity) to examine the causal relevance of the hypothesized MOD-1 receptor. Yet, admittedly, the understanding is far from complete. The correlation is statistically relevant but not complete: many details are missing and a whole bunch of neural and molecular interactions that are also considered critical to explain the phenomena are still poorly understood. Only one of the causally relevant mechanisms has been discovered. This is a great achievement but is far from a complete and satisfactory explanation of the processes that make aversive learning possible. In fact, despite the paradigmatic pressure to localize functions and achieve a clear reductionist account of the functioning of the system, the impossibility to create a simple one-to-one mapping between ‘molecular structure’ to ‘behavioural function’ is a generalized situation along the literature. Although specific functions were, for simplicity, attributed to specific neurons throughout our journey with *Elegans*, most of the literature generally uses terms like “neuron/molecule/gene X is required for behaviour Y” or “neuron/molecule/gene X takes part on behaviour Y” or “behaviour Y is partially/normally/completely impaired under ablation of neuron X” or “performance of Y behaviour is reduced when X is ablated or when gene/molecule X is absent”. The effect of molecular deficits and genetic mutations is highly limited and only occasionally determinant of a specific phenotypic change (Lewontin 2000). This is due to the fact that many developmental processes are robust to specific mutations due to redundancy and many other molecular effects are pleiotropic (have multiple effects) and reconfigure the phenotype restabilizing the organism on alternatively viable configurations.

A different, and complementary approach to the genetic/molecular correlational study is to build a dynamical model at a higher level of detail. This is the alternative taken by Dunn et al.'s approach to chemotaxis and thermotaxis mentioned previously (Dunn et al. 2004, 2006, Dunn and Lockery 2007). A relatively abstract dynamical model of *C. elegans*' NS is created where the equations governing the activity of the neurons approximate the behaviour of the real neurons. Next, from a well defined sensory and motor dataset the parameters of the neurons and their connectivity are optimized to match the data. A number of solutions are available after the optimization procedure is repeated. All these solutions are clustered into a reduced number of dynamic networks motifs. Each of these motifs represents variations of the same dynamic strategy required to match the recorded sensorimotor correlations (i.e. chemotaxis or thermotaxis). At this point the resulting motifs can be tested in a number of ways. For instance some could be ruled out due to the impossibility of the real system to achieve certain parametric configurations. The motifs can be, and are in fact, tested, embodied and situated in simulated bodies and environments to see if they match qualitatively the behaviour of real nematodes under different varying conditions. After these tests have been carried out the amount of potential candidate motifs is very reduced, to a number of two or three. What is lacking at this point is the capacity to match the motifs with specific and testable mechanisms. However, as is made explicit on a recent paper (Dunn et al. 2007), despite their similarity the nodes of the network should not be confused with real neurons. Dunn et al. distinguish at least three possible mechanisms that the each node of the dynamic model may be representing: a) a group of co-active neurons, b) different steps in the subcellular signalling pathways of a neuron, and c) a higher level combination of a and b and other factors that may contribute to the dynamic organization of behaviour. Although not explicitly addressed by the authors among this additional factors other non-neural components could equally be included: the body's physical response delay, its embeddedness on specific hydrodynamic conditions etc. (but I shall take care of this feature on the next section). What remains important is that what has been modelled is the dynamic functional organization of behaviour: an abstract generative system whose level of abstraction does not necessarily match the molecular level but still captures the relevant causal organization of behaviour to the extent that it qualitatively and quantitatively matches the observed behaviour. Once the necessary mechanistic details are available to disambiguate between the available dynamic functional models it is the organization of behaviour of that model that matters and, conversely, only when a dynamic model is available can all the mechanistic molecular, physical and environmental factors be functionally integrated in a model. To sum up, the NS shows itself as a decoupled domain whose modelling permits, and requires, the abstraction from local mechanistic-molecular details and the specification of the NS as a high dimen-

8. THE EMBODIMENT AND SITUATEDNESS OF NEURAL DYNAMICS

As we have seen, the appearance of multicellular organisms endowed with a (sub)system allowing fast, efficient and plastic agency, was necessarily accompanied and enabled with other important changes in their internal multicellular organization. At this stage of our morphophylogeny we can make explicit some more fundamental properties of natural behavioural processes that are found to be at the basis of all cognitive processes. Some of these properties may already appear on earlier (even unicellular) agents, specially in those based on motility and endowed with sensory and motor mechanisms, but their full fledged significance shows up only in neurally guided behaving systems: sensorimotor embodiment and situatedness. It would be misleading to reduce the dynamic organization of behaviour to the NS or even to try to model the dynamic organization of a NS in isolation. Any adequate model of agency needs to include not only the sensorimotor context in which the agent is situated but also the body. If we retake the example of the “memory problem” of *C. elegans*, other solutions not involving operations into the dynamics internal to the NS are feasible. For instance the nematode could regurgitate the infectious bacteria together with some intestinal acid and aversively associate the odour of the bacteria with its smell. This way the impairment of reception of ADF serotonin release can be circumvented with an alternative associative mechanism that is achieved, and this is the important point, *through the consequences of the actions of the agent on its environment*. In such a case the solution does not come “just” from a change on the neurodynamic organization of the agent but through the way in which the agent exploits its interaction with the environment by releasing some chemicals that it can in turn sense. Alternatively a change on sensory modality may provide a good solution (an artificial sensor could be implanted on the intestine) or a body response like a muscular shrinking could be sufficient (provided that proprioceptive mechanisms are also available and that new associations can be made upon such proprioceptive sensing). In both cases it is a body change what becomes crucial to solve the behavioural “deficit”: i.e. to implement or perform the same dynamic function. Any agential organization involves a considerable degree of recurrent interactions between body and world. In other words the NS does not operate in isolation, behaviour is a feature of the coupled NS, body and environment systems and their interaction need to be called into any explanation of the organization of behaviour (Beer 1995, Chiel & Beer 1997).

Little attention has been paid to the body throughout the explanation of *El-e-gantix* behaviour. This was partly due to the widespread assumption, in the most of the neuroscience literature dealing with *C. elegans*, that motor neurons “execute different motor programs” for each type of behaviour (forward

mouvement, omega turns, pirouettes, etc.). An auxiliary assumption is that how the body is differentially stretched and shrunk by the activation of muscles and body morphology is subsumed within a *queteris paribus* clause: it doesn't change or its change is regular enough so that it can be functionally ignored as not affecting the operations of the NS. Yet "what a motor program is" can not be explained without reference to the body shape, muscle disposition, etc. There is no little pilot waiting for abstract motor commands (turn left, right, pirouette) to read and execute them into the stream of possible system-to-environment effectors. There are, however, some very illustrative models of how different motor neurons can affect muscles to produce specific coordinated movements through the shape and articulation of the body in the context of a given environment (Neibur and Erdös 1991, Ferree, Marcotte and Lockery 1997). The body is maintained cohesive thanks, among other things, to a collagenous cuticle which is secreted by the epithelium. The interior of the worm has a high hydrostatic pressure so that the volume of its body does not change very much. As a result, the differential contraction of ventral and dorsal muscles produces the bending of segments of the body which, in friction with the environment, give rise to undulatory forward and backward movements. Thus, in more general terms, the NS is embodied on specific tissues capable of channelling metabolic energy into efficient mechanical energy (muscle contraction) and a system of fixation (along with the mechanical cohesion of the body) transforms these contractions into unitary directional movements. All this is achieved tangentially to the continuous process of metabolic self-maintenance, reproduction and morphological transformations that the organism undergoes by means of digestion, cell replication and growth. The evolution of agency does not but increase this tendency shown in *C. elegans*: a specific articulated systems of fixation develops (the skeleton) and muscles achieve modularized, specialized and more sophisticated forms.

Hans Jonas wonderfully summarized the implications for agency of this mechanical embodiment of animal motility in comparison with that of plants:

[A]nimal motion is more than merely an intensified or magnified case of organic motion in general. We may note its difference from vegetative motion in these physical respects: in speed and spatial scale; in being occasional instead of continual; variable instead of predefined; reversible instead of irreversible (...) The next obvious distinction of animal from vegetative motion is its variability and the revocable nature of the changes which it performs. A process of growth or formation runs, on the whole, on predetermined tracks, and its results are there to stay (to be added to, perhaps, but not to be reversed). Animal motion, on the contrary, consists in the operation of structures whose future condition is not prejudged by their acts: throughout the train of action and at the end of it the condition remains what it was at the start, viz. that of free motility. The motions of limbs are reversible, and the change which they bring about is not a new state of the animal but a new state in its relations to other things. It is an unaltered animal which proceeds from this to the next motion, and the limb resumes its proper function each time under identical structural conditions. In other words,

the operation of a limb does not entail its own and thereby the agent's transformation in the process. This is equivalent to saying that a limb is a mechanical structure, and animal motion is outward motion, which as such changes only spatial relations and leaves the nature of the agent unaffected (Jonas 1968: 248—249).

In this passage Jonas perfectly captures the transition from the notion of generic biological agency to the more specific one of embodied behaviour characterized by mechanically articulated, reversible and spatially structured interactions. Only under these conditions can the environment for the system become a geometric space where objects can be freely explored. When robotic engineering is downplayed in contrast with the organic complexity of biological bodies we tend to forget a crucial aspect. What seems almost straightforward for human engineering capacities (the construction of wheels, articulated arms, etc.) was an astonishing achievement for nature: the generation of mechanically articulated motion from the coordination of millions of biochemical cellular units!

But mechanical embodiment, as I will call it, does not only free the organism from irreversible morphological transformations: it also limits the movement, biasing it towards specific interactions. Thus, the body's physical and mechanical properties shape possible interactions and relative positions through enabling biomechanical *constraints*. The space of motor outputs to be instructed by the organism is not a uniform multidimensional space defined just by a number of degrees of freedom. On the contrary, embodied motion defines a biased "landscape" within that space; determined by the shape, elasticity of joints, relative orientation and a host of alike body constraints. An extreme case of embodied motion is given by dynamic walking models (McGeer 1990) where, even in the absence of neural control, a mechanical system (a pair of legs) determines a well structured environmental coupling with the ground's surface, giving rise to coherent and robust walking behaviour. Instead, in the same situation, a "disembodied approach" to behaviour would have required an exhaustive and detailed motor output, resulting from an expensive processing of anticipatory trajectories drawn upon a representational modelling of the surface features, the angles of the joints, etc. This type of computational control of behaviour is characteristic of robots in television series and popular culture (and many research labs). Yet when humans try to imitate such stereotypically over-coded movements it turns out to require a high controlling effort and practice, in contrast with the synergetic coupled behaviour that results when we "naturally" walk or move on everyday life, in the absence of a conscious effort to explicitly control each movement. None, or very little, of this control is required if one takes into account the mechanical embodiment of behaviour. The "righting" behaviour (the change from downward fishing position to the upward swimming position) of *Aglantix* provides an illustrative example of this type of fluid embodied interaction,

where a simple asymmetric bending of the margin of its bell triggered a full process of turning round as a result of the hydrodynamic properties of its body.

Yet the mechanical is not the only relevant aspect of embodiment, the body is also the surface from which the agent is linked to its environment through sensing. And here we encounter again a biased transformation of the coupling due to body and sensor properties. Embodied sensory surfaces are not continuous full-range measuring devices but, on the contrary, appear limited and specialized on specific ranges, transformations and filtering of sensory perturbations. In the case of *C. elegans*, for instance, the distribution of mechanosensory detectors along the body is strategic. When touched on its head *Elegantix* made an omega turn, and the process did not require that the nematode had to represent where the stimulus came from, calculating a complex function on the process. Nor it had to decide or to calculate the trajectory of the appropriate movement to make, because the sensor is located in the anterior part and a direct signal to specific neuromuscular structures suffices, through its embodiment, to turn around 180 degrees. Chemosensation is also very specialized for a limited, yet rich, variety of chemical components that bind to specific molecules in the membrane of sensory cells, thus providing relevant cues for adaptive interactions. In more sophisticated sensory systems embodiment is also crucial. For instance, the position and shape of the ears, the cochlea, the position of the eyes (focused to the front in predators and displaced on the sides on preys, for example) become embodiments of sensory surfaces that exploit physical and relational features to transform environmental perturbations into functionally pre-structured signals, and they already modulate and sustain the behavioural flow in a certain manner.

Finally, both sensory and motor embodied surfaces appear highly intertwined due to the circular and recursive nature of sensorimotor interactions that have evolved and developed together. I will call these *enabling constraints* because they bias the potential dimensionality of the sensorimotor coupling of the NS so as to enable or facilitate self-organized developmental and adaptive interactions (cross-modal sensorimotor spaces, developmental scaffolding by bodily changes, structural adaptation with certain object size and shape, etc.). From a computational perspective, embodiment also means that much of the cognitive processing is carried out as embedded on the structure and mechanical functioning of sensorimotor processes. For instance information theoretic measures of how diverse sensory morphologies (specifically the distribution and density of sensory units) affect situated information flow has been precisely quantified on artificial systems (Lungarella & Sporns 2006).

As a consequence of their sensorimotor biomechanical embodiment behaving systems are **situated systems**, their relation with the environment is rel-

ative to their situation on it and this situation is both the effect and cause of its sensorimotor coupling. In other words, the agent is in continuous feedback with its environment and its engagement can provide as much information as can neurodynamic modelling in order to explain behaviour. An extreme case is found in the nematode *Ascaris* (Neibur and Erdos 1991, 1993). When moving in soil its neuromuscular system produces regular burst so that the body segments can bend coordinated. Yet, when moving in water a change of neuromuscular activity is required so that movements can keep coordinated due to the mechanical properties of the medium, the same neural motor signals that propagate when the nematode is situated on a soil environment will not lead to a coherent coordination. How is this adaptation achieved? Interestingly, it is the viscosity of the water what makes movements slower; this in turn delays the proprioceptive feedback that self-regulates the waveform signals that control body segments for coordinated movement. In *C. elegans*, on the contrary, it is hypothesized that the nematode senses in which environment it is situated and changes the neuromuscular waves accordingly. Yet *Ascaris*' embodied situatedness does not require specific perception of the aqueous environment to change neuromuscular rhythms for coordination, the situatedness of its *body in mouvement* suffices for adapting to the new medium. Just put the nematode in water and the very dynamics of movement in the new liquid environment will be adjusted without the mediation of specific sensors and neural processes recognizing that the environmental context pertains to a different type and that changes need to be made accordingly. Randall Beer (1990) has provided beautiful examples of similar and more complex adaptive motility in exapod walking models where the adaptation to leg disruption, differences on surface friction etc. are the result of the feedback-between different legs *via* the environment and the full body (without mediation of functionally specialized regulatory mechanisms).

Yet, one of the most relevant aspects of situatedness is that of relative position in a geometrical space. In this sense, and contrary to the most generalized examples when talking about dynamical modelling of cognition (van Gelder 1995, Chiel & Beer 1997, Bechtel 1998, etc.), the behaving organism is not coupled to the environment as a Watt Governor is coupled to a water flow²¹. Sensory input is not only a function of the environment (as a set of dynamic variables) but a function of the controlled *relative position* of the agent in its environment. Simply put, geometry cannot be formally, analytically, reduced to differential equations. And sensorimotor coupling is, primarily, a coupling between a geometrical space and a dynamical system. The coupling

21 Chiel and Beer (1997) do not make use of the metaphor of a Watt Governor but their paper is probably one of the main references on the conceptualization of cognition as coupled dynamics and no reference is made to the special status of the environment as a geometrical system.

is asymmetric on that: a) whereas the modelling of the environment must include the agent (at least its position and orientation) the modelling of the agent does not include the environment (only its sensor and motor transfer functions)²² and b) whereas we need not model the internal geometrical space of the agent we do in fact need to do it for the environment and the agent within it. Yet geometry, and the particular situatedness that it permits, is not a feature imposed only by the environment itself. We “know” as external observers, that the environment has certain properties (such as a number of coordinates, objects, gradients, etc.) and for ease of interpretation and familiarity, we tend to model it conserving its geometric structure. However, from the point of view of behavioural dynamics in many occasions, even involving motility, the environment could be reduced to a string of values with no geometric structure. For instance in the case of *E. coli* and *C. elegans* doing chemotaxis in an homogeneous gradient the environment can be modelled as a variable that has the following changing function: if the bacterium or the nematode are moving on “running” mode (i.e. forward motion) then $y = gx$ (where g is the slope of change of the gradient, y the local concentration on the immediate surrounding of the agent and x the velocity of straight movement). If the bacteria has engaged on “tumbling” mode or the nematode on “turning” mode then we reassign a random value for $g = [-G, G]$ where G is the maximum gradient in the environment. Due to the internal dynamics of the control system of both bacterium and the nematode we know that y will tend to have, statistically a high value. No geometry is required, the fact that the bacterium is situated in a geometrically distributed gradient is absolutely irrelevant from the point of view of the dynamics of behaviour. Certainly the environment and behaviour can become more complex, for instance when the gradient is not homogeneous and the above simplification

22 For the modeller the following argument may be of help. The idea is to make explicit what is implicit in most simulation models of adaptive behaviour uncovering some hidden assumptions about the relationship between dynamics and geometry: We say that A (the agent) is coupled with E (the environment). However, looked up more carefully, when programming a simulation, it turns out that we consider to be a system E is really a combination of E and A . For instance... if A is considered a dynamical system that is coupled to E as another dynamical system then, is the position of the agent a parameter in the equations governing the state of the environment? No, because the environment is not really like a dynamical system it is rather a geometrical system. See it this way: the environment (as space) is not defined by coupled differential equations, what you are doing in the simulation is to define coordinates (which are independent variables). You define objects by their shape and their position, but the position is not a state. Imagine that they were states. Let us consider that the environment is a set of states of position variables, pairs (x,y) for objects A (agent) B and C ; so $E = \{Ax, Ay, Bx, By, Cx, Cy\}$. Now where is the coupling between these variables and the sensory (S) and motor (M) parameters of the agent? Nowhere as such. For each time we are to calculate the state of S we need to calculate geometrical, spatial relationships between Ax, Ay, Bx, By, Cx and Cy . So now let's do it, let's say that the environment is this set of distances so we can really couple them to S . $E = \{Dab, Dbc, Dac\}$... now what is the coupling between the agent and these new variables of the environment? There is none as such, the motor parameter does not couple to Dab, Dbc or Dac , we need to calculate the coupling through the old $E = \{Axy, Bxy, Cxy\}$ system. And we have not really faced the problem of orientation or embodiment of the agent, just a problem of position in two dimensions.

might not hold any more. However, the example was provided as an illustration of how the fact that an agent's behaviour occurs in a geometric space is not sufficient to deduce that the agent's environment, from its own perspective, has any relevant geometric properties. The argument can be extended, more generally, for any other properties of the environment that the scientist might model: such properties are not a real part of the environment of the agent independently of the structure of its behaviour. As Poincaré noted more than 100 years ago (Poincaré 1895) the geometric properties of its world are due to the agent's sensorimotor situatedness on a geometric environment *and* its capacity to enact geometric properties (such as continuity of space, dimensionality or homogeneity), through certain sensorimotor invariances (like active visual tracking, reversibility of perceptions and invariance of shape upon movement around an object).

Thus, the situatedness of an agent is not merely the trivial result of its sensorimotor embodiment but also, and crucially, of its neurodynamic capacity to exploit its sensorimotor coupling with the environment. When this occurs the organization of behaviour appears distributed on a set of feed back loops between the NS, the body and the environment. This coupling between a geometric space and a situated agent has more consequences than what it might become trivially apparent. For instance, functional behaviour cannot be taken to be exclusively the result of extracting statistical properties or patterns from a string of predefined sensory inputs and the production of an adequate response output. Situatedness provides much more complex and flexible possibilities for functional action. For example, a non-situated agent whose control architecture is reactive (i.e., whose output is determined by the instantaneous input by a non-modifiable internal structure) cannot solve a non-markovian task, i.e., cannot successfully classify an environmental condition if its detection requires to extract a sequential (timely) order, when the condition of the environment cannot be reduced to an instantaneous sensory value. On the contrary, a situated system with a reactive controller can transform non-markovian tasks into markovian tasks only by means of exploiting its relative position on its environment (as beautifully shown in a conceptual model by Izquierdo-Torres and Di Paolo 2005). A similar conclusion was reached by Clark and Thornton (1997) by exploring how, through situatedness, previously intractable problems could be easily transformed on tractable ones for certain types of neural network architectures.

This is mainly possible through the geometrical situatedness of an agent, what it senses is not just a feature of the environment but a feature of its relative position on it and this position is in turn a function of its motor activity. This circularity between the state of sensors and motors may happen in cases where no geometric coupling occurs. For instance in the case of a biochemical agent affecting its environment through chemical effectors that can have re-

active effects on its environment that can in turn be sensed. But it is in the motile agent, coupled to a geometric environment, where this sensorimotor coupling through the environment is particularly unavoidable and essential; almost any movement will have immediate effects on its sensors. As Powers noted long ago: “What an organism senses affects what it does, *and what it does affects what it senses*. Only the first half of that commonplace observation has been incorporated into most psychological concepts of nervous system organization. The effects of behavior in altering subsequent stimuli, and even in directly causing stimulation, have certainly been noticed, but there has as yet been no correct analysis of this in any fully developed psychological theory” (Powers 1973:41). Powers is perhaps one of the earliest cognitive science researchers on stressing the importance of action for perception in terms of control theory and dynamical systems, to the extent that he conceptualized a whole theory of cognition based on the control of perception at difference levels of abstraction. But Powers was certainly not the first to address the importance of the sensorimotor cycle. Among others, American pragmatists based their conception of mind on action and inverted the traditional conception of sensation as the unidirectional cause of action. John Dewey synthesized this view very clearly: “Actions are not reactions to stimuli; they are actions into the stimuli.” (Dewey 1914)²³ or “The only events to which the terms stimulus and response can be descriptively applied are to minor acts serving by their respective positions to the maintenance of some organized coordination” (Dewey 1893)²⁴. Equally phenomenologists, like Merleau Ponty, also stressed similar conceptions of perception and action as circularly intertwined (Merleau-Ponty 1942, see Thompson 2007 for a current upgrade of Merleau-Ponty’s ideas). And also Piaget (1967/1969) stressed a unified view of the sensorimotor cycle against the stimulus-response paradigm. Following these authors it can be said that environmental invariances are often “*created*” through sensorimotor coupling and not just passively perceived by the senses or extracted through complex inferential and representational procedures.

Thus, situatedness permits to achieve functional behaviour exploiting relative position and orientation and, in more abstract and general terms, exploiting the effect of motors into sensory states through the properties of the environment it is coupled with. In this sense motor activity can be interpreted as

²³ Quote borrowed from Freeman (1997b).

²⁴ The following passage deserves a full-length quote: “The ordinary interpretation would say the sensation of light is a stimulus to the grasping as a response, the burn resulting is a stimulus to withdrawing the hand as response and so on. There is, of course, no doubt that is a rough practical way of representing the process. But when we ask for its psychological adequacy, the case is quite different. Upon analysis, we find that we begin not with a sensory stimulus, but with a sensori-motor coordination, the optical-ocular, and that in a certain sense it is the movement which is primary, and the sensation which is secondary, the movement of body, head and eye muscles determining the quality of what is experienced. In other words, the real beginning is with the act of seeing; it is looking, and not a sensation of light.” (Dewey 1893:358)

controlling the value of sensory variables or paraphrasing Powers' seminal work: behaviour is the control of perception. This feature makes possible what has been called emergent functionalism (Steels 1991) or interactive emergence, i.e.: the dissolution of functional structures in highly interactive loops between agent and environment (Clark 1997, 1996). I shall explore the consequences of situatedness in more detail later. What should remain clear at this point is that neural dynamics are not only constrained by sensory and motor embodiment but also and importantly by the situated closed loop between sensors and motor that the environmental coupling permits or affords.

However, the NS is not only sensorimotorly embodied and situated in an external environment. As mentioned before, the hierarchical decoupling of the NS from the rest of the organism includes an additional coupling or embeddedness on its metabolic substrate: what I shall call **biological embodiment** (including both metabolic and reproductive embodiment²⁵—I will, however, for the sake of simplicity, focus on metabolic embodiment). Our journey with *Elegantix* has provided a number of illustrative examples. For instance, when the nematode encountered food, neuron NMS from the pharynx released serotonin producing a slowing down of movement to stay in place and explore the local environment while ingesting. Equally some means of body-to-NS communication was required for the aversive learning of odour preference when the infection took place (presumably some neuromodulator release under the presence of infectious bacteria in the pharynx). Thus, a second sense of situatedness needs to be called into play, that which appears due to the biological embodiment of the agent, sensory (and effector) cells coupled to an “internal environment” that in turn is coupled to the external environment: the content and state of the pharynx and the intestine are the result of the behaviour of the system, but they also activate and modulate behaviour.

I will distinguish the new form of adaptive agency based on motility of those multicellulars endowed with a NS controlling a mechanical body as properly **behavioural agency**²⁶. It is precisely the hierarchical decoupling of the NS

25 By this term I mean not only metabolic self-maintenance but a number of organizational constraints derived from the evolutionary dimension of living beings (reproduction, kin caring, sexual selection and mating, etc.).

26 It is the sense of behaviour as sensorimotor coupling with the environment that I want to highlight as different to other uses of the term “behaviour” such as the behaviour of clouds or planets as mere change of position. Instead of “behaviour”, “comportment” is probably a more appropriate term to use Merleau-Ponty's terminology (1942). Moreno & Etxeberria (2005) have termed “Neural Agency” what I have here called “Behavioural Agency”. Yet, although made possible by the NS, it is not the neuronal aspect what is characteristic of this type of agency but rather its embodiment and situatedness. After all, neural like processes are present in plants and other organisms that do not show Adaptive Behaviour as I have defined it.

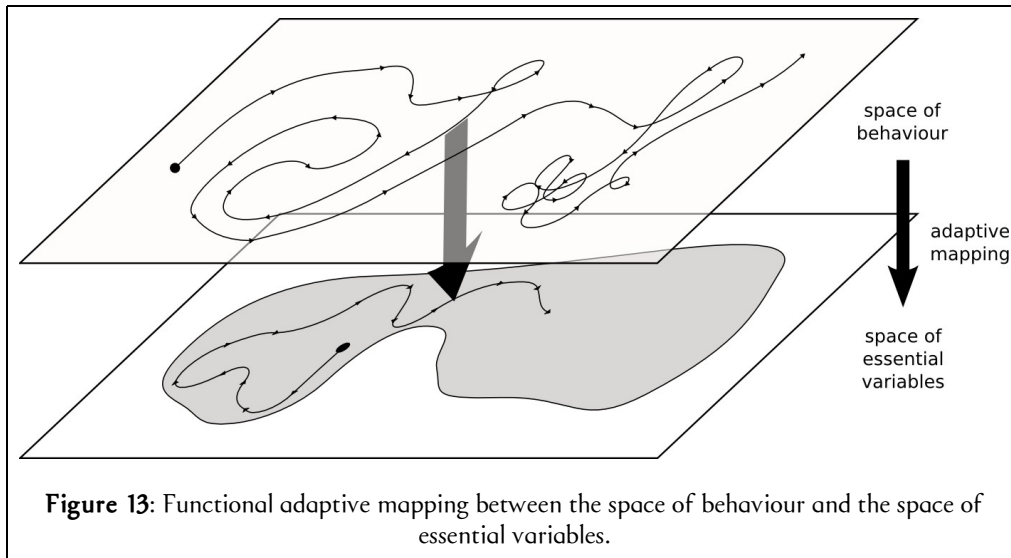
and its sensorimotor coupling with the environment what permits to study adaptive behaviour in terms of sensorimotor dynamics (as it is the case in several fields such as robotics, cognitive neuroscience or embodied psychology) and qualifies behaviour as a specific phenomenon distinct from generic biological agency. In contrast, the explanation of the interactions of plants with their environments would require the introduction of additional constraints such as the rate of growth through cell replication according to exposure to the light, availability of water in the immediate surroundings and a host of alike agency-metabolism interdependencies; making plant movements not as reversible as the animal movement. Although plants can be considered chemical agents, even embodied agents, they do not have a bodyplan capable of behavioural agency²⁷. But, as we have seen, a mechanical sensorimotor body by itself does not suffice, the dynamic organization of the subsystem in charge of interactive control needs to be able to exploit the possibilities of such embodiment and situatedness and this can only happen with the appearance of the NS and its characteristic dynamic properties. Thus, behavioural agency can only develop and evolve above certain degrees of complexity in organisms endowed with a NS and a particular bilateral and mechanically articulated bodyplan.

9. THE ADAPTIVE ORGANIZATION OF EMBODIED AND SITUATED NEURODYNAMICS.

We have justified a dynamical modelling of neural activity and provided support for the crucial role played by its embodiment and situatedness. It is now time to have a look at how the dynamics of behaviour are organized and structured along brain, body and world. We know that the organisms endowed with a NS severely depend on behavioural agency for their ongoing existence, since their biological organization is based on neurally guided sensorimotor interactions that satisfy a number of metabolic, reproductive and bioregulatory functions (energy and matter flow, temperature constancy, physical integrity, etc.). More specifically and recovering our previous definition of adaptation, the normative function of the NS activity (hierarchically decoupled from other organismic functions, sensorimotorly and biological embodied and situated on its characteristic environment) is the adaptive maintenance of essential variables under viability boundaries through the neural sensorimotor control of the interactive coupling with the environment (Ashby 1952, Beer 1997, Barandiaran 2004, Barandiaran and Moreno 2008). A qualitative geometrical picture of adaptive behaviour is a mapping between the

²⁷ Obviously photosynthesis and other properties of plants make them viable organisms but not through behavioural agency.

dynamics of behaviour and the dynamics of the essential variables of the organisms so that behavioural dynamics render possible the evolution of the essential variables within their viability boundaries (see figure 13).



But the fact that the global dynamics of NS-body-environment produce an adaptive maintenance of essential variables under viability constraints does not specify or explain *how* this functionality is achieved. Since it is impossible for metabolic or other biological “needs” alone to instruct functionally such a potentially high dimensional space, the understanding of behavioural agency requires that we make explicit what kind of constraints act on the embodied NS generating functional order. So, the relevant question now becomes that of how is this function or mapping achieved. There are many ways in which such mapping could be satisfied and even more ways in which the system could fail to do so. So the question can be reframed as how are the dynamics of a potentially high-dimensional system constrained, and thereby organized, so as to satisfy the requirements of its organismic embodiment.

From what has been already said in the previous sections we can provide a tentative answer. We have built a conceptual picture of behaviour as the dynamics cutting across brain, body and environment. The properties of the NS allow for a potentially high-dimensional rich and plastic space of behaviour to occur. In addition, we know that from this conceptual and abstract space of possibilities the activity of the NS is limited so as to satisfy certain conditions, particularly those of adaptation for its biological embodiment. Thus, this potentially open dynamic domain appears constrained, biased, canalized or pre-determined in a number of ways; i.e. there are sources of functional order

that need to be specified in order to explain how this dynamic domain is functionally (i.e. adaptively) organized. Some of these sources of order can be considered exogenous, i.e. generated from outside, in the specific sense that they are not the result of behaviour per se: to account for them we need to bring other (non behavioural) dynamics or structures into play (genetic-developmental, bioregulatory and/or environmental). The other type of order can be found in principles of self-organization (both at the neural network and at interactive levels) and we can consider it as an endogenous source of order. Yet, the term self-organization is often used with abuse and does not always help to clarify the issue. What I mean by self-organization here is nothing more than the fact that the states taken by a dynamical system are the result of its own activity. This is, of course, trivially true for any mathematical dynamical system that is deterministic and closed, yet it is not so for a system that is decomposable and/or analytically tractable and conceptualized as linear aggregation of different procedures, algorithms or mechanisms; nor is it evident for a system whose functions have some level of stochasticity and is open to an additionally stochastic or noisy environment. But I will come back to this issue later. By now the main idea is that the way in which the dynamic domain of behaviour is organized can be, in principle, divided on two main groups. Whereas the first can be considered imposed or given from outside the dynamics of behaviour the second can be considered internal or the result or product of behaviour itself. It turns out, however, that, from the point of view of a differential equation model, the source of order is always exogenous if we consider that the functions defining the system are pre-specified. In such a case the system, simply put, is nothing but following those laws or rules. We can think on a system where the parameters of the function are activity dependent (i.e. are modified by the activity of the systems) so that, to a certain extent, we can say that the functions that govern the systems are modulated or (re-)defined from within. Yet, from a modelling perspective, this may turn out to be trivial: the parameters become variables of the system (understood as a whole) and these variables are, in turn, governed by functions that need to be specified. So ... can we say that the system is constrained by endogenous sources if we consider that the functions governing the change of variables are exogenous? Strictly speaking, and from the point of view of dynamical system modelling, we cannot. An alternative, however, is to differentiate between systems that are analytically tractable and those which are not analytically tractable and thus require the reproduction of the states of the system; i.e. systems in which the evolution of behaviour depends on the activity and history of the system. So we can say that the order is determined from outside, that it is exogenous, if we can deduce it when it is reducible to the operations of the parts taken in isolation, and endogenous when we need to consider the collective activity of all the part within an organized whole (it is

in this latter sense that I will use the term self-organization in the neurodynamic and behavioural domain). Now, we are ready to start making distinctions over the types of constraints or sources of functional order in adaptive behaviour (see figure 14 for a diagram of the following classification).

The first type of exogenous constraints²⁸ can be called architectural naming what is given, not variable, in the dynamic modelling of the NS and its interface with the environment: sensorimotor embodiment, neuronal type properties and the connectivity matrix (together with the relevant time delays representing spacial distribution, etc.). Simply put, in terms of dynamical modelling: the parameters and functions of the NS and its coupling transformation functions with the environment. These constraints are the result of genetically triggered developmental processes that have been selected in relation to the behaviour they produced when coupled to the environment. So, it is to be expected that at least to some degree, they are able to assure the adaptive mapping. On the one hand sensorimotor embodiment constraints the coupling between NS and the environment reducing considerably the modes of interactive control (some features of the environment would not be possibly sensed and degrees of freedom of the body will be constrained). Yet, how the body is articulated and muscled, what type of sensors (modality range, filtering, transduction, etc.) it is endowed with, etc. is exogenous in relation to the behavioural dynamics of the agent. These features are not (to a certain degree) the result of its activity as an agent and are often considered congenital. On the other hand, given a specific sensorimotor embodiment, the circuit diagram can completely determine behaviour: such is the case of the so called “hard-wired” circuits: escape behaviour in *C. elegans* provides an example (recall the example of the 180 degree turn made by *Elegantix* when touched on the head) but any reflex-arch in higher vertebrates is also an instance of it. The term “hard-wired” is, however, of little help and in most cases can lead to confusion²⁹. For instance a continuous time recurrent neural network (CTRNN

28 Together with the term self-organization that of constraint can be equally confusing and misleading. What I mean by constraint here is, simply put (and following Ashby 1962), that the uncertainty over the abstract space of all possible values of the observables of the system (e.g. membrane electric potential) becomes reduced via correlations or conditioning (i.e. non independence) between the variables.

29 The widespread use of this term is probably due to two main factors. One is the widespread use of feed-forward neural networks to illustrate the functioning of the NS (particularly on philosophical approaches). These networks have no feed-back loops and the state of the network (and ultimately its output) is completely determined by the input and the values of the connection weights. Early dynamic and recurrent neural network model such as that of Hopfield networks (Hopfield 1982) were not really much different. The dynamics of the network settle down in a particular attractor and this attractor is again determined by the input and the network parameters, the network is disembodied and there is no room for functionally relevant transient dynamics. The other factor is that search for plasticity in neurobiology has been paradigmatically focused on synaptic plasticity and dendritic growth as mechanistic instantiations of behavioural plasticity so that networks without synaptic plasticity are taken to be severely constrained on their capacity to provide behavioural flexibility. There are good paradigmatic “reasons” for doing so: methodologically speaking when neural activity is studied in

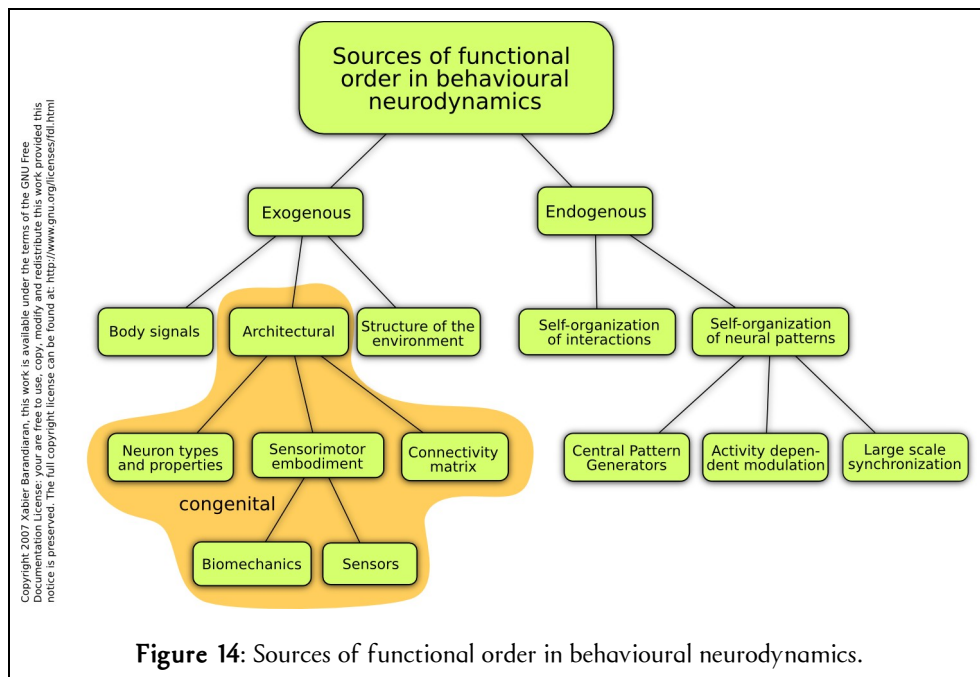
hereafter) can be completely hard-wired in the sense that the parameters of the connection of the variables (the nodes) can be completely fixed and yet the internal dynamics could be extremely plastic and complex. For instance, it can be capable of learning, different forms of adaptation, multiscale dynamics, etc. (Tuci et al. 2002, Izquierdo-Torres & Harvey 2007). Simply put, a CTRNN can approximate just any kind of dynamical system, however hard-wired we want it to be. What matters, in terms of how much of behaviour can be specified by architectural constraints (particularly neural ones), is the *effective* dimensionality of the system, the shape of its internal landscape (number of attractors, etc.) and the degree of decomposability of the system. For instance, in the most simple case, if the system has just a single attractor and a uniform basin of attraction, behaviour would be completely specified by its architectural constraints. Another example can be found when the network is highly modular and decomposable. Here the phase space could be broken down into more simple (lower dimensional) sub-spaces and linked to each other via effectively low dimensional channels that activate or deactivate different modules. In such a case the effect of behaviour on its own organization is “just” limited to the activation or deactivation of such modules and their internal dynamics and, again, if the dynamics of each module is simple enough so that modular input-output relationships follow predetermined regularities, we could consider that neural architectural constraints almost fully specify adaptive behaviour. Thus, for example, certain mechanosensory circuits in *C. elegans* are often taken as modularly organized in the sense that specific sensory neurons activate different pathways leading to specific and highly stereotyped responses.

Obviously there is much room for complexity even in a “simple” NS such as *C. elegans*. The examples and descriptions of architecturally specified behavioural interactions were given in order to illustrate what is meant by the type of external constraints on behaviour that I called architectural. It is a question of degree to which extent can externally defined architectural constraints determine the course of behaviour under specific environmental conditions. But one way or the other the architecture of the NS constraints to a certain degree what the system is capable of doing. On top of this type of constraints other operations and processes (also external to the activity of the NS) can be added. The second type of external constraints, or modulation of constraints, we can call into play are those that have a very special type of effect upon the architectural constraints: those produced by strong perturbations of neural dynamics through specific signals (pain, hunger, satiation, pleasure, nausea, etc.). They operate modulating the overall activity of the NS (generally

isolation voltage dependent dynamics are very volatile (in the sense that global brain dynamics are absent) and research tends to focus on the interaction between two levels: membrane action potentials and synaptic modifications.

through specific neural modulatory pathways such as the dopaminergic system). I will call them *body signals* since they are generally originated on different states or conditions of the body (e.g. metabolic or sexual needs). These signals can be considered external from the point of view of the activity of the NS: when do these signals appear and how they operate does not *directly* depend on the behaviour sustained by the NS. In fact, from the point of view of mathematical modelling, these signals should be considered inputs, rather than constraints. The role of endogenous body signals is to *regulate* the behaviour sustained by neural activity in relation to the adaptive needs of the organism: acting as switches, changing the membrane properties of neurons and activating dendritic growth among other things. For instance, in *C. elegans*, egg laying behaviour is considered to be regulated through the action of serotonin (an important neurotransmitter that also has the capacity to change neural membrane properties) over the neural circuits controlling the neuromuscular structures of the vulva. Equally serotonin was the main neuromodulator for aversive learning to pathogenic bacterial odour in *C. elegans*. In more general terms what I have here termed modulation of the NS architecture by signals has been intensively studied under the label of reinforcement learning when the signals involve a long term functional-adaptive change of the architecture (Staddon 2003) and as internal drives in the classical literature in ethology (McFarland 1985). The name “value signals” has also been used to name the modulatory adaptive effect of signal originated in the body and certain neural structures (Edelman 1987).

Finally, environmental external constraints need to be taken into account: what the environment affords in terms of the sensorimotor engagements it elicits. In the most simple case a given condition or property of the environment (temperature, presence of a chemical compound, etc.) will be transduced by the sensors and through the network into a particular response. Yet, it could equally be said that the organism selects a given environmental cue to satisfy some internal need, in the sense that it is not the environment imposing upon the organism a particular behavioural response but the organisms defining its own internal state what determines which of the current environmental variables may serve to guide its behaviour. However, from the discussion on situatedness described along the previous section, we know that in many cases it will not be a simple stimulus response reflex but a continuous feed back involved on the production of behaviour. Thus, it is not this or that particular feature of the environment (understood as a “stimulus”) but the overall structure of the environment what becomes an external source of order or constraints over possible interactions. Gibson’s ecological realism (Gibson 1979) is probably one of the strongest defences of the importance of the structure of the environment for the explanation of behaviour. Thus, and putting aside the effect of more sophisticated forms of behaviour, such as niche construction



(e.g. a spider web), the environment can be taken as a source of *external* constraints that collaborates on the production of behaviour without it being the product of behaviour other than by the selective engagement with some of its features.

But in most of the known neural systems the complexity of the possible neural dynamics appears under-specified by these constraints: just by looking at the neural architecture, the type and effect of endogenous body signals, and the structure of the environment the observer scientist might be unable to predict what is going to happen by purely analytic means. Nonetheless we still find functional order, the organism behaves coherently in relation to its adaptive demands. Therefore, different (non-external) principles of order are required to explain it. It is the creation of internal and interactive constraints that needs to be invoked here. In short, the dynamics of the NS enter a process of *local* and *interactive* self-organization through the recursive activity of neural dynamics and sensorimotor interactions. The problem lies, as I mentioned before, on the difficulty of ascribing this form of order either to the activity itself or to architectural and environmental constraints which will ultimately determine the activity of the system. The way out of this dilemma, however, is to establish the distinction between determination and specification. The notion of determination, if we take the universe to be composed of laws and particles subject to such laws, is not of much help if we are looking for a criteria to make distinctions over types and sources of order. The notion

of specification on the other hand (with a stronger epistemological load) is of more help, because it implies, as suggested above, what needs to be done in to understand, explain or predict the functioning of the system as a whole (it is the Aristotelian notion of causality that is invoked here: i.e. what is necessary to bring into the explanatory process to understand the system under study). If what is required in order to model the system is to decompose it into analytically well distinguishable and understandable functional parts, then we may well say that the system is fully specified by its architecture in combination with environmental and internal conditions. If, on the other hand, understanding of the system requires that we reproduce (through simulated models—numerical methods) the dynamical unfolding of the coupled agent-environment system, then we can say that the system is not fully specified by its architecture and that it is the activity of the system what additionally specifies its functioning.

A paradigmatic example of local self-organization in neural systems is that of CPGs (or Central Pattern Generators): CPGs are neural ensembles that create multistable patterns (systems with different attractors) that generally control motor neurons or directly the muscles (Arshavsky et al. 1997). The generation of rhythmic patterns by a neural ensemble is a necessary requirement to constrain the degrees of freedom of the muscular system and create a coherent and coordinated movement (we found examples of this type of structure as early as in Cnidarian rophalia but the best understood and studied CPG is that of the crustacean stomatogastric system). Some of the neurons that compose the network generate oscillatory rhythmic patterns on their own and coupled with the rest of the neurons of the circuit, they are able to generate a global stable pattern of oscillations. The CPG is capable of maintaining this pattern under a wide range of internal and external perturbations including lesion of part of the circuit. But a CPG combines stability and variability along different robust attractors, allowing different neuromodulators and neighbour neurons to affect the circuit and shift to a different patterns according to the different motor needs of the organisms: the same circuit may thus instantiate different functions. In addition, even if the CPG of two members of the same species generally produce very similar patterns, the circuits may be remarkably different in structure. Thus, different circuits can generate the same pattern, and the same circuit can generate different patterns. This property is taken to be characteristic of self-organized systems (Kelso 1995: 239—243). These patterns can be thought of as determined by the architecture of the circuit plus some external modulation that marks the transitions from one attractor to another. However, neither patterns nor transitions can be directly mapped to a functional decomposition of the architecture. The patterns emerge from the collective activity of the neurons that compose the CPG and their connections. In this sense, in close analogy with the chemical domain,

the activity of the circuit is said to self-organize on a set of stable macroscopic patterns.

But self-organization also appears at the level of the coupling between NS, body and environment. We have seen how situatedness provides opportunities for self-organized processes to occur through the environment: the effect of sensory perturbations propagates recurrently through the network generating muscle contractions, which in turn feeds back to sensory neurons both through the changes that movement induces in the immediate sensory environment and through proprioceptive feed-back. The recurrent embodied coupling of the NS to the environment results in adaptive behavioural patterns whose functional stability is the result of the dynamic integration of neural, body and environmental features. Examples of this kind of interactive self-organization are given by active categorical perception, optic flow navigation, swimming, and a set of well known and studied phenomena. Situated and autonomous robotics (Maes 1990, Brooks 1991, Clark 1997, Pfeifer & Scheier 1999) has provided a set of insightful models of embodied and interactively self-organized behaviour: obstacle avoidance (Brooks 1990), wall following (Steels 1991), active categorization (Cliff et al. 1993) and a number of other interactive behavioural phenomena that exploit recurrent interactions with the environment.

So far, thus, neural dynamics can be captured through the specification of a) a neurodynamic architecture or connectivity matrix, b) a set of endogenous body signals with high modulatory capacity that reach the NS through specific channels c) the structure of the environment, d) local self-organization of neural circuits and e) interactive self-organization of behaviour. So, the simplest picture of adaptive behaviour may be seen as that of the embodied and situated dynamics of the NS constrained by its architecture and modulated by body signals to generate, when coupled to the environment, a (self-organized) behaviour that maps into the maintenance of a number of essential variables within viability boundaries. The mapping is assured by a) the history of natural selection operating upon the developmental processes that lead to an architecture capable of satisfying that mapping when coupled to a characteristic environment in which the organism has evolved and developed; and b) by the regulatory modulation that the endogenous body signals exert upon the architecture and the activity of the network in relation to the adaptive needs of the organism. This is thus, how we get to genuinely **adaptive behavioural agency**.

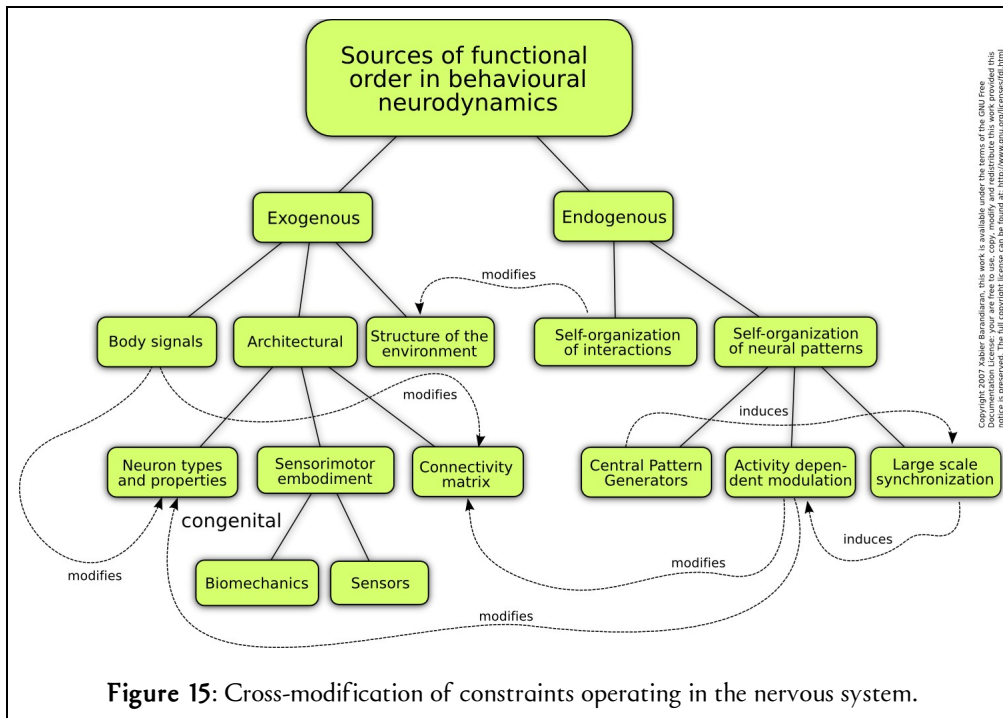


Figure 15: Cross-modification of constraints operating in the nervous system.

However, the picture I have drawn over the type of constraints and dynamics of behaviour is still extraordinarily simple (a more accurate diagram may be that of figure 15). Even with the notions that I have introduced in this section there is room for much more sophisticated interactions: for instance neural modulation can be triggered by neural activity itself, not only by endogenous body signals, thus allowing for an activity-dependent restructuring of the architecture. On the other hand neurodynamic self-organized processes need not be exclusively confined to local circuits (like CPGs) and can extend along the full organization of neural dynamics, including the environment and the body (embodied sensorimotor constraints evolve and develop together with behaviour). As a result, the modes in which the organization of behaviour may lead to biological adaptivity can become far more complex than what I have depicted above, different processes of learning, regulation, hierarchical control, modular specialization and architectural development can appear and be modified according to complex correlations involving bioregulatory, sensorimotor and internal dynamics. However, before I attempt to specify more complex forms of organization, I shall recapitulate to see how adaptive behavioural systems qualify as agents.

10. RECAPITULATION: AGENCY REVISITED THROUGH ADAPTIVE BEHAVIOUR.

We have made a long journey from minimal autonomous systems to neurodynamic embodied agents. From the origins of life and agency, as that of an autonomous organization of a chemical reaction network separated from its surroundings by a self-enclosing membrane, we have scaled up to the level of multicellular agents endowed with a mechanically articulated body controlled by a potentially high dimensional dynamic domain and coupled to its environment through specific sensors and motors. Those readers that found a proto-cellular system pumping ions across its membrane not to be even remotely close to anything like a cognitive system may feel now more comfortable with the adaptive behavioural agency of *C. elegans*. Yet the conceptual trip from basic autonomous systems to adaptively behaving multicellular organisms has uncovered a number of crucial transitions without which adaptive behavioural agency could not be well grounded: from the constitution of a physico-chemical identity and normative functionality to the requirement of a decoupled and open, potentially high-dimensional subsystem in charge of controlling the interactive requirements of its biological embodiment. Through the morphophylogeny of agency we have discovered how the bottlenecks of structural stability, chemical diffusion processes at big sizes, epithelial conduction and radial bodyplans have lead to a number of organizational transitions. These transitions ended up with a Bilaterian bodyplan in which a differentiated neurodynamic subsystem came to be coupled to a mechanical and metabolic body and to the environment to generate adaptive behaviour. However, the case of *C. elegans* may seem, from an agential point of view, not really so distant from that of the motile bacteria except for a set of crucial differences. On the one hand its body is mechanically organized so as to allow a more sophisticated full body bending and coordination. But most importantly, the organization of the adaptive subsystem of the agent (the embodied NS) can potentially encompass (at least in evolutionary terms) an almost open ended dynamic complexity with the subsequent expansion of its agentic capacities. In turn this potentiality can now be exploited by a powerful multicellular developmental process. For many we have reached the point where we can genuinely talk about intelligence and cognition, as the unfolding of an increasingly complex organization of adaptive behaviour (see next chapter for references). Our departure point was to take agency as the foundation for a bottom-up theory of cognition and we defined agents as *systems doing something by themselves according to certain goals or norms within a specific environment*. It is now time to recapitulate how our behaving systems qualify as agents on their individuality, normativity and causal-asymmetry conditions, and discuss if the way in which adaptive behaviour satisfies this

conditions becomes sufficient for a well grounded notion of cognition or if some additional condition or component might still be required.

In the case of *C. elegans* biological embodiment provides the grounds for individuality, very much in a similar vein as it did in the case of minimal autonomous systems: the body of a natural behaving agent is an organismic body, actively defining its boundaries and internal organization through a continuous process of far from thermodynamic equilibrium self-maintenance (digestion, metabolism, repair, etc.) and the construction of an epithelial cuticle that defines its boundaries. Yet, unlike unicellulars based on motility, neurally organized multicellular systems possess a much more complex individuality subject to a sophisticated developmental process, requiring some means for intercellular coordination and cohesion, the maintenance of a homeostatic internal milieu, an immune system capable of defining and defending the internal milieu from external cells, etc. This type of organismic individuality depends to a high degree on the activity of the NS both at the level of internal developmental and organismic regulation and at the level of behavioural adaptivity so that its behaviour and its identity as an organism appear intimately intertwined.

As it happened with the case of minimal agency the environment of the behaving organism is also partly co-defined by its form of individuality. Thanks to the endogenous body signals the sensorimotor flow is in continuous feedback with the internal milieu of the agent. Thus, the world of a behaving organism is the expanded spatio-temporal space that its motile situatedness opens, a world of sensorimotor correlations made possible by the NS, able to coordinate full body movements in a fast and high-dimensional manner. And this environment remains modulated by the “internal” world, bringing forth what von Uexküll (1940:1982) called the *Umwelt* of each organism: the dynamics of the constructive-metabolic cycle “expressed” through body signals harnessing neural dynamics for the active self-maintenance of its organization (in terms of temperature, inflow of matter and energy, outflow of waste products, etc.) and its preservation from the threads that could potentially destroy it (predators, collisions, toxic environments, etc.). This is why certain environmental objects, processes and situations are said to be meaningful for the agent, they jeopardise or enhance their organismic autonomy: a colony of bacteria provides an opportunity for feeding and recovering the energy required and spent on metabolism and movement, changing the direction of movement upon mechanosensation prevents being eaten by a predator or colliding with an obstacle. We have here entered the domain of *normativity*: the agent is autonomously following some norm, what is good or bad, adapted or maladapted is defined by itself, by its own organismic embodiment, without reference to an external observer. Yet, it is within its biological embodiment that this happens. It is the organism as a whole, as a living system,

that becomes a genuine agent in terms of normatively loaded regulation of its interactions with the environment.

Finally the causal asymmetry condition needs to be re-evaluated. The case here is very different from the case of a minimal proto-cellular agent whose actions were clearly against the thermodynamic drive. At the level of minimal prebiotic (chemodynamic) autonomy, energetic considerations made the active vs. passive distinction clear and unambiguous. We could also make indirect use of this criterion for the case of behavioural agency since its embodiment implies a mechanical articulation able to channel metabolic energy into movement. Yet, unlike the case of the minimal agency of BAS the source of energy is external to the dynamic modelling of behaviour. A robotic arm, for instance, capable, through solar panels, of synthesizing its own energy but remotely operated, would not by virtue of its internal energy source be called an agent. Thus, the causal asymmetry condition in the case of behaving agents demands some reference to the issue of control or some sort of dynamical causal asymmetry of behaviour laden to the side of the agent. However, the emphasis put by some adaptive behaviour researchers on the issue of situatedness and interactive self-organization somehow seems to threaten the causal asymmetry conditions. If behaviour is the result of a causal democracy of distributed processes extended along the brain-body-world continuum, it seems impossible or at least difficult to say that any internal factor can be privileged as the origin or cause of behaviour. The locus of agency seems lost. The active and selective sensing of the behaving agent over its environment, the behaviour seen as the control of perception, could provide a tentative solution to the problem. Yet, when modelled as dynamically coupled, the formalization of this agent-laden selective sensing is not a trivial issue, because it is difficult to justify formally whether it is the agent who selects its sensory input or is input that selects a particular response. However, the remark that situated motile agent are coupled not to a Markov chain (a sequence of inputs) nor to a dynamical system but to a geometrical system provides a stronger case for causal asymmetry: it is in virtue of its situatedness that a behaving agent can control its perception. In addition, as we have seen, endogenous body signals, external to the brain-body-world continuum (yet internal to the agent) play a crucial role shaping sensorimotor dynamics. In this sense, even if behavioural interaction may be pictured out as coupled brain-body-world dynamics it is the agent as an organismic individual who modulates which of the potentially available sensorimotor engagements with the world is carried out.

Whereas energy considerations were central for the characterization of agency in basic autonomous systems, the crucial factor that characterizes causal asymmetry in adaptive behaviour is the asymmetry in the dynamic complexity of the interaction, laden to the side of the autonomous system: i.e.

the behavioural mechanism of the system is more complex than the coupled process it sustains. More complex means that the set of internal variables that define the agent are more integrated and functionally differentiated (Tononi *et al.* 1996) than those of the environment involved in the system-environment coupling (for a quantitative comparison between internal and behavioural dynamic complexity see Seth & Edelman 2004). As a result of this complexity asymmetry, the system can be said to be controlling the process and not the environment. Other information theoretical approaches have been recently proposed to measure this causal asymmetry (Bertsinger *et al.* 2008, Seth 2007). For instance Bertsinger and colleagues provide a measure that quantifies the interactive dimension of autonomy as a combination of two factors: (a) non-heteronomy (an autonomous system is not determined by environmental states) and (b) self-determination (the evolution of an autonomous system is determined by its past states, i.e. it is not a random system). The measurement is drawn making use of probabilistic observational means (observing how much of the internal variability is causally correlated with external variability) and, more importantly, combined with causal-interventionist procedures that make explicit how much of the system-environment correlations are due to the pro-active nature of autonomous system (by manipulating internal and external states and observing the result of such manipulations on the production of behaviour). Anil Seth (2007) has proposed a metric (G-autonomy) to measure the extent to which a system is determined (i.e. its model can predict future states) by its own past states compared to the extent it is determined by external states. In accordance with our intuitive notion of agency, it turns out that the more an agent's behaviour is specified by endogenous activity the bigger the causal asymmetry on the production of behaviour and the more agential it becomes.

Does adaptive behaviour provide a minimal example of cognitive agency, a minimal mind? The next chapter will be devoted to analyse in detail this very idea, a widespread position in contemporary cognitive science.

PART III

The Mind has a Life of its own

The question as to the nature of life, I believe, has been finally resolved, and is no longer a philosophical question. I hope something like this will happen to the so-called mind-body problem in the twenty-first century.

JOHN SEARLE

Chapter 7: Failed intentions. Adaptive behaviour as intentional agency?

An adaptive behavioural agent defines its own identity as a biological far-from-equilibrium individuality, in doing so it defines a set of conditions for self-maintenance: a set of needs and threads. When sensorimotor mechanisms permit the system to manage the satisfaction of those conditions through environmental interactions we have an adaptive agent. The evolution of multicellular adaptive agents with a bilateral musculo-skeletal body able to sustain reversible mechanical motion and a nervous system that coordinates it, brings us close to what may be considered as a genuinely cognitive system. And it is tempting to claim that we have reached the final destination of our morpho-phylogeny of agency towards cognition, pictured out as increasingly complex adaptive behaviour.

Under the equation “Cognition = Adaptive Behaviour” (C=AB hereafter) I shall capture the statement that a neurodynamic, embodied and situated model of adaptive behaviour (or even of adaptive agency more generally) constitutes an adequate minimal model for minds. It is certainly the point where we should evaluate such an account of mind and cognition and put it in the broader context of contemporary cognitive science. For doing so I will split in two the hypothesis that “Cognition = Adaptive Behaviour” into “Cognition = Behaviour (sensorimotor loop) + Adaptive (biologically grounded norms that the sensorimotor loop satisfied)”. I make no harm to the plot by anticipating that the hypothesis will run into several problems to ground intentional agency (as the characteristic form of minimal mindfulness). But I will expose those problems by steps and following, to some extent, Di Paolo’s analysis and refinement of autopoietic theory of cognition in relation to teleology, adaptivity and agency (in fact this chapter may be thought of as a direct response to Di Paolo 2005; and partially to Thompson 2004, 2007).

First, I will analyse the “Cognition = Behaviour” and show how it is insufficient to account for cognition as a specific phenomenon different to any dynamical systems (call it “agent”) coupled to another one (call it “environ-

ment”). Then I will introduce the idea, defended by many, that the biological grounding of sensorimotor interactions can solve the problems posited by a purely behaviouristic account of cognition. Next, I will try to show that the existing formulations of the necessity and sufficiency of biology for cognition have not been formulated with precision and that additional requirements need to be added. Thus, I will expand the “Cognition = Adaptive Behaviour” into “Cognition = Adaptive Behaviour + X + Y + Z” where X, Y and Z represent those additional requirements that bring adaptive agency closer to intentional agency. I will try to argue that, despite the adding up of new conditions, the approach remains essentially problematic. To account for these problems I shall rely on a benchmark test: the skyline of our own subjective experience of intentionality-in-action (or intentional agency). Finally I will end up proposing some in-principle reasons for this failure and how we can still save part of the basic intuition by moving the baby to a different bathwater: from biological autonomy to sensorimotor autonomy, from the self-maintenance of metabolic organization to the self-maintenance of the behavioural organization.

1. THE PROMISED LAND?

We have reached the point where the constituent basic/metabolic autonomous organization of living systems has developed a decoupled sensorimotor structure that qualifies the organism as a genuine adaptive behavioural agent. Our bottom-up travel from the origins and minimal organization of life towards the mind has reached a point where a significant number of scientists and philosophers consider that a minimal stance of mindful systems is already at stage. We are thus dealing with the issue of the minimal expression of cognition (or mindful properties) and its relationship with life. This demands that we halt our bottom-up trip to evaluate if we have really reached our final destination: if we have finally entered the lands of cognition within the living continent. If such is the case we would already be in place to draw the boundaries of the cognitive lands. Or alternatively, if its enclosure does not allow for an all-or-nothing distinction, if cognition is to be taken as a mountain whose hillside extends without clear demarcation, the final task we confront would be that of specifying some form of conceptual altimeter that could point towards an increasingly mindful peak (or crest).

After the demarcation is complete, the lands of cognition would obviously still be open for exploration into a further sophistication, to explore a quantitative increase on the properties and capacities that are already found at the bottom level of adaptive behaviour. But the minimal case would be already clearly stated, the essential features or mode of organization characteristic to

mind would be disclosed and the problem of concern in our quest would be solved. The atom model of the mind would be available for further contextual expansion and additional integration of components: other protons, neutrons and electrons could be included on the model (different sensorimotor modalities, learning and imitation capabilities, extended technologies, etc.) giving rise to new molecular and chemical properties and their higher level organization (cultural and social cognition, techno-scientific research, etc.). It thus becomes of fundamental importance that we evaluate in detail if adaptive behaviour is sufficient for cognition.

For many, we have in fact reached the promised land. This is the land of adaptive behaviour (for some it even extends to the earliest lands of motile agency in bacteria) and its placard claims that “Cognition = Adaptive Behaviour¹”. From what I have explained so far about adaptive behavioural agency, it shall suffice to state that this position conceives cognition as stirring from an ongoing sensorimotor dynamical coupling between an agent and its environment. This sensorimotor coupling, however, is not sufficient. The embodiment of the agent serves as a biological grounding so that not only does the system define its own individuality (and therefore distinguishes itself from its environment) but its different sensorimotor engagements become *functionally relevant* for the ongoing self-maintenance of its being and can thus be said to follow a norm defined by the very agent. *C. elegans* moves around its environment sensing different chemicals, associating them with a metabolic value as a food source or as a toxic milieu. It follows adaptive *norms* learning new correlations depending on the effect of environmental interactions on its own body, etc. The integration between ongoing sensorimotor activity and biological self-maintenance is often said to provide genuine intentions to sensorimotor engagements and meaning to its encounters: the organism displays fine grained behavioural strategies *in order to* survive, endowing environmental situations with a *meaningful valence* in relation to its struggles. This intentionality-in-action (if it really is so) is what characterizes and distinguishes cognitive or mindful systems from equally coupled, dynamic and motile artefacts (such as robots). This is the statement that we need to evaluate in detail. For doing so that we can subsume the statement under the equation “Cognition = (biologically grounded) Adaptive Behaviour” (C=AB hereafter).

Among others, neurobiologists like Humberto Maturana and Francisco Varela (1980), Rodolfo Llinás (2001) or Antonio Damasio (1999); biologists like Jackov von Uexküll (1982), Lynn Margulis and Dorion Sagan (1995) and John Stewart (1996, Bourguine & Stewart 2004); roboticists and engineers like

1 Our characterization of behaviour was certainly more demanding (including a nervous system and a mechanically organized body) than what many authors defending this position will require. Throughout this chapter the term “adaptive behaviour” will be used more liberally including what I have termed more generally as “adaptive agency”.

W. Ross Ashby (1956), Randall Beer (1990, 1997), Luc Steels (1991), Tom Ziemke (2001) or, to some extent, Ezequiel Di Paolo (2005); and philosophers like Hans Jonas (1966/2001), Evan Thompson (2007), Fred Keijzer (Keijzer 2006, van Duijn et al. 2006), Pamela Lyon (2006), Wayne Christensen and Cliff Hooker (2000), Mark Bickhard (2004) or Alvaro Moreno and Arantza Etxeberria (2005) can be said to have defended, with some important differences but fundamental similarities, this same basic theme. The differences stir from the additional requirements that are added to a plain system (anticipative character, neural control of behaviour, representational structure of the controlling system, plasticity and learning capabilities, etc.). The point of similarity that permits to unify the above authors around the equation $C=AB$ is the rooting of the intentional (representational or non-representational) character of normativity on the biological organization of the agent and, more specifically on the metabolic self-organized nature of biological systems. It is this unifying theme what is at stake in this chapter.

We have long postponed a dialogue with many of these authors that were side-referenced along our journey and we have now reached the point where, along the well established and specified grounds of adaptive behavioural agency, we can start such conversation. In order to proceed carefully and situate this approach within the larger context of cognitive science I will split the dialogue into the two main components of the thesis: sensorimotor coupling (behaviour) and its (adaptive) biological embodiment.

2. COGNITION AS CLOSED SENSORIMOTOR LOOP

It is moving around, avoiding obstacles and following a light source while compensating different disturbances: it behaves. Whatever *it* is (a robot, an insect, a pet or a human being) we tend to attribute cognitive capacities to this type of systems: we say that it *perceives* the obstacles *in order to* avoid them, that its *goal* is to follow the target system, etc. After all, its behaviour can easily be described by using an intentional vocabulary: as if it had goals, believes, intentions, etc. And, not only described but usefully predicted on the basis of intentional attributions. Philosopher Daniel Dennett (1987) has defended, not without controversy, that there is nothing more to intentionality (i.e. to the essence of cognition) than this possibility of describing the system *as if* it had intentions. To be more precise his claim is that there is no principled way to draw the distinction between “real” and “as-if” intentionality and that therefore, from an instrumentalist position, the distinction is illegitimate. The purpose of this section is to show that this is not indeed the case, that just by looking to behaviour (despite its affording for an intentional attribution) cognition cannot be properly characterized. To say it in other words,

that the equation “Cognition = Behaviour” is not sufficient. And yet behaviour and behaviour generating mechanisms, stand as a crucial departure point for cognitive characterization, in clear contrast with traditional approaches to cognition that took the logical structure of our inner thoughts as a departure point.

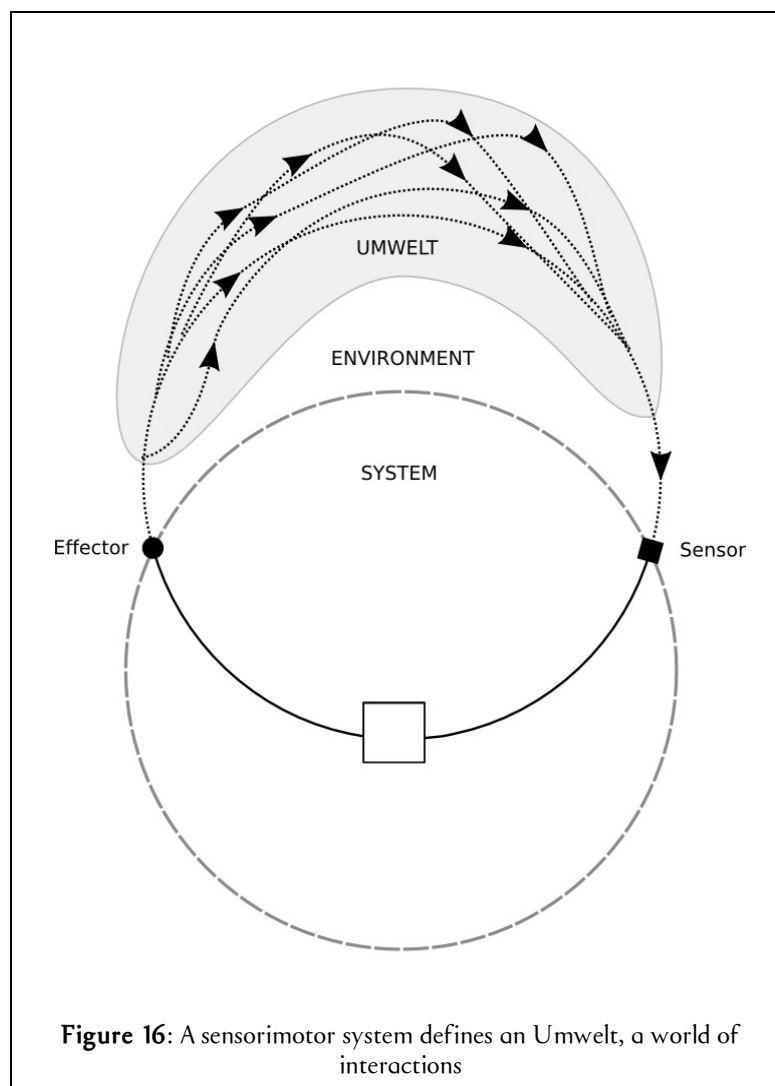
2.1. The sensorimotor “world”

A widespread reference of the sensorimotor coupling as cognitive grounding traces this tradition back to von Uexküll's theory of meaning and the *Umwelt* of animals (von Uexküll 1940/1982). According to the German biologist the environment of a living animal (capable of selectively perceiving environmental features and to “respond” with specific actions) is cut off from a background of undifferentiated environmental properties (or, if one wishes to put herself on God's eye, from the set of all possible relational properties). This selective cut-off, from the array of all possibly detected features of an absolute observer to the functional domain of perception-action cycles, constitutes the *Umwelt* of the organism. What may appear for a human observer as a given object with its *objective* properties appears in fact as carrying a different “meaning” for different animals according to the different modes of engagement that the very same object might elicit for them. Thus, for instance, catching up one of von Uexküll's favourite examples, a flower has a completely different meaning for a cow or an ant. Whereas for the first the flower is a “meaning-carrier” that triggers the action of eating it, for the ant it might just elicit a walking path. Thus, for both animals, different properties of the flower become *perceptual-cues* to carry on different responses or *effector-cues*.

Every action, therefore, that consists of perception and operation imprints its meaning on the meaningless object and thereby makes it into a subject-related meaning-carrier in the respective *Umwelt* (subjective universe) (...) Because every behaviour begins by creating a perceptual cue and ends by printing an effector cue on the same meaning-carrier [object of the *Umwelt*], one may speak of a functional circle that connects the meaning-carrier with the subject. (von Uexküll 1940/1982: 31).

Thus, for von Uexküll, cognition (or a meaningful system-environment relationship) is primarily about behaviour, about how different environmental features are integrated on the sensorimotor cycles of animals cutting-off a world of interactions (see figure 16).

In a similar vein Gibson's theory of ecological perception (Gibson 1979, see Chemero 2008 for an insightful and fresh upgrade of this approach within contemporary cognitive science) conceives cognitive operations as occurring primarily over on-line, although not necessarily continuous, sensorimotor engagement within different environmental contexts. The structure of the environment (endowed with lawful regularities) appears for the agent as eliciting



affordances, opportunities for action, in accordance with the agent's abilities (Chemero 2003). As Chemero puts it (2008), cognition is not about mental gymnastics (about deep thoughts and computational symbol crunching) but primarily about dealing adequately with environmental situations, directly perceiving possibilities for action, contextual opportunities to exploit in embodied interaction. This is how animals (including humans) do it almost all the time. And, if we want to find the roots and minimal expression of cognition, it is upon behaviour and interactive strategies that we should rely as a departure point.

All this is no surprise for us since we have built a picture of adaptive behaviour where what matters is the dynamic organization of behaviour as it cuts

across neurodynamics, body constrains and the environmental structures that the agent can engage with from its situated perspective. In fact, this primacy of know-how, of the embodied on-line ability to deal with context dependent situations, is now a common theme for contemporary embodied and situated cognitive science, from developmental psychology to robotics (see Cliff 1991, Clark 1997, Pfeifer and Scheier 1999, see Wheeler 2005 and Chemero 2008 for a recent philosophical review).

Yet this approach did in fact came to surprise or unbalance the foundations of cognitivist (representational-state-functionalist) approaches to mind. Cognitivism is based on a clear demarcation between subject and object and characterizes cognition as the manipulation of symbolic (or even subsymbolic connectionist) representations of the external world inside the agent, often under the form of a context-independent general purpose reasoning. Much of the effort of dynamicist and situated cognitive science was, and still is, put on convincing or confronting strong cognitivist positions while providing examples of how a dynamical approach cutting across brain, body and world is better suited to explain, model and predict how cognitive performance is achieved in real (i.e. natural) intelligent systems and not in chess-playing or database management computational systems. The embodied and situated approach defends the primacy of a closed sensorimotor loop against the primacy of abstract reasoning. We need not postulate abstract representational units, processed according to rules of reason, in order to model cognitive systems². We can model animal behaviour and build robots that are capable of producing adaptive or cognitive-like behaviour without using internal representations. As roboticist Rodney Brooks concluded: “When we examine very simple level intelligence we find that explicit representations and models of the world simply get in the way. It turns out to be better to use the world as its own model” (Brooks 1991:140).

Whereas some of the problems of designing and understanding cognitive systems came to be solved by bringing them down to the world (to the direct sensorimotor situatedness of action) some problems were also transferred to it. As we saw in chapter 3 dynamicism, and embodied approaches for the same sake, does not solve the problem of demarcation of cognition on its own; by the mere situatedness of action. Conceptualizing cognitive systems as dynamically engaged on continuous sensorimotor interaction does not provide *ipso facto* a solution to how that engagement may become meaningful

2 This is certainly not the place to develop a full critique to the computational-functionalist approach to cognitive science. This critique has been done elsewhere (Varela, Thompson and Roch 1991, Brooks 1991, Bickhard & Terveen 1995, Clark 1997, Ryle 1949, Wheeler 2005, Chemero 2008, Keijzer 2001, etc.). My goal is just to put this discussion on its broader context. The problem of intentionality and meaning is not a specific problem of behaviour based robotics or dynamical cognitive science but a problem of cognitive science and philosophy of mind more generally. A problem that remains with us independently of the substrate (behavioural or symbolic, computational or dynamic) that different theoretical approaches take to support cognitive processes.

or intentional for the agent as distinctively cognitive; distinct, for instance, from the coupling between a Watt Governor and a water flow.

Going back to von Uexküll, he considered meaning to play a crucial role on the constitution of an Umwelt: “each Umwelt forms a closed unit in itself, which is governed, in all its parts, by the meaning it has for the subject” (von Uexküll 1982: 30). If the parts of an Umwelt are “governed by meaning” that implies that there is a causal role played by meaning and that cognition cannot be just reduced to a mere species-specific circularity between perception and action. However, the language of meaning, that the Umwelt is embedded with in von Uexküll’s account of behaviour, seems difficult (if not impossible) to naturalize by sole reference to sensorimotor cycles.³

But this problem of meaning is not new, it was certainly not alien to the cognitivist functionalist-representationalist paradigm where it was formulated by John Searle (1980) and later by Stephen Harnad (1990) as the problem of how may a symbol or a representational unit acquire any meaning or semantics if the rules governing their manipulation are purely computational or syntactic (i.e. causally independent from the specific content of such representations). Despite the fact that artificial computational systems could mimic (and even outperform) human rational and linguistically structured thinking, a problem remained over how could symbolic computational states acquire or be endowed with intentional or semantic properties that referenced to external states of affairs. John Searle’s Chinese room argument (Searle 1980) became a keystone of a critique to the computational conception of mind. A human subject is isolated into a room, so Searle’s thought experiment goes, following instructions to manipulate incoming Chinese symbols to deliver different symbols to an output window. Despite the fact that, seen from outside, “the room” may be showing linguistic or cognitive competence (having an “intelligent” conversation about psychoanalysis, diagnosing a brain tumour or winning a chess game) nothing like an “understanding” of the meaning of the symbols being manipulated inside the room is taking place. If understanding and meaning something (not just performing as-if) is the hallmark of the mind, Searle concludes, syntactic or computational manipulation of symbols does not suffice for mindfulness.

Dynamicist situatedness in turn puts at centre stage the entanglement of meaning with action and not with representational units (cf. Chinese symbols). Thus, the problem is transformed and liberated from the tyranny of symbols as units of cognitive organization whose relationship with external events is to be established and endowed with a semantic content (as if cogni-

³ When any explicit reference is provided to account for its meaningfulness in terms of how or why should that inclusion be termed “meaningful”, von Uexküll’s discourse excludes itself from any explanatory value through recurrence to admittedly metaphysical and unobservable “ego-qualities”, “non-physical energies”, “all-embracing master plans” or the metaphor of counterpoint of melodies of unknown origin and compositional source ultimately played by an “invisible hand”.

tion were to be a process of manipulation of boxes containing a message to be read and understood by an homunculus inside the agent). But if one were to reframe Searle's Chinese room argument for the situated, behaviour-based and dynamicist approach, the problem seems to have advanced little from its previous formulation. A person pulling and pushing robs and springs, following the numerical calculations of a set of differential equations, making a wheeled room move around an environment, does not seem to be more genuinely cognitive than its symbolic counterpart manipulating Chinese token according to syntactic rules.

2.2. Meaningful problems: behaviour without frustration

Thus, the embodiment and situatedness of cognitive agents does not seem at first sight to solve the problem of meaning but just to reformulate it: What is the relationship between a continuous sensorimotor engagement and meaning? Why should something (a situation, an object, the activity of a sensor) be meaningful at all just by the fact that it be integrated on a sensorimotor loop? A first reasonable step to make is to consider that not all sensorimotor loop is meaningful or intentional on its own, by a mere act of coupling, but only certain forms of engagement, by virtue of a certain characteristic mode of behaviour. For instance, it is common to rule out cases of reactive behaviour in which, for any given stimulus the same response is always determinately fixed, no matter what preceded the stimulus, what the environmental context is, what the internal state and history of the agent be⁴. Thus, if not any kind of behaviour is to qualify as cognitive or intentional, a *special kind* of coupling may still be able to do the job. Not a mere reaction nor necessarily a highly abstract behaviour mediated by inner thoughts or human symbolic capacities, but probably something in between.

It is precisely along the early rise of differential equation modelling carried out to understand biological and cognitive systems that such an attempt was made by pioneer cyberneticians Rosenblueth, Wiener, and Bigelow (1943)⁵.

4 The case of reactive behaviour cannot that easily ruled out as trivial and uninteresting; particularly after embodied and situated approaches to cognition. Even if the sensorimotor connection is fixed on the side of the agent its relative spatial location, its body morphology, momentum etc. may provide opportunities for considerably complex behaviour to emerge. Yet we are just dealing with a first approximation at this point and just describing what counts as a widespread, although not explicitly justified, assumption for many.

5 That we move back to 1940's, back to the pre-AI era, is no coincidence. When the functionalist program came to substitute the behaviourist program in psychology, behaviourism was radically dismissed and put aside in mainstream cognitive science. The situated, embodied and dynamical approach to cognitive science has revitalized old problems once the reference to internal representations (the functionalist hammer against behaviourism) is put into question. The reference to pre-functionalist accounts of cognition seems natural in this context and some of the old behaviourist attempts to explain mindful properties need to be re-evaluated. See Keijzer (2005) for an insightful reappraisal of the relationship between behaviourism and situated-embodied-dynamical cognitive

The aim of Rosenblueth and colleagues was to reduce or translate the terms purpose and teleology (and the mental vocabulary altogether) to the language of cybernetics and its mathematization of behaviour. (On what follows I would preserve the use of the terms purpose and teleology of their time where they could equally have argued today in terms of intentionality and meaning.)

In their seminal paper, Rosenblueth and colleagues defined behaviour as “any change of an entity with respect to its surroundings”. Purposeful behaviour was taken to be a subdivision of active behaviour, where “the term purposeful is meant to denote that the act or behaviour may be interpreted as directed to the attainment of a goal—i.e. to a final condition in which the behaving object reaches a definite correlation in time or in space with respect to another object or event” (p. 18). Teleology, in turn, is defined as purposeful active behaviour controlled by negative feedback. The term negative feedback is used to name the behaviour of an object A in relation to a condition or reference state B (named goal), so that the behaviour of A is governed by the minimization of the error in relation to B. The paradigmatic example is that of a missile-tracking device that adjusts its trajectory so as to correct the deviations from its moving target. The purposeful (goal oriented) and teleological (error correcting) behaviour can be described and accurately characterized in purely behaviourist terms; i.e. without reference to the mechanisms and structure that gives rise to it. Rosenblueth *et al.*’s proposal is that intentional behaviour can be taken to be of this kind. Think for instance on your desire to grab and drink a glass of water standing on the table. Without appealing to internal ghost-like entities such as desires or mental images of the goal, the behaviourist argues, it seems reasonable to think that one could define intentional behaviour by just looking at the trajectory of your arm: if you really intend to catch the fly or the can of beer then you will (through visual monitoring) direct your arm towards the can, correcting deviations from it (e.g. if I move the glass of water somewhere else), until you get it. Just like the missile-tracking device, like any system whose behaviour is governed by negative-feedback.

Yet, this approach, despite the initial enthusiasm that the notion of feedback generated on the understanding of living and interactive systems, was far from providing a definitive answer to the problem of meaning or intentionality. In *The Phenomenon of Life* (Jonas 1966/2001) the phenomenologist Hans Jonas strongly criticized the cybernetician, purely behaviouristic, approach⁶. He argued by *reductio ad absurdum*: if something is purposeful, as

science.

6 The behaviourist approach to cognition received a strong attack from different disciplines. But for a long time only the functionalist critique made by some analytic philosopher’s of the mind was considered important, due to the research program (based on the computer metaphor and functional representationalism) that such critique was able to develop for Artificial Intelligence and Cognitive Science. However, a number of phenomenologists (cf. Mearleau-Ponty and Jonas among many others)

far as it can be interpreted as directed to the attainment of a final condition of “definite correlation ... with respect to another object”, then, Jonas argues, all objects have the purpose of “running down” or decreasing (minimizing) entropy, since the only objectively sustainable final state of a machine is its disintegration. This leads to the unavoidable and uncomfortable conclusion that disintegration is the main and ultimate purpose of all machines and organisms. Richard Taylor (1950a) also argue in a similar vein using the example of a clock, whose definite correlation with a final state is, sooner or latter, to be found on stopping to operate due to any sort of failure.

One may, however, argue that the term “purpose” to denote a “definite correlation” is too far a generous excess in the early cyberneticists’ terminology. What remains important, however, is the notion of negative feed-back. In this sense it is clear that organisms of any kind do not maintain a negative feed-back interactive control towards disintegration but, on the contrary, it is precisely in the direction of their survival that negative feed-back control is executed. It is, however, the externalist nature of their approach that Jonas was arguing against. The root of this failure to account for purpose and teleology is the fact that behaviour (and thus purpose, teleology and, we could add, intentionality and cognition) is defined exclusively in terms of external relationships: “change of an entity with respect to its surroundings (...) by the behavioristic method of study that omits the specific structure and intrinsic organization of the object [or subject]” (Rosenblueth *et al.* 1943:18). According to Jonas, a behavioural account of teleology (or intentionality) cannot capture the intrinsic properties that define it⁷. Unlike the early cybernetic machines, genuine purposeful entities (organisms) have intrinsic, rather than extrinsic, ends. This intrinsic character of intentionality shows up in the *awareness of failure* that natural purpose is embedded with.

To further develop this idea Jonas’ considers the case of analogy between a cerebellar patient and a missile tracking device introduced by Rosenblueth and colleagues on their original paper. The servomechanical system, being inadequately damped, “misses its goal” over-correcting the shoots with higher and higher oscillations. The cerebellar patient who, asked to carry a glass of

had something to say against behaviourism without appealing to internal representations as vehicles of intentionality. It is their critique that is revitalized today, after the foundations of functional representationalism has been put into question by situated, embodied and dynamical approaches to cognitive science.

- 7 Richard Taylor (1950b) also pointed to a similar argument: “Now I submit that, from observable behavior alone, one cannot certainly determine what the purpose of the behaving object is, nor indeed, whether it is purposeful at all. Surely the observable behavior of the car and its driver might be exactly the same, whether the purpose is, as supposed, to overrun a pedestrian, or merely, as a joke, to frighten him, or, indeed, to rid the car of a bee, the driver being in this case wholly unaware that his car is endangering another person. If, however, purpose were definable solely in terms of observable behavior, as these writers suppose, then any driver who appeared to behave as if he were trying to run down a pedestrian, but who yet pleaded that he had no such intention, would not simply be probably lying, but could not possibly be telling the truth.” (p.328).

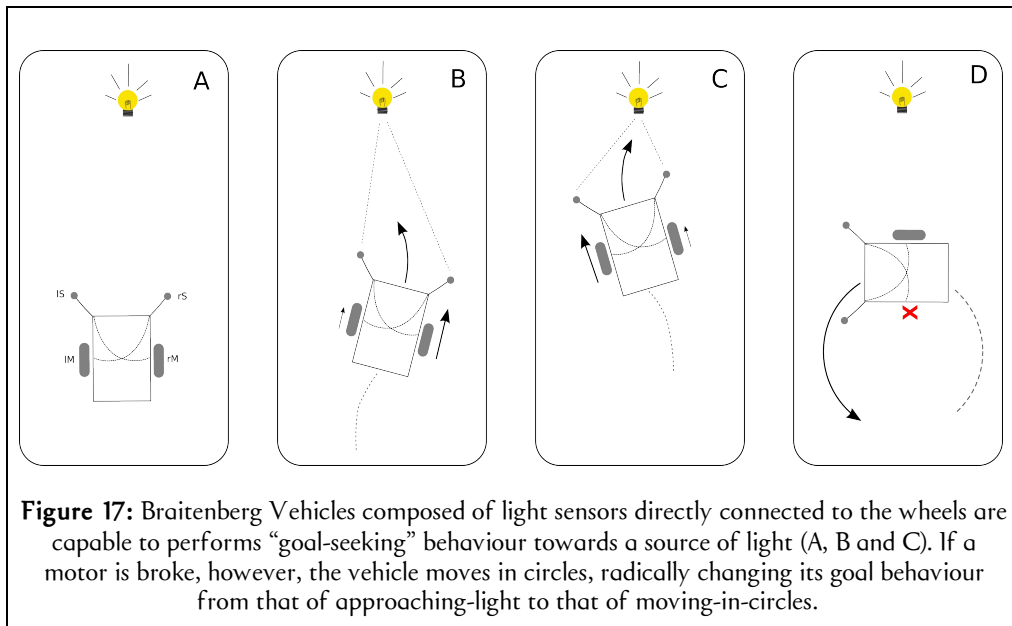
water from the table to his mouth, generates an analogue over-oscillatory pattern spilling the water, and it was said to instance of the similar failure as the negative-feedback system. The analogy of the dysfunction is meant to support an analogy of the mechanisms involved. There seems to be no objective difference between the servomechanical and the organic intentionality. To which Jonas' responds:

It is contended that the two cases are “strikingly similar”. However that may be, one point at least should be clear from the outset: the patient himself *wills* to bring the glass to his mouth, that is, he wants it there. This *end*, motivating the action from the start, is intrinsic in all the part-motions, providing the reference by which they are in themselves failures and make the whole undertaking a failure. Presumably the patient finds his inability to perform distressing. But the machine, for all we know, may just as well be said, instead of being distressed, to abandon itself with relish to its wild oscillations, and instead of suffering the frustration of failure, to enjoy the unchecked fulfilment of its impulses. “Just as well” amounts of course to “neither”. Manifestly neither “distress” nor “enjoyment” fits the *modus operandi* of a machine—not even as an operational analogy, since the machine is equally “satisfied” in each and any single step of behavior as it occurs, the occurrence as such being its own sole and sufficient vindication. In the case of the machine “missing a goal” means, of course, missing our goal, the goal for which it has been designed, namely by us, it “having” none itself; whereas the unfortunate patient truly misses its goal, which is his not because he has been designed for it but because he has formed and entertained the design. (Jonas 1966/2001: 112, italics on the original).

Whereas the patient (probably with a Parkinson's disease) finds its inability distressing and leading to frustration, the machine, designed in terms of the negative-feedback that was taken to be necessary and sufficient for teleology (intentionality-in-action) does not show any “sign of preoccupation”, no possibility for frustration. The problem of a behaviourist account of intentional agency is clearly shown in relation to human built devices, specially robots, that come to satisfy the situated and embodied sensorimotor condition and yet seem to fail on engaging in intentional behaviour.

Consider the additional example of a Braitenberg vehicle shown in figure 17. The name “Braitenberg vehicle” is used to refer to a class of robotic systems first proposed by Valentino Braitenberg (1984)⁸. One of the most simple forms of his vehicles consists of a robot with two wheels and two light sensors (see figure 17-A). The light sensor that is situated on the left side of the robot's front (lS) is directly connected to the right motor (rM) and the light sensor on the front-right (rS) connected to the left motor (lM). Activity on the light sensor induces a proportional activity on the motor. The simple (figure 17-A) case is straight-forward: if the robot is positioned facing directly a source of light it will move straight line towards it. However, this is a very unlikely case since the robot will, sooner or later, deviate from the straight line

⁸ Although pioneering work was already advanced by Grey Walter's experiments with *machina speculatrix* (Walter 1950—see Holland 2003 for a review).



(due to differences in motor friction, the inclination of the surface, etc.). Consider now the case of the vehicle standing slightly inclined to the right in relation to the straight line towards the light source (figure 17-B). In this position the robot receives a higher light intensity on its left side than what it receives on its right side. Accordingly its right motor (to which the left sensor is connected) will receive a higher activation than that of the left motor. An overall forward movement is produced while slightly turning to the left. As it turns to the left, however, the robot’s right sensor is now closer to the light (figure 17-C), this produces a “correcting” response since the right sensor will more strongly activate the left motor now. The net behaviour is that of a negative feed-back movement towards the light, its behaviour corrects deviations from the straight line while moving towards the light (its “goal”, in the cyberneticists’ jargon).

A simple Braitenberg vehicle satisfies both the closed sensorimotor loop condition and the cybernetic negative feed-back condition. Does it suffice for intentionality? Can we justifiably claim that it intends to approach the light? According to Jonas we cannot, intentionality cannot be pictured exclusively in an externalist behavioural descriptive fashion. The failure is best seen when considering the effect (or lack of effect) when the machine fails to achieve its “pretended goal” by some mechanical disruption. For instance if the left motor breaks (figure 17-D) the vehicle will turn around counter-clockwise rather than moving forward towards the light, endlessly rotating without achieving “its goal”.

To be precise the situation is even worst for the cybernetician approach; simply put, the robot cannot even fail! On the example of the Braitenberg vehicle endlessly moving in circles, according to the cybernetic approach, the light cannot be said any more to be the goal of the vehicle, since the notion of goal was defined exclusively in terms of a behavioural description. Failure of a motor has suddenly transformed the goal of “approaching the light” into that of “rotating counter-clockwise under the presence of any light source”. This qualifies as a new “goal” since the robot will maintain it stable, correcting deviations we may impinge upon it if, for instance, we push it forward. A completely different regime seems to take place in organic behaviour. If a dog is starving, breaking its leg while approaching a food source is certainly not going to make it change its goal automatically from that of approaching the food to that of moving in circles. The animal will adapt to the new condition and crawl towards the food, the intention being the source of its reorganization of action and frustration being the result of repeated failures.

Thus, although the domain of sensorimotor cycles seems the right place to pose the problem (in opposition to a higher level symbolic domain of language-like capacities) there seems to be little, if any, entailment of meaningfulness on the mere externalist description of sensorimotor closure. A missile tracking device or a wheeled and double wired phototactic robot (a Braitenberg vehicle) do satisfy the sensorimotor coupling condition. They have their own Umwelt (if defined purely in terms of sensorimotor cycles) and they can even achieve some types of apparently intentional behaviour. But they fail to count as intentional. Obviously von Uexküll was surely not ready to admit such devices as meaningfully engaged on their own Umwelt; yet he failed to provide a fruitful explanation of how precisely did meaning arise from organic behaviour. von Uexküll was, however, in line with Jonas, pointing towards some type of internal property of organisms that could endow perception and action with a meaningful, intentional character; with a *norm* capable of providing an intrinsic sense of failure and intentional readjustment.

3. BIOLOGICAL GROUNDING OF THE SENSORIMOTOR LOOP

Jonas’ requirement that intentionality be originated on internal rather than external grounds is what biological embodiment may come to satisfy. Our account of adaptive agency (in the preceding chapters) was precisely formulated in this direction. Interactions become functional insofar as they are integrated on the self-maintaining organization of the biological autonomous individuality that supports them. The norm is defined from within, as a boundary from which to remain safe: the viability boundaries for self-main-

tenance and preservation that the metabolic far-from-equilibrium organization defines.

The interactions in which robots (as we know them) are engaged in cannot be said to respond to an intrinsic normative functionality because the consequences of their actions have no effect on their organization. It is the designer that specifies the goal of their sensorimotor coupling, which remains contingent to their being. Whereas organisms' sensorimotor coupling with their environment becomes crucial to ensure their ongoing existence. When we see a dog approaching a piece of food we may judge that it is concerned by its goal, that the result of its behaviour bears significance for it, for its survival, and that it really has the intention to approach the food. The problem now is to precisely derive or ground the intentionality of its action on its biological organization and to clarify how to assess if this biological grounding of intentionality is adequate or possible at all.

3.1. Living as a process is a process of cognition

One of the most cited formulations of the $C=AB$ hypothesis goes back to Maturana and Varela's theory of autopoiesis: "The domain of all the interactions in which an autopoietic system can enter without loss of identity is its cognitive domain" (Maturana & Varela 1980: 119). A more explicit formulation goes back to Maturana's previous work (1970):

A cognitive system is a system whose organization defines a domain of interactions in which it can act with relevance to the maintenance of itself, and the process of cognition is the actual (inductive) acting or behaving in this domain. *Living systems are cognitive systems, and living as a process is a process of cognition.* This statement is valid for all organisms, with and without a nervous system. (...) *The nervous system expands the cognitive domain of the living system by making possible interactions with 'pure relations'; it does not create cognition.* (Maturana 1970: 13, italics in the original).

As I quickly mentioned in previous chapters Ashby proposed that adaptive behaviour should be modelled as that of a system capable of maintaining homeostatic (i.e. stable under perturbations) a set of "essential" variables within viability constraints (understood as boundaries). Maturana and Varela's original contribution lies on the conceptual formulation of systems that what maintain homeostatic is precisely their own organization: autopoietic machines (see figure) or what I have more explicitly termed as basic autonomous systems (based on Ruiz-Mirazo & Moreno 2004). Thus, it is not mere sensorimotor coupling what occurs and characterizes living behaviour but that the mechanisms generating such behaviour (e.g. the NS) be "dynamically subordinated to the autopoiesis of the organism which it integrates" (Maturana & Varela 1980: 127). As Randall Beer puts it clearly: "An animal's behavior is adaptive only so long as it succeeds in maintaining its trajectory

within this viability constraint, that is, only so long as it succeeds on maintaining the conditions necessary for its continued existence". (Beer 1997: 265). We can now provide a more explicit formulation of a basic or plain C=AB hypothesis in terms of adaptive agency as follows:

Cognition = *the sensorimotor coupling that an autonomous agent establishes with its environment in order to (or as long as it manages to) maintain itself under viability boundaries*

Many other authors have explicitly endorsed this notion of adaptive behaviour in order to characterize or ground cognition or intelligence. For instance Luc Steels writes: "The behavior of a system is intelligent to the extent that it maximizes the chances for self-preservation of that system in a particular environment." (Steels 1995: 78). Or Bourguine & Stewart affirm that "A system is cognitive if and only if sensory inputs serve to trigger actions in a specific way, so as to satisfy a viability constraint" where the viability constraint is derived from the system's conditions to maintain its autopoiesis (Bourguine & Stewart 2004:327). Similarly, the observation that it is the living embodiment of organisms what provides their behaviour with and intrinsic normativity has pushed other authors (Ziemke 2001, Moreno and Etxeberria 2005, Haselager & Gonzalez 2005) to state that robots will ultimately fail to be genuinely cognitive unless they are intrinsically endowed with living and self-producing organic bodies.

Yet, despite an increasingly widespread endorsement of this position there is little explicit justification over why should self-producing or autopoietic capacities endow a sensorimotor agent with genuinely *intentionality*⁹. On the work of Maturana and Varela cited above cognition is just stated to be coextensive with living (i.e. biologically adaptive) behaviour. The thesis is declared and some of its consequences elaborated but not explicitly justified. It is implicitly assumed that adaptive behaviour (and more generally "living as a process") already contains all the necessary "ingredients" for anything that is generally attributed to the domain of the cognitive: a world of perception and action that appears, for the observer, as in congruence with the environmental conditions that are relevant for the organism's self-maintenance. In Maturana's words:

9 This is not to say that the question has not been treated in length, but its various treatments can be grouped into three main categories: a) the focus is on the distinction between robots (or human built artefacts in general) and organisms, b) the goal is to trace back the origins of cognition (and, arguably, the engineering principles of AI and robotic) to the lower level behaviour of animals (in contrast with higher-level symbolic architectures for behaviour engineering) and c) a description of living organization is provided, its contrast and consequences regarding other approaches to cognitive science highlighted and the full approach declared as an alternative approach. However, in all three categories of contributions, little effort is dedicated to the demarcation problem and particularly to justify why and precisely *how* does cognition arise exactly, making use of explicit mechanistic details that could disclose the minimal (necessary and sufficient) conditions of cognitive behaviour in a natural model organism or in a complete (autopoietic and behaving) simulation or mathematical model.

[I]f we see a living system behaving according to what we consider is adequate behavior in the circumstances in which we observe it, we claim that it knows. What we see in such circumstances, is: a) that the living system under our attention shows or exhibits a structural dynamics that flows in congruence with the structural dynamics of the medium in which we see it, and b) that it is through that dynamic structural congruence that the living system conserves its living (Maturana 2002)

Yet it is still to be proved, or convincingly argued, why should mere *congruence with conservation* of life be quantifiable as knowledge, why should interactive contribution to self-maintenance be sufficient for cognition or, ultimately, how could biologically grounded normativity (the norm of self-maintenance) bring *intentionality* into sensorimotor interactions. (As we will see it is this last requirement, that the proposed model of adaptive behaviour suffices for *intentional* agency, what better meets the demands for mindfulness established by traditional and contemporary philosophy of mind, in line with Jonas' critique and analysis of frustration, the counterpart of intentional action.)

Despite Maturana and Varela's original endorsement of a purely mechanistic and operational explanatory framework (and subsequent rejection of functional or teleological vocabulary¹⁰), Varela's latter work (Varela 1992, Varela 1997 and, particularly, Weber & Varela 2002) explicitly addressed the issue of teleology, intentionality and sense-making in an attempt to ground the mark of the mental (the most characteristic features of mind and cognition) in the realm of living organization. I shall follow their argumentative line, and some latter interpretations (Di Paolo 2005, Thompson 2004, 2007), as exemplifying one of the most elaborate developments of Maturana and Varela's original formulation of the hypothesis in relation to the issue of intentionality.

3.2. Making it explicit: adaptive mechanisms need find their place

Weber and Varela's analysis of teleology in organisms departs from Kant's *Critique of Judgement* where he acknowledges the impossibility of understanding (self)organized systems (i.e. living systems) by purely Newtonian-mechanical principles due to the circular entanglement between causes and effects in organisms, whereby parts are causes of other parts but also effects of the others. To say it in other words, Kant acknowledges the need to approach organisms as circularly networked, so that parts of the organisms are both means and ends of one another and the organisms (unlike human made artefacts) maintains and generates itself as a unity: parts are products and producers of other parts. This feature of organisms (or, Kant would rather argue, the way they

10 Maturana and Varela state: "[T]he notions of purpose and function have no explanatory value in the phenomenological domain which they pretend to illuminate, because they do not refer to processes indeed operating in the generation of any of its phenomena." (Maturana & Varela 1980: 86)

appear accessible to our Judgement) endows them with *natural intrinsic purpose* since their parts, or the organism as a whole, are not means for an external agent (that makes use of them for its own ends) but ends on themselves and for themselves. Current scientific understanding of complex and self-organized systems and, particularly, the autopoietic theory of life, Weber and Varela claim, provide today a naturalized account of organisms' intrinsic teleology (something that Kant took only to be acceptable as a regulatory concept valid to guide human understanding but not that clearly constitutive of natural systems).

The scientific unravelling of the characteristic circularity of organisms makes possible to treat their teleology as constitutive (not merely as a regulatory concept necessary to make then understandable) and, I shall evaluate, to provide a naturalist account of their purposive or intentional behaviour. We have followed a very similar line when characterizing basic autonomous systems and explaining their intrinsic teleological nature in terms of normative functionality. But the notions of behavioural purpose, sense-making and meaning, the language of cognition, slips into Weber and Varela's account when jumping into the realm of system-environment interactions:

The key here is to realize that because there is an individuality that finds itself produced by itself it is ipso facto a locus of *sensation* and agency, a living *impulse* always already in relation with its world. There cannot be an individuality which is isolated and folded into itself. There can only be an individuality that copes, relates and couples with the surroundings, and inescapably provides its own world of *sense*. In other words by putting at the center the autonomy of even the minimal cellular organism we inescapably find an intrinsic teleology in two complementary modes. First, a basic *purpose* in the maintenance of its own identity, an affirmation of life. Second, directly emerging from the aspect of concern to affirm life, a *sense-creation purpose* whence *meaning* comes to its surrounding, introducing a difference between environment (the physical impacts it receives), and world (how that environment is evaluated from the point of view established by maintaining an identity). (...) The organic coupling and change must, according to its self-constitution, be always directed to maintain the process of self-realization. An autopoietic system is necessarily referred to itself: its actions consist in establishing the dynamical processes of staying alive. Stimuli from outside enter the sphere of relevance of such a unit only by their *existential meaning* for the keeping of the process of self-establishment. They acquire a *valence* which is dual at its basis: attraction or rejection, approach or escape. (Weber & Varela 2002: 117, italics added.)

In a similar passage Varela claims:

The difference between environment and world is the *surplus of signification* which haunts the understanding of the living and of cognition, and which is at the root of how a self becomes one. In other words, this surplus is the mother of *intentionality*. (...) There is no food *significance* in sucrose except when a bacteria swims upgradient and its metabolism uses the molecule in a way that allows its identity to continue. (...) In brief, this permanent, relentless action on what is lacking becomes, from the observer side, the ongoing *cognitive* activity of the system, which is the basis for the incommensurable difference between the

environment within which the system is observed, and the world within which the system operates. (Varela 1992: 7—8, italics added.)

Italics in the above quotes are meant to highlight how a mentalistic language permeates the interpretation of living behaviour. It is difficult, however, to disclose how this jump from intrinsic teleology (the normative functionality of living organization) to a “meaningful” and “sense-creating purpose” is carried out by Weber and Varela. Two main attempts have been made to render this issue more explicit, to show some implicit assumptions and disclose the argumentative line that justifies the equation $C=AB$.

Evan Thompson’s (2004) reconstruction goes as follows:

1. Life = Autopoiesis
2. Autopoiesis entails the emergence of a bodily self
3. Emergence of a self entails emergence of a world
4. Emergence of self and world = sense-making
5. Sense making = cognition (perception/action)

In this sense the equation $C=AB$ seems to have collapsed into a more basic one. From Thompson’s analysis, if 1 to 5 holds, it follows that “Life = Cognition”. Cognition, to use Weber and Varela’s phrase, “*directly* emerges from the concern to affirm life”.

The second attempt to clarify the issue, is due to Ezequiel Di Paolo’s unravelling of the implicit entailments that tie the theory of autopoiesis with its grounding of cognitive processes (Di Paolo 2005: 434):

- A. Self-construction is a process that defines a unity and a norm: to keep the unity going and distinct,
- B. Encounters with the external world are “evaluated” by the system (through the self-constructing machinery) as contributing or not to the maintenance of autopoiesis, consequently
- C. Autopoiesis implies sense-making, an intrinsic perspective of value on the world.

“The crucial point here” Di Paolo follows “is B. and whether it can be derived from the original formulation of the theory [of autopoiesis]” (p.434). Our reconstruction of the morphophylogeny of agency has provided enough arguments to show that metabolic self-maintenance or autopoiesis does not necessarily entail B and that, following Di Paolo, additional mechanisms of regulation of interactions are required to achieve what I have termed “adaptive agency”. Equally, it is the entailment of steps 3 and 4 in Thompson’s argumentative line that appears incomplete: the emergence of a self does not entail the emergence of a world and the co-definition of self and non-self does not imply sense-making. The mechanisms to achieve adaptive agency are

missing from the autopoietic framework account and they crucially shape the feasibility of the whole hypothesis.

Consider the case of a cell that stands close to a source of radiation for a few hours. The bio-molecular organization of the system determines that radiations are capable of destroying it, thus making radiations normatively “dys-functional” for its continued existence. But from this mere fact we can derive no necessary sense-making or intentional relationship towards the source of radiation. The cell (and any organism for the sake of the argument) would remain completely unaware of its progressive disintegration *in relation to the source of radiation*. The radiation, despite bearing a devastating existential consequence for “the keeping of the process of self-establishment” cannot be said to acquire “any valence of attraction or rejection, approach or escape” unless, *at least*, mechanisms exist that are capable of detecting the source of radiation and make the system act accordingly. Our careful analysis of the set of organizational steps required to achieve agency has made it clear that the emergence of a metabolic self only creates a set of boundary conditions that do not qualify as environment (not to speak of the term “world”) until the mechanisms are in place able to engage on an interactive cycle that expands its boundary conditions to create an environment. In Di Paolo’s words¹¹:

We find that sense-making in organisms needs both autopoiesis and adaptivity. Autopoiesis provides a self-distinct physical system that can be the centre of a perspective on the world, and a self-maintained, precarious network of processes that generates an either-or normative condition. Adaptivity allows the system to appreciate its encounters with respect to this condition, its own death, in a graded and relational manner while it is still alive. (Di Paolo 2005: 439)

It seems that we have moved in circles from “Cognition = Adaptive Behaviour” to “Cognition = Life” and back to “Cognition = Adaptive Behaviour”. What we have gained in this round about is the rooting of intentional action into the teleology (normative functionality) of metabolic or autopoietic organization expressed through their sensorimotor coupling. Thus, sense-making (as the most elementary form of intentionality) may be understood, according to the C=AB hypothesis, not as a mere encounter of environmental factors with an autopoietic (living) organization nor as a mere sensorimotor coupling, but as some kind of *projection* of the inherently teleological nature of living organization into adaptively regulated environmental interactions. It is the connection between behaviour and the norms dictated by its biological embodiment that seems able to account, on a yet not fully examined way, for the intentionality of adaptive behaviour

A number of naturalist philosophers have developed an account of cognition in a very similar fashion, inspired-on and expanding-from Maturana and

¹¹ As we shall see latter Di Paolo’s position is still more elaborated that what the following quote might seem to imply. I will proceed by integrating additional levels of complexity step by step analysing the requirements one by one.

Varela's original theory of autopoiesis and autonomy though a theory of interactive adaptivity. Christensen and Hooker (Christensen 1999, Christensen & Hooker 2000), for instance, consider that cognition arises as an expansion and further complexification of the adaptive capacities of an autonomous system in four complementary dimensions: a) anticipatory capacity, b) reduction of local-context dependency, c) modulation of interactions over larger time-scales and d) construction of anticipative models and explicit goals. Mark Bickhard (2004), on the other hand, relies on a non-encodingist, interactive and forward-oriented concept of representation to characterise cognitive capacities. His concept of representation and representational normativity is grounded on the adaptive autonomy of the system (on its recursive self-maintenance, to use his terminology) so that representational content can be evaluated and updated according to the consequences of the interaction on the self-maintenance of the system. What biological autonomy provides for these authors is a conceptual ground to explain the internal emergence of norms that successive adaptive mechanisms should satisfy, thus naturalizing the concept of norm and intentionality in biological organization.

This account of adaptive behaviour seems now to be safe from Jonas' critique to the externalist cybernetic conception of purpose and teleology: the norms or goals to which negative feed-back (and additionally sophisticated) adaptive mechanisms follow are immanent to the agent. In living agents Ashby's viability boundaries of viability, that define the adaptive norm of the agent, are not capricious limits imposed or defined extrinsically by some external agent but generated from within as a result of the far-from-equilibrium conditions of their organic infrastructure.

Following the terminology that I have developed through our morpho-phylogenetic reconstruction of agency, the basic autonomous metabolic closure (autopoiesis) provides an identity and norm for regulation, a set of constitutive boundary conditions and an inside-outside dichotomy. Adaptive mechanisms, on the other hand, operate regulating interactive processes according to this norm and, the proponents of the C=AB thesis will endorse, the self-organized source of this norms provides a teleology to be projected into environmental interactions thus making reasonable the use of a cognitive or intentional terminology in the description (sense-making, value, meaning, purpose, ...).

Note, however, that the nature of this *projection* remains still to be fully specified. We have an adequate biological grounding for the norm and we have a set of adaptive mechanism regulating sensorimotor interactions in accordance or correlated with this norm. But why and how can this correspondence or correlation be sufficient to *project* intentionality or sense-making into the environment? Is the correspondence between the norm (dictated by the

self-maintaining requirements of metabolism) and adaptive mechanisms sufficient to account for intentionality as a form of projective teleology?

4. PHENOMENOLOGY MEETS THE MECHANISMS OF ADAPTIVE BEHAVIOUR

Abstract and complex as it is and full of highly inspiring and radical consequences, the autonomous nature of living systems elicits an empathic projection of our own inward experience onto their (our) struggle for self-affirmation and the interactions that living agency (including ours) establishes to do so. However, avoiding the temptation of a jeopardizing over-projection should be a major philosophical concern. It remains a central claim for this project that “autopoiesis [and, together with it, the theory of biological organization rooted on the notion of autonomy] is a singularity among self-organizing concepts in that it is on the one hand close to strictly empirical grounds, yet provides the decisive entry point into the origin of individuality and identity, connecting it, through multiple mediation, with human lived body and experience, into the phenomenological realm.” (Weber & Varela 2002: 116). But if we are to connect empirical theories of living organization with the phenomenological realm, all the burden of the connection rests on the details of the “multiple mediation”. And this is not always sufficiently addressed. Unfortunately, claims of minimal cases of adaptive behaviour as genuinely intentional (in Varela 1992, Weber and Varela 2001, Thompson 2004, 2007 or Di Paolo 2005) lack any detailed reference to specific mechanism and how they are suppose to generate intentional agency. A further development of the $C=AB$ hypothesis would require to make as explicit as possible the connection between the adaptive mechanisms of biological organization and the phenomenological experience of intentionality.

But before doing so we need first spend some time explaining some methodological aspects of this connection and, particularly, to carefully design the way in which phenomenological insights can be put at work for this task. In fact, at this point, we cannot elude for longer what constitutes a benchmark test for a mechanistic-organizational model for minds and this is where phenomenology may helpfully enter into the scene.

A common strategy in philosophy of mind is to appeal to some kind of phenomenological test, to compare our inward experience of “what is it like to be a mind” with the properties and capacities that a proposed mechanisms is able to display. Different types of strategies have been used but I will focus on two. The first is based on abstracting a structure, a template, of mind processes (intentions, thoughts, etc.) and then testing whether a proposed mechanisms matches that template or is capable of generating the properties or

structure of such template. Fodor and Pylyshyn's strategy (1988) provides a fruitful example: they abstract certain structured properties of thought (compositionality, systematicity etc.) and tests whether certain proposed mechanisms (e.g. connectionist architectures) match this structure¹². There is a second widespread use of phenomenology as a benchmark test to rule out negative cases. One can check if the *consequences* that can be derived from the analysis or synthesis of proposed mechanisms are consistent with our inward experience. Searle's Chinese room argument falls under this category together with Jonas' analysis of failure and frustration of intentional behaviour. In turn, both strategies are deemed to work together in the long run if they are to deliver a satisfactory account of the mind. The process of building a model often involves a circulation between a proposed model, its phenomenological validation or refutation and the inclusion of additional changes until the final structure is not only internally consistent but also consistent in relation with our experience. In turn, the process is often split in two: the *top-down* structure of the model is generally developed at a very abstract level, linguistic-conceptual and/or functional, and then used to guide or test the building of *bottom-up* mechanistic models.

We have made use of a similar strategy proposing a rather abstract notion of agency and building, from the bottom-up, a generative model of the organization able to match that conceptual structure. It is time, however, to move a step forward and make explicit, in more detail, the structure of intentional agency as the minimal expression of mind. For this task Jonas' reflections on the intrinsic origin of intentionality, and its counterpart of frustration, brings us the opportunity to specify in more detail the minimal structure of intentional agency as the benchmark test for evaluating the adequacy of the C=AB hypothesis. The notions of meaning and sense-making, in turn, will rest on that of intentional agency. (I shall latter make explicit how the second strategy will be used to refute or reconfigure proposed mechanisms.)

4.1. Towards a minimal structure of intentional agency

I shall assume the shared agreement among philosophers of mind that intentionality is the most characteristic mark of the mental. This consensus is broken, however, when trying to come into a unified definition, the nature of

12 To the extent that some of the properties of thought can be externalized on linguistic structures, and even formalized, it may seem that the "structure-template" strategy involves no phenomenological commitment. This is due to the fact that the distilled formalized template might not require a necessary reference to our own inward experience to justify its consistency (i.e. it is formally well founded, rigorously defined or computationally tractable on its own). However, why and how a certain structure has been chosen (e.g. among the infinite number of well founded formal systems) needs always to be justified on our experience of reasoning, perceiving or acting according to or in analogy with that structure. There is always an unavoidable phenomenological dimension on the justification of such structure.

intentionality and how it should be understood and conceptualized to provide a benchmark test for cognitive models. The philosophical marketplace offers a huge variety of options and I will dedicate a few words to introduce and make explicit the rationale for our favoured choice.

The notion of intentionality that was inherited in mainstream cognitive science and philosophy of the mind is due to Brentano who, in turn, borrowed it from the scholastic tradition of medieval philosophy (Jacob 2003). Under this tradition the term “intentionality” refers to the *aboutness* or semantic content of thoughts or ideas as the characteristic feature of the mind. If you now close your eyes and think about freedom and democracy George W. Bush will surely come into your mind. Obviously it is not George entering your skull (that would be painful! —specially for Bush’s peculiar style of democratizing other-minded territories): it is rather the capacity of your thoughts to be about things out there, even kilometres away from where you stand, even if you cannot see them now, even if they don’t exist. Tones of ink and computer memory have been spent on explaining this feature of cognitive processes and the intricate relationship between thoughts and the object they refer to (semantics). But we need not get so far, appealing to abstract ideas and thoughts to start solving the question of intentionality, at least if our goal is just to provide a minimal model for minds as intentional agents. As Searle puts it:

If you ask, how is it possible that anything as ethereal and abstract as a thought process can reach out to the sun, to the moon, to Caesar, and to the Rubicon, it must seem like a very difficult problem. But if you pose the problem in a much simpler form, How can an animal be hungry or thirsty? How can an animal see anything or fear anything? then it seems much easier to fathom. We are speaking (...) of a certain set of biological capacities of the mind. And it is best to start with the biological capacities that are primitive—for instance, hunger, thirst, the sex drive, perception, and intentional action. (...) Once we demystify the problem of intentionality by removing it from the abstract, spiritual level down to the concrete level of real animal biology, I do not believe that any unsolvable mystery remains about how it is possible for animals to have intentional states. (Searle 2004:115).

The notion of intentionality that I shall advocate here, “down to the concrete level of real animal biology” will be that of intentional agency. So I will not deal with any kind of ideatic content or mental imagery. Rather (recall that the C=AB hypothesis is primarily about closed sensorimotor loops) I should use the term **intentionality** on the domain of behaviour and particularly biological behaviour generating mechanisms. This sense of intentionality is in fact closer to the original Latin verb *intendere* which means *to be directed towards* some result, goal, object or situation. Note, in addition, that the origin of the term is, in fact, a verb: i.e. a mode of action.

Searle’s account of intentionality (and that of most contemporary analytic philosopher’s of mind) relies on the concept of representation and propositional or conceptual content. Unfortunately the propositional or conceptual

structure that serves as a template, to be filled in with intentions, is an extremely abstract notion that depends (for its modelling and understanding) upon linguistic constructs of the sort “I believe that drinking the glass of water will satisfy my thirst” or “I desire to satisfy my thirst”. Atomized elements or items (e.g. glass, thirst, water, etc.) are tied together by syntactic or logical relationships and tied to the world, in Searle’s account, through the notion of “conditions of satisfaction”. In particular the notion of what Searle calls “intentionality-in-action” is tied with the world through an action: e.g. my intention “to grasp the glass of water” is satisfied by the action of grasping the glass of water. In Searle’s approach, for something to be an intentional agent its actions must be caused by a propositional representation of the action’s conditions of satisfaction. The difficulty to naturalize this account of intentionality is apparent if one tries to translate propositional terms into mechanistic ones. But the problem is not just that of translating or finding what might count as an instantiation of a propositional representation in a mechanistic model amenable to biological modelling. After all, we should be ready to face this difficulty if we are to complete the project of making explicit a model for minds (in fact, philosophical departments, journals and publishing companies alike have made a huge investment on providing potential solutions to mechanistic or naturalized accounts of representations). The problem, and the opportunity that is whereby opened, is whether representational or propositional structures are really required to account for intentionality. The question is whether mindful processes do really take the form of representational items tied by syntactic or logical relationship of a language-like nature or something else, and exactly what, can and does the job in everyday living and lived intentionality.

Hubert Dreyfus (1972, 1993, 2002), reclaiming the phenomenological tradition inaugurated by Heidegger (contra Husserl) and developed by Merleau-Ponty, has repeatedly argued that reference to representations is not necessary to account for intentionality-in-action. One needs not postulate the presence of an inner explicit representational image of a glass of water in order to have the intention to catch it. Rather, most of our intentional behaviour takes the form of an involved or absorbed coping; a fluid interaction dynamics whose intentional dimension comes forth when a breakdown or a failure occurs.

According to Merleau-Ponty, in everyday absorbed coping, there is no experience of my causing my body to move. Rather acting is experienced as a steady flow of skilful activity in response to one’s sense of the environment. Part of that experience is a sense that when one’s situation deviates from some optimal body-environment relationship, one’s motion takes one closer to that optimal form and thereby relieves the “tension” of the deviation. (Dreyfus 2002: 378)¹³.

13 The discussion over the requirement of conceptual content or a propositional structure for mindfulness is a hard philosophical issue. I am not pretending to settle it here.

This form of engagement with the world is typical of skill full behaviour, of the everyday know-how that Ryle took to be primary and characteristic of mind. In order to make explicit the structure of this process Dreyfus brings into play Merleau-Ponty's notions of *maximum grip* which "names the body's tendency to respond to these solicitations in such a way as to bring the current situation closer to the agent's sense of an optimal gestalt" (Dreyfus 2002: 367-8) or "directly sensing one's current experience as a deviation from a norm" (p. 376).

As a guiding example think again on yourself being thirsty and trying to grasp a glass of water standing in front of you. Both your thirst and the opportunity to satisfy it with the glass of water brings about a tension to grab the glass. To some extent, that you intend to catch the glass of water shows up on counterfactuals (what-would-happen-if). If the glass moves (e.g. if you are on a plane and the surface where the glass stands becomes inclined) you will change the trajectory of your arm accordingly. If your shoulder is in pain you will change the approach with an alternative movement of your arm. If the pain persists, or your arm does not respond, then you will most probably try to catch the glass with your other hand or ask someone else to do it for you. But these counterfactuals cannot be exhausted on a list that could constitute a full definition of intentional behaviour for each instance of it. This is why an externalist-behaviouristic account will not do for explaining intentionality (as I argued previously). The internalist-situated alternative highlights that all the counterfactuals stir from an inner (yet contextually situated) source of intentionality that will re-organize action to achieve the goal under different circumstances. It is the inner tension, that initiates and defines the conditions of satisfaction of your action, what characterizes it as intentional. And if you fail on grasping the glass of water (e.g. if the glass falls down to the floor and the water spills), if you fail to release that tension, you will find yourself on a characteristically mindful situation: you will feel frustrated.

The minimal structure of intention-in-action can thus be pictured out as that of an *inward tension* (an in-tension¹⁴) between initiation and satisfaction of an action (your perceiving the glass of water and your reaching it). The conditions of satisfaction of the interaction are given by a *norm* defined by the agent, and the tension is the result of the deviation from this norm. Yet, neither norm, deviation nor tension need to be cast in terms of itemized representational units with a propositional structure and, more importantly, they appear closer to our dynamicist modelling of adaptive behaviour, amenable to the epistemological constraints stated at the beginning of this work. A key aspect of this minimal structure of intentionality-in-action is that the tension is

14 Unfortunately the term "intensionality" has been charged with different connotations to that of intentionality in the philosopher's jargon. It would otherwise have been a beautiful term to label the (internal)-tension of intentional action.

constitutive of the very action in some inner, subject dependent, manner. This subject-dependency does not imply that the action be exclusively *triggered* by internal factors as if occurring in isolation from an environmental context. The environment may well solicit a response or provide an opportunity that, in turn, initiates or triggers the tension (e.g. the stewardess bringing the glass of water to your seat). However, *which* environmental situations solicit a response or become an opportunity (and opportunities for *what*) is, in any case, determined by the agent, by its background of abilities and needs, its embodiment and dynamics, by its organization and, particularly, by what this organization does determine as a satisfactory outcome of the interaction; thereby defining a tension and its conditions of satisfaction (i.e. compensation).

I would like to add a final proviso in order to keep the structure of intentional agency at its minimum and avoid potential misunderstandings on what follows. We certainly gain phenomenological insight into the intentional structure of our action through becoming aware or conscious of its occurrence (or by consciously recalling the experience of it). Yet I should avoid the temptation to include any form of conscious deliberative or self-monitoring awareness of behaviour to our minimal structure of intentionality. It is not the nature of consciousness what is at stake here but the nature of what we are conscious of (if we are conscious at all) when we are involved on intentional behaviour.

Thus, in what follows, intentions will be attributed to specific modes of action (or dispositions for acting) and in relation to norms that originate on the internal organization of the subject of such actions. How the norm is originated and how it relates to behaviour still remains to be explained, but this is precisely what I will investigate on what follows. Our sketch of a (rather explicit but still intuitive and pre-operational or pre-mechanistic) minimal structure of intentionality-in-action will suffice to depart.

Additionally we are now in the position to provide a set of additional distinctions that surround that of intentionality throughout the literature that advocates for the C=AB hypothesis. **Meaning** can be distinguished-from and be related-to intentionality in a relatively straight-forward manner. An object or event can be said to become *meaningful* for an agent insofar it is integrated in an intentional sensorimotor coupling, either triggering it, mediating it or otherwise. The stewardess approaching through the corridor becomes meaningful in relation to the opportunity for asking a glass of water. The glass of water is meaningful in relation to your intention to drink. The fact that the table gets suddenly inclined due to a turbulence on the flight becomes a meaningful or significant event in relation to your intention of grasping the glass and so on. In turn, **sense-making** may be used as the most elementary form of intention-in-action, in which no *directionality* of action (in the sense of a specific outcome) needs to be involved as such, but just a *sense* of attrac-

tion or rejection, two poles of valence that split an otherwise neutral space of possible courses of action into positive and negative (Weber & Varela 2001, Di Paolo 2005)¹⁵.

These provisional clarifications shall suffice to move on and evaluate how biological embodiment may provide sensorimotor interactions with an inner norm, from which a tension and conditions of satisfaction can emerge, able to make adaptive behaviour intentional and a minimal instance of cognition. But first, I would like to introduce the second mode of turning into to phenomenology for refuting or reconfiguring a model of intentional agency.

4.2. Phenomenological refutations

Phenomenological approaches are currently reclaimed as the ultimate ground of cognitive explanations, starting from the very origins of intentionality in living organization (Jonas 1966/2001, 1968, Weber & Varela 2002, Thompson 2007). Yet this phenomenological turn, necessary as it may be for an ultimate grounding of any possible explanation and also as a valuable heuristic source into the philosophy of mind and cognitive science (Petitot et al. 1999), introduces no less traps on the reconstruction of the organizational requirements necessary and sufficient for cognition to arise. We need to be careful on this strategy. For instance we all have an experience of a “temporary gap between need and satisfaction that is provisionally spanned by emotion (desire) and practically overcome by motion”, to use one of Jonas’ phrases. It is easy to project our experience into organisms that meet the above description. But we need to proceed with care. Such projections need to be supported by a circulation between phenomenology and empirical modelling. Unfortunately this circulation cannot be made in positive terms: we are not bacteria and we cannot know what is it like to follow a sugar gradient with a two component signal transduction mechanism (TCST) composed of phosphotransferase and methylation reactions. Nor can bacteria report any state of awareness or intentional insights about their behaviour.

This is not to say, however, that there is no room for phenomenology. As I explained above we can appeal to the strategy of abstracting a conceptual or logical structure and comparing it with available mechanistic models. But we can also elaborate on the consequences that unfold from the analysis of adaptive mechanisms and negatively check if those consequences are consistent

15 For clarity, on what follows, the term **teleology** will be detached from any sense of inwardness or perspective of an agent upon its environmental interactions (a purpose), from any “sense” of the situation or deviation from the norm. In fact, teleology can well be understood as a more general category of which intentionality is but a subclass. It might be equated with the less demanding notion of function or normative functionality in the sense of a certain norm or result be implicit on the working of a part or mechanisms within a whole organizational context. To some extent the whole issue of the feasibility of the C=AB hypothesis pivots around the question of whether the interactive dimension of biological teleology suffices for minimal mentality as intentional agency.

with our inward experience of intentionality. Yet, only negative cases can be decisive, since congruence with our inward experience does not imply that necessary and sufficient conditions are being met; only that no essential structure of intentionality has been violated or that no inconsistency exists between our experience and the functioning or malfunctioning of proposed models. It is only when our experience of intentional action conflicts with some of the consequences of the operations of adaptive mechanisms (or with the way in which we expect them to be supporting intentionality) that we can make use of phenomenology to rule out the case as a genuine instance of intentionality¹⁶.

The case of a source of radiation next to a bacteria provided a paradigmatic example. We can easily imagine being destroyed by a radiation source without being aware of it, without the impact of radiations having any significance for us despite its destructive effect on our bodies (in fact, much more came to be required for scientists and physicians to fear and protect themselves from radiations in the XX century). This way, attribution of meaningfulness or intentionality to a mere disruptive perturbation into autopoietic organization was ruled out. As a result I introduced an additional requirement, that the agent be endowed with an adaptive subsystem capable of generating a behaviour that is in congruence with the maintenance of its biological organization, in relation to the environmental event or conditions that is said to be meaningful for it. Similarly, a new set of requirements might follow if we pay additional attention to the real mechanisms of adaptive behaviour that are found in nature, and their congruence (or rather lack of congruence) with our inward experience of intentionality.

5. INTERROGATING MECHANISMS: ADAPTIVE BEHAVIOUR AND INTENTIONALITY

Since a minimal form of biologically grounded adaptive agency is already found on *E. coli* chemotaxis, it constitutes, for many of the authors mentioned above, a point of departure and most (if not all) of the essential features of

16 We are here following a very similar strategy to Searle's Chinese room argument, which relies ultimately on a phenomenological justification. The fact that the person manipulating symbols while reading instructions does not "really" understand any Chinese (thus providing evince for the failure of the functionalist program) can only be ultimately justified on the basis that if the reader were following those rules she would have no understanding of Chinese. The argument relies on reference to the subjective phenomenological experience that one projects upon the rule-manipulating situation that the thought experiment proposes. The phenomenological contrast between the experience of speaking and understanding a language vs. the experience of manipulating symbols following rules inside a room is what makes the thought experiment powerful. As John Searle recognizes: "The strategy of the argument is to appeal to one's first person experiences in testing any theory of the mind" (Searle 2004: 62).

cognition are taken to be present on it. We have, in previous chapters, analysed in detail the workings of such adaptive mechanisms so it provides a good opportunity for our task. The example of *C. elegans* moving up the gradient of bacterial odour will also provide a useful instance of a minimal case of potential sense-making. The goal of this section is to test whether available mechanistic models of agency are capable of satisfying the minimal structure of intentionality-in-action (as depicted in the previous section) and whether they appear in congruence with our own experience. For doing so I will start from the most simple mechanisms adding increasingly sophisticated operations as *necessary* to approach the structure of intentional agency and evaluate whether they are *sufficient*.

5.1. The case of metabolism-independent chemotaxis

Chemotaxis to aspartate in *E. coli* is one of the most minimal instances of the basic C=AB hypothesis. *E. coli*'s interactive mechanisms (transmembrane Tar chemoreceptors, TCST and flagellar set-up) are tuned so that the resulting movement is that of climbing up an aspartate gradient (a metabolizable substrate for its constituent living organization). It seems reasonable to assume (according to the plain C=AB hypothesis) that *E. coli*'s chemotaxis is a minimal instance of intentional behaviour since it maintains essential variables (in particular the variable *concentration of metabolites*—which would otherwise be continuously decreasing) within viability boundaries. In particular, it seems that an analogy holds with the minimal structure of intentionality described earlier: the interaction corrects an internal tension, the decrease of the concentration of metabolites, back to the norm.

We can consider the following cases for analysis:

1. *E. coli* performs “normal” chemotaxis to aspartate,
2. *E. coli* performs chemotaxis to a non-metabolizable analogue of aspartate (e.g. chemotaxis to D-Fucose, a non-metabolizable analogue of galactose—Adler 1969), or
3. For some reason (e.g. a mutation¹⁷) *E. coli* cannot metabolize aspartate but it performs chemotaxis towards it.

It is important to note that chemotaxis to aspartate has not been shown to be metabolism-dependent (whether the binding substrate, aspartate, is metabolised or not, does not affect chemotaxis)¹⁸. The two component signal transduction mechanism (TCST) is dynamically decoupled from metabolism on its operations (except for the one-directional production of ATP that continu-

17 Pioneering work by Adler (1969) shows how W4690 and SU742 mutants did not metabolize galactose but did perform chemotaxis towards it.

18 Alexandre and Zhulin (2001) only report *E. coli* metabolic dependent chemotaxis for proline glycerol and succinate.

ously feeds the phosphotransferase pathway). This is why cases 2 and 3 can in fact occur, despite type 1 cases being generally the norm.

I will try to show how cases 2 and 3 render the attribution of intentionality problematic (according to the $C=AB$ hypothesis). If intentionality is made dependent upon the compensatory relationship between behaviour and the deviation from the norm defined by metabolism, then the very same conduct, generated by the very same internal mechanism, within the same biological organization, is intentional in case 1 and non-intentional in cases 2 and 3. No behavioural difference exists, no mechanistic difference can be appreciated on the operations of the TCST mechanisms, the only difference is that the molecule that binds to tar detectors is not metabolizable in cases 2 and 3. If we are ready to admit that this is the only difference between the three cases, then it seems that we are making intentionality depend on the effect that an external molecule may or may not have on metabolism. Yet the sensorimotor mechanism operates independently from whether the molecule in the environment is functional for metabolism or not. So, if we are to endorse the plain $C=AB$ thesis, we end up with the paradox that, for this type of cases, attribution of intentionality is independent from the mechanisms that generate behaviour.

One could counter-argue that it is the totality of the organism what makes chemotactic agency intentional and not the TCST mechanism or the nature of the molecule that binds to tar receptors or the metabolic network of the bacteria when taken in isolation. But even if we assume an holistic perspective... How are we to interpret cases 2 and 3? A straightforward answer could be to interpret those cases as failures, after all, intentional behaviour is subject to failure. For instance, the glass of water that we intent to drink may contain Vodka instead of and isotonic liquid, failing to satisfy our thirst. But, the counterargument goes, the failure does not make our behaviour less intentional, so, there we go, this is precisely what cases 2 and 3 illustrate: that the behaviour of the organism fails to bring the situation close to the norm, not that the organism engages in non-intentional behaviour. *E. coli* has an intention but fails to meet the conditions of satisfaction. This is a perfectly valid view of functional failure as judged by an external observer. But from the point of view of the system, this interpretation is not without its problems. In which sense, can cases 2 and 3 count as failed intentions, as frustrated attempts, from the point of view of *E. coli*, if no mechanisms exists for detecting such failure? The only alternative is to appeal to the organismic totality: *E. coli* following a gradient of a non-metabolizable substrate will, sooner or latter, “notice” the effect of such failure due to a progressive decrease of metabolites on its organization. Yet, and this is where we meet a crucial challenge, whether the cause of this progressive decrease originated on an interactive failure or in some internal obstacle to metabolize, is absolutely indistinguishable for the

organism. Therefore, attribution of intentionality or sense-making to its actions, to say for instance that “aspartate is meaningful for *E. coli*”, turns out to be problematic, to say the least.

At least for this type of cases of adaptive behaviour there seems to be no possibility for failure or frustration at the behavioural level (i.e. at the level required to ground intentional agency). Despite Jonas', and others', emphasis on the lack of “awareness of failure” that non-organic behaviour suffers, many organisms equally fail to be aware of such failures, independently of their body suffering the consequences. Life, by itself, does not provide a source of frustration or awareness to failure. There are, for living systems, objective conditions of satisfaction of their being as basic autonomous systems. But life, *per se*, does not provide a sense of conditions of satisfaction of particular *actions* and, therefore, no a priori guarantee for genuine intentionality-in-action.

It turns out that the requirements a) that there be an autonomous or autopoietic organization and b) that mechanisms exist that generate behaviour that is in correspondence or correlated with the norm, seem not sufficient for cognition as intentional agency. At least some organisms that behave adaptively (in the sense that requirements *a* and *b* are met) cannot be interpreted as intentional without entering into severe paradoxical problems. Thus, not only the equation “Cognition = Life” does not hold (if by life we mean metabolism encapsulated on a self-produced membrane) but neither “Cognition = Adaptive Behaviour” holds for those cases where the relationship between the norm (defined by metabolism) and the functional operation of interactive mechanism are dynamically decoupled (thus rendering the intentional failure, and together with it the very possibility of intentional agency, difficult to justify). The case of Cnidarian agency studied in the previous chapter (in the absence of further information about the relationship between metabolic functions and behaviour in *A. digitale*) would also fall under this category. Escape behaviour and tentacle bending for food ingestion are decoupled from metabolic regulation and evaluation. The neural mechanisms that make this behaviours possible are fixed and stereotyped, independently of the specific ontogenetic consequences (for the organisms) of the interactions they sustain.

5.2. The case of adaptive *evaluation* of behaviour

However, following Ezequiel Di Paolo (2005) I have defined adaptive agency in a slightly more sophisticated way, introducing an additional requirement: c) that mechanisms exist for monitoring the state of essential variables so as to evaluate responses according to their trajectories in relation to the boundaries of viability (see figure 18). As Di Paolo explicitly recognizes no meaning

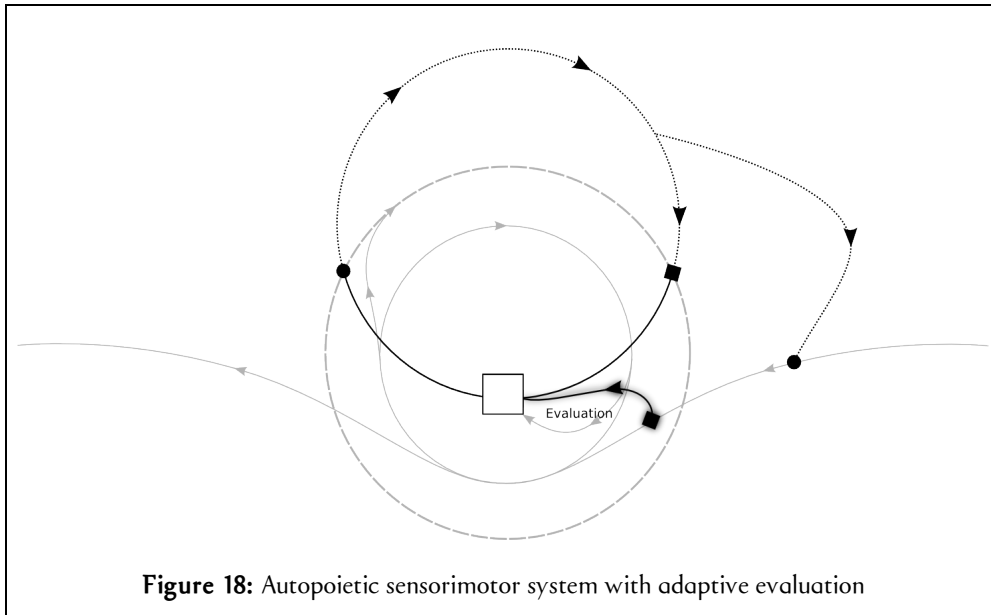
can arise without self-monitoring and regulation (ruling out the cases of chemotaxis that I described above):

Both elements, self-monitoring and appropriate regulation, are necessary to be able to speak of meaning *from the perspective of the system*. Self-monitoring without the right response is (apart from useless) meaningless, since significance must relate to a referential totality (in this case a totality of operations internal to the system). This *paradigmatic* aspect of meaning is provided by the actions of the counteracting mechanisms which differ in degree or in kind for different encounters. Events that provoke the same regulative response are not meaningfully distinguishable. Similarly, regulation without the guidance of self-monitoring is (apart from not possible) disconnected from the source of *syntagmatic* meaning that links the right concatenation of responses to the right situation, that is, its neutralizing or ameliorating effect and subsequent evaluation as such. (Di Paolo 2005: 438–439, italics in the original).

Now, the question is whether adaptive mechanisms of monitoring and evaluation can genuinely ground intentionality on the intrinsic normativity that metabolism provides. Although metabolism-dependent chemotaxis for certain attractants in *E. coli* (as described in chapter 5) may be an instance of a case of monitoring and evaluation, the detailed mechanisms are still unknown (even for *A. brasilense* whose behaviour has been shown to be metabolism dependent—Alexandre *et al.* 2000). We can, however, refer to our case study of *C. elegans* chemotaxis towards bacterial odour (studied in detail on the previous chapter) to analyse this more sophisticated case of adaptive agency.

I shall start noting that the case of *C. elegans* is slightly different and richer in adaptive sophistication than the case of bacterial chemotaxis, for a number of additional features. On the one hand the interaction is not gradually adaptive. No positive change of essential variables can occur until *C. elegans* finds the bacterial colony; unlike the case of *E. coli* progressively increasing the level of metabolites as it moves up the sugar gradient. This is so because behaviour is mediated by indirect perception (bacterial odour is not bacteria) whereas *E. coli* directly detects the metabolite. As we analysed in detail through our journey with Elegantix (see previous chapter) a cognate response is associated with detection of molecules that are correlated with *Serratia marcescens* bacteria. *C. elegans*' NS performs a differentiation function to produce gradient ascent towards the source of the odour.

On the other hand, when the bacterial colony is found, mechanisms exist that can detect the new environmental condition, as a result of which the worm engages in “dwelling” behaviour and starts ingesting bacteria. Thus, unlike the case of *E. coli* described above, *C. elegans*' sensorimotor mechanisms can evaluate its behavioural achievement (that it has reached the food source): the conditions of satisfaction of its interaction are accessible to its behaviour generating mechanisms. Additionally, if bacteria turns out to be infectious, aversive learning is achieved through the release of serotonin on the differentiation function circuit, transforming its odour preferences and avoid-



ing attraction towards them. Connections between neurons of the pharynx (the worm's stomach) and neurons controlling sensorimotor coupling play a crucial role on this “negotiation” between the structure of behaviour and its effect on essential variables (expressed through the activity of the pharynx—a digestive and metabolic proxy). It seems that this example provides a much deeper case of adaptivity, in the sense advocated by Di Paolo, and could come to satisfy what minimal intentionality requires.

According to the amended $C=AB$ hypothesis it is in virtue of its contribution to metabolic self-maintenance, *evaluatable* through the monitoring of the change of essential variables via the pharynx, that moving up the bacterial odour gradient becomes intentional for *C. elegans*. Yet, once again, the above formulation is open to different interpretations. The term “evaluatable” is a key notion here. If we were to interpret it as a definite condition (as *evaluated* rather than *evaluatable*) a quick analysis seems to force us to say that either a) the interaction becomes intentional only after its effect on essential variables has taken place or, rather, after this effect has been “noticed” by the pharynx (backwards-intentionality) or b) the interaction is intentional because of the effect it will have on the future (forward-intentionality). Both alternatives lead to unsatisfactory accounts of intentionality. If we attach to the first (backward-intentionality), an interaction is not intentional but a posteriori, once it has achieved its result and it has been evaluated by the internal monitoring. As a consequence, for the duration of the interaction, there is no intentionality whatsoever; which seriously conflicts with our experience. On the other hand, if we attach to forward-intentionality it is the future that decides upon

the intentional status of an action that takes place in the present. And this alternative is paradoxical, to say the least.

The above interpretation may seem to depend on the assumption that intentionality is an instantaneous thing, which is certainly questionable. So, a way out of the above dilemma, may just require to expand the temporal granularity of the time-scale at which behaviour should be judged or interpreted as intentional. As Di Paolo puts it:

The adaptive event (or act) may be formed by the concatenation and parallel co-ordination of many other regulatory events, but there will be a point below which no further de-composition will be possible without losing the time-structure of the act. At that point what remains are raw processes. There is consequently a minimum temporal granularity in adaptivity. By way of example, it is not possible to judge if the sudden overproduction of a metabolite is part of an adaptive response unless the analysis is extended to a minimal period of observation spanning the immediate past and future. (Di Paolo 2005: 444)

That there is a times-scale below which no intentionality can be said to occur seems out of question, on the light of neurophenomenological research studies showing that, at least in humans, awareness of intentional/voluntary movement takes on the order of a few hundred milliseconds to be built (see Eagleman 2004 for a short review of the subtleties involved). Even if we were to leave aside the issue of conscious awareness in order to evaluate the nature of intentional behaviour, it is out of question that any adaptive event is temporarily extended, that no adaptation takes place instantaneously. The very notion of adaptation involves at least a change that takes time to occur. The question is what counts as an appropriate time-scale for deciding upon the intentional character of an adaptive event. Should intentionality (or lack of intentionality) be attributed to *C. elegans* chemotaxis only at the scale, and never below, of the complete action, including the evaluation or lack of evaluation of having approached and digested bacteria? This would require no reference to backward or forward intentionality since there would be no privileged instant from which a forward or backward could be established. And yet, this interpretation seems to mismatch with our first person experience of intentional action. It is not when I satisfy my thirst or I feel the glass in my hands that my grasping the glass of water becomes intentional. Even if we accept a certain granularity of time for the formation of an intentional act, from its very initiation and through the course of action we perceive and conceive it as intentional.

We can also relax the condition of a definite evaluation and open it to “the capacity for evaluation” as the term “evaluable” in our formulation above denotes. Yet, the capacity without its occurrence is difficult to analyse. One alternative is to extract the consequences of the opposite: the incapacity to evaluate. Luckily we can refer to *mod-1* mutants (reported in Zhang et al. 2005) which are otherwise completely functional but remain, however, unable to

evaluate the infectious nature of their food source and perish inadvertently. Unfortunately such cases make us return to the previous example of *E. coli*. Two equivalent organisms (except for the gene that regulates the expression of a single molecule that is synthesized or activated after the interaction has taken place) perform the same behaviour, making use of exactly the same mechanisms (the same neurons performing the same dynamic motif). Yet, according to the amended $C=AB$, one turns out to be intentional (its interactions are evaluable) while the other is not (its interactions are not evaluable, due to the lack of MOD-1 receptor on neurons AIY and AIZ—see chapter 6 for a detailed explanation).

5.3. The case of adaptive *initiation* of behaviour

It seems problematic to attribute the burden of intentionality to evaluative mechanisms but they seem a necessary condition to model intentional agency if we are not to fall into the paradoxes that arose from the metabolic-independent chemotaxis in *E. coli*. Some complementary capacity or mechanism may be additionally required to fully account for the intentional character of adaptive behaviour. Our debt with Di Paolo's definition of adaptivity (quoted in chapter 5) needs to be finally resolved for it implied that interactions need to be *initiated* as compensatory regulations to the monitored outward tendency of essential variables (not only correlated with a norm and evaluable accordingly)¹⁹. It is time to recapitulate upon the conditions I have been adding to the plain or basic $C=AB$ hypothesis (see figure 19):

- a) that the system be an autonomous or autopoietic organization (defining an identity and a norm)
- b) that mechanisms exist that generate behaviour that is in correspondence or correlated with the norm, and
- c) that mechanisms exist for monitoring the state of essential variables so as to initiate *and* evaluate responses according to their trajectories in relation to the boundaries of viability

The last requirement (that of initiation) would seem to solve, at least, the problem posed by backward and forward intentionality since we can now justify the intentional nature of the interaction on its origin, not only on its result. In turn, the inclusion of this additional condition seems to bring us closer to the minimal structure of intentionality: it is the deviation from the norm measured by the adaptive subsystem that initiates the behaviour and contrib-

¹⁹ Note that the strong version of adaptivity just outlined rules out a big amount of cases of adaptive behaviour (understood in the basic or plain sense) in which interactions are not originated or initiated as a compensation for the deviation of essential variables but on a certain external cue that is dynamically decoupled from any internal monitoring mechanism that links it to its biological (metabolic) relevance. What may be gained in the process is an approach to the structure of intentionality-in-action.

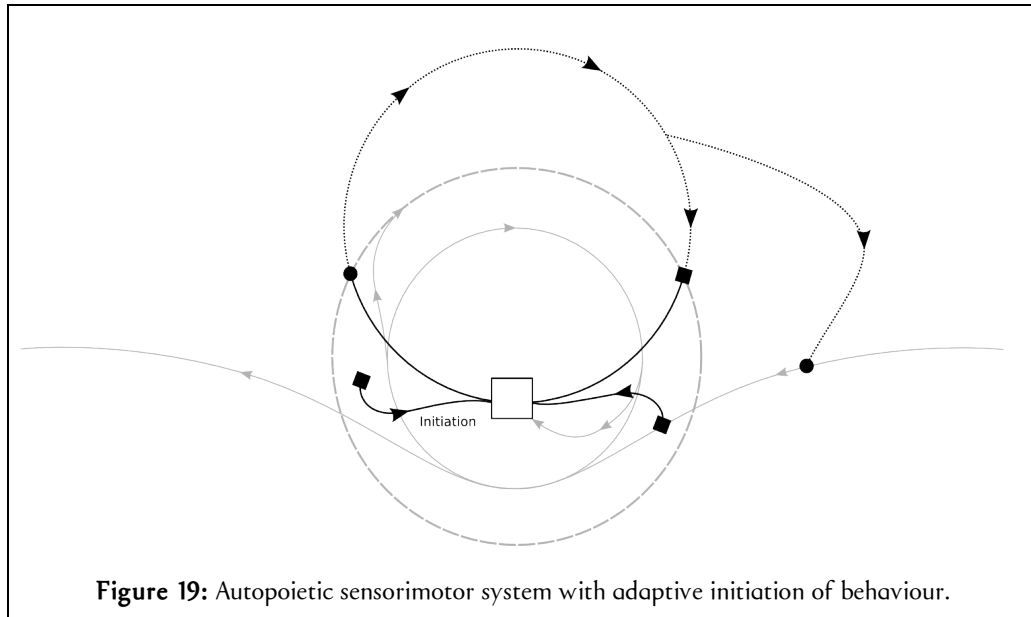


Figure 19: Autopoietic sensorimotor system with adaptive initiation of behaviour.

utes to its intentional character. In addition, the interaction could be said to fail the intention that the organism had when initiated and reshape or correct its intentional behavioural dispositions if mechanisms exist for evaluating and correcting it (as when *C. elegans* encounters an infectious bacterial colony and mechanisms of aversive learning permit to modify the interactive mechanism that lead to the infectious bacterial colony and avoid such bacterial odour in the future—unlike the case of metabolic independent chemotaxis in *E. coli*).

Let us suppose then that the climbing up of a bacterial odour gradient by *C. elegans* is originated on some neural “monitoring” or “measuring” of the falling down of internal energy and nutrients (which seems structurally analogous with our experience of thirst and intending to take a glass of water). Part of this “monitoring” of the essential variables, according to recent studies, may well be instantiated in the pharynx muscles and specifically, for the issue of initiation of food-search behaviour on the activation of MAPK (mitogen-activated protein kinase) (You et al. 2006). We can, for the purpose of this argument, hypothesize that the pharynx muscles, or the neurons controlling it, under the activation of MAPK, induce some change on the behaviour generating neural circuits and that these changes give rise to the initiation of a food-search type behavioural coupling. It seems that this last piece of the mechanisms controlling food-search behaviour (together with the rest of elements we have previously made explicit) will suffice to achieve a mechanistic model that, finally, makes justice for a minimal model of intentionality

But ... What if we were to artificially induce a change on MAPK (and this is not an armchair though experiment but a real condition met in the laboratory and producing exactly the same behavioural response as when real starvation is taking place—You et al. 2006)? Would that triggering or initiation of the behaviour produce genuine intentionality? If the answer to this question is “yes” then ... what does the whole metabolic or autopoietic story really add to the intentional nature of the behaviour that is thus initiated? If we take artificially induced MAPK initiated behaviour to be genuinely intentional, it turns out that the intentional character of behaviour would be independent of its correlation with metabolic or biological norms. On the contrary, if we are to rule out genuine intentionality from a behaviour that is initiated by artificially activating the behaviour generating mechanism we need to postulate that the origin of the activation makes a difference (a crucial difference indeed!) to the issue of intentionality. And yet, to use von Foerster’s phrase, “[t]he states of a nerve cell do not encode the nature of the cause of its activity” (von Foerster 1974: 233)²⁰. How could the *origin* of a signal, which is absolutely causally irrelevant for the dynamic generation of an action, determine its intentional nature?

Think on your being thirsty again. Your thirst may be artificially induced (e.g. by some medicament or by the induction of electrical activity on a certain neural ensemble in the hypothalamus) so that your thirst is not correlated with a real deficit of water on your body. Yet, this lack of correspondence does not preclude your intention to drink and grab the glass of water. That the correspondence between biological norm and interactive monitoring/regulation be lost does not downplay the intentional character of your grasping the glass of water. The temptation arises, from these case studies, to conclude that the biological grounding of intentionality is superfluous; suggesting that it is the *effect* of a perturbation on the behaviour generating mechanisms what is relevant for intentionality (not its origin or correlation with biological norms). I will come back to this idea shortly

6. SOME IN-PRINCIPLE PROBLEMS FOR THE BIOLOGICAL GROUNDING OF INTENTIONAL AGENCY AND A DISJOINT CONTINUITY

We have pointed out a series of problems that any account of the C=AB hypothesis needs to address if the intentional character of the sensorimotor loop is to be grounded in biological normativity. It turned out to be much easier to appeal to our phenomenological experience in order to project intentionality

²⁰ Cited in Ziemke (2005).

into biologically evocative examples of adaptive behaviour than to provide a model of how mechanistic processes are precisely concatenated or networked so as to generate intentional agency. Despite the addition of increasingly demanding conditions for adaptive behaviour to satisfy the requirements for intentional agency, problems of interpretation came repeatedly to the surface when explicit reference was made to particular mechanisms able to sustain it and, particularly, when these mechanisms failed to be correlated with the norms defined by the organism. My hope is that the analysis developed through this chapter can contribute to clarify both the very hypothesis and its feasibility.

As the research on integrated and holistic models of both interactive and constructive cycles in model organisms advances, we shall be able to assess in more detail the relationship between metabolic self-maintenance and the regulation of sensorimotor interactions. Unfortunately, research programs devoted to this task (specially those that highlight or make explicit their holistic interdependencies) are scarce. On the other hand, conceptual simulation models of adaptive behaviour usually take biological organization for granted and tend to model sensorimotor interactions in isolation, optimizing conditions that are externally imposed on the system (e.g. in the form of a fitness function in a genetic algorithm)²¹. Meanwhile some advances can possibly be made at the theoretical level. The issue will certainly benefit from in principle reasons that would clear away or systematically explain the problematic examples exposed above. In order to move forward I shall expose a number of candidate arguments that pinch into key issues that still require further theoretical development.

6.1. Intentional guts?

A significant in-principle problem for the task of grounding intentionality in biological organization has to do with the nature of the qualitative distinction between internal and interactive adaptive processes that makes the latter be intentional and preclude the former to achieve such status. If biological teleology were to imply intentionality then cognition should not only appear at the level of behaviour, for organisms are much more than behaving or interacting systems. Any process that participates on the holistic and circular organization of natural organisms, regulating deviations from a norm, should also be understood as intentional. What makes regulatory system-environment interactions qualitatively different (in relation to the nature of intentionality) from internal regulatory processes? Why should a bacteria or a worm swimming up a sugar gradient be intentional and not a mitochondria's opera-

21 I see Ikegami and Suzuki's model (2008) as motivated by this very issue. Yet, as analysed in chapter 5, their model falls short to address the kind of organization able to encompass increasingly complex agency on that mechanisms of motility are but a differential growth of the metabolic machinery.

tions or any other of the internal adaptive processes integrated on the organization of the system? If perturbations of internal processes (due to infection, heat, intoxication, collisions, radiation, etc.) gives rise to the initiation and monitored regulation of compensatory processes it seems that they should equally qualify as intentional. This, surely, many of the authors defending the $C=AB$ hypothesis will not be willing to accept. But even if adaptivity is taken in the strong sense advocated by Di Paolo no qualitative distinction (from the organismal point of view) seems possibly be made between those adaptive processes occurring as a result of internal changes and those resulting from environmental interactions. Except, of course, for the fact that system-environment interactions are much closer to our sense of intentionality and thus affords for an intentional empathic projection. Note that I am not denying that there are indeed differences between mechanisms in charge of sensorimotor coordination and mechanisms of internal regulation. The previous two chapters were explicitly dedicated to making explicit those differences, to explain how the environment of the organism is expanded and shaped by them, etc. What I am questioning here is whether those differences can provide sufficient reasons for distinguishing internal and interactive compensations *in relation to metabolism as a source of intentional normativity*.

6.2. The problem of dissociation between behavioural mechanisms and metabolic norms

There is, however, a deeper problem than that posited by internal regulatory processes. A quick overview at the problematic cases that we have seen throughout the previous section calls for an in-principle reason that might be thought to underlie the failure of the $C=AB$ hypothesis: *The ascription of an intentional character to adaptive behaviour fails because the operations of adaptive mechanisms are dissociated from the source of the norms that emanate from basic (metabolic) autonomy*. We could call this “the argument of dissociation” between behaviour and norms, stating that adaptive behaviour fails to meet the level of integration required to appropriately intertwine behaviour and norms in order to produce genuine intentional agency.

I will recur again to the example of bacterial chemotaxis to illustrate the nature of the dissociation. The TSCT pathways of chemotaxis to aspartate are dynamically decoupled from metabolism (the chemical reactions involved are causal-dynamically independent from the metabolic cycle—except of the use of ATP). Nonetheless they are both functionally coupled on that the operations of TSCT provide metabolism with the required concentration of aspartate and, in turn, metabolism produces ATP and synthesizes the molecules that conform the TSCT pathway, tar detectors, flagella and motors. Biochemical reactions within the metabolic network are relations of production (that alto-

gether constitute a self-producing network) where dynamical (mechanistic) and functional (normative) domains are unified (the rate of change of a chemodynamic variable has a direct involvement on the self-producing processes). On the other hand, the relation between the operations of TSCT pathways and basic autonomy is not a relation of production but one of transformation of boundary conditions mediated by displacement: dynamical and functional regimes are dissociated²².

We can generalize the above interpretation. Basic autonomy (with its metabolic networked circularity and self-production) provides a genuine source of normativity. However, due to an intrinsic organizational bottleneck on the possibilities for complex self-organization of chemical networks (as analysed in previous chapters), adaptive behaviour must be generated by decoupled sensorimotor mechanisms. This decoupling introduces a fissure on the holistic organization of adaptive agents whereby the generation of norms and the generation of behaviour is dissociated into different regimes. These two regimes appear in turn functionally corresponded. We know that the occurrence of a functional link or correspondence between behaviour and norms can be traced back to random variations (genetic, cytoplasmatic or otherwise) retained at the scale of differential reproduction over evolutionary scale. Therefore, the fixation of the functional correlation may not involve the operational integration of the full organism as a unified dynamical system but the tinkering of different subsystems operating rather independently. These subsystems may appear dynamically coupled at different levels by means of internal detectors and effectors (like the internal monitoring of the effect of behaviour in metabolism). And yet, the form of this coupling does not but reveal the nature of its deep dissociation for it requires the construction of specific transducer channels. The hierarchical decoupling of the NS from metabolism (and its mechanical embodiment) makes this dissociation even stronger for the case of multicellular adaptive behaviour.

The dissociation shows up with crudity on cases of mismatch and incapacity of the organismic totality to assure the correspondence between behaviour and the norm. These are not cases of breakdown of behaviour due to a disruption of sensorimotor coupling nor cases of intentional failure, but cases of mismatch between behaviour and norm, whereby behaviour is completely integral as sensorimotor coupling and yet “inadvertently” maladaptive. If any

22 An increase in metabolite concentration gradient across the membrane has a direct functional relevance for the metabolic organization of a unicellular organism. But a mere correlation between binding to aspartate receptors and flagellar rotation does not, unless it is additionally correlated with a directly functionally relevant change of boundary conditions. When *E. coli*'s behaviour fails to contribute to metabolisms (the functional link or coupling is broken), it is not because its behaviour was “inadequate” but because the concentration gradient is “inadequate” in relation the metabolic organization, and only the former is judged as failure by means of the latter. It is evident that the mode of contribution of behaviour to self-maintenance (its normative functionality) is of a radically different nature than that of a component process of the metabolic cycle.

kind of frustration can be said to occur it is only at the metabolic level and blind to the behavioural “responsibility” of its cause. The dissociation between behaviour and norm makes the normative nature of the sensorimotor coordination invisible for the behaviour generating mechanism thus precluding intentionality in the sense of a genuine interactive compensation of a deviation from a *norm*. This is so because, if the nature of the norm resides on the conditions of self-maintenance of metabolism (on the biophysical and biochemical viability boundaries of the system—its biological essential variables) then, from the point of view of the operations of the behaviour generating mechanisms, deviations from the norm are nothing more than deviations from an arbitrary state (that is fixed by processes that go beyond the operational domain of the organism—i.e. evolution).

The dissociation argument is a strong challenge for the very idea of grounding intentional agency on biological normativity. But ... would the whole problem vanish if it is shown that the full organismic totality is dynamically involved on the production of behaviour and through it on the assurance of the functional integration? If empirical modelling finds that such holism does really take place (at least in relation to some sensorimotor processes) it could provide a compelling counterargument. Even if some correlations fail, the counterargument goes, the whole organisms will soon readjust to recover functional integration. Sensorimotor interactions would not be blind to the norm because the norm would be showing up itself on its capacity to reorganize the mechanisms that generate adaptive behaviour. While the adaptive mechanisms is functioning adequately, it remains decoupled from the totality in the sense that no reorganizing action takes place affecting it. This is why it seems to be strongly dissociated and this is why some researchers can momentarily provide a mechanistic explanation of its functioning independently of the whole organismic context. But as soon as it doesn't fit the norm, the counterargument requires, a global readjustment should take place. Not a readjusted at the evolutionary scale, which is alien to the lifetime of the organism and its operations, but at the ontogenetic time. It is not natural selection that will take care of removing the dysfunctional organisms, it is the organismic totality that will preclude dysfunctionality to appear and correct it if it does. It is not only the sensorimotor subsystem that monitors essential variables and acts accordingly, it is also the metabolic organization that monitors and regulates sensorimotor mechanisms at a different time-scale and assures its normative functioning integrating the behaviour the organisms into a unified totality capable to adjust continuously to its self-generated norms.²³

²³ Fortunately, this issue can be empirically settled down. The problem, again, is that we are lacking models that emphasize the robust and interdependent organization of organisms. There are important paradigmatic reasons for this absence. For instance, only mutations with specific dysfunctional effects are useful for the reductionist scientists and the rest of mutations (the majority) are immediately

Some research areas seem specially well suited to provide new insights into the present discussion. At the unicellular scale the bioenergetic organization of ionic and ATP currents in cellular processes (including its integration in metabolism-dependent chemotaxis) is a crucial issue that could illuminate some holistic interdependent aspects of bacterial metabolic organization and agency (Alexander & Zhulin 2000, Harold 2001, Plaetzer et al. 2005). At the multicellular scale the integration of neurodynamic plasticity with metabolic regulatory processes (both at developmental and behavioural scales) remains understudied. In this sense, the interaction between the neuromuscular structures of the pharynx and AIY and AIZ interneurons in *C. elegans* (during development and adaptive behaviour) seems a promising avenue for research. Unfortunately, there is, however, still little understanding of these key issues in terms of the dynamical and functional integration of mechanisms that span across behaviour, bioregulatory subsystems and constitutive metabolic and bioenergetic processes.

However, there are material and organizational reasons to believe that the dynamic integration of metabolic processes with interactive ones will never be sufficient to ground intentional agency in biological teleology, at least at the level required for sensorimotor interactions to be responsive to the teleological nature of their operations. A full integration will put too strong a demand on metabolic organization and it is reasonable to assume that evolution has moved through the space of sticky functional articulation rather than full dynamical integration (assuming that internal and environmental conditions are sufficiently stable to attain a workable cheap but efficient correspondence between the operations of behaviour generating mechanisms and norms). Empirical research holds the key for a definite answer but the very question was, and still is, in need of further theoretical development²⁴ and computer simulations that integrate interactive and constitutive aspects of biological organization will be crucial on this task.

Two additional problems would still call for clarification even if holistic integrated models prevail. The first has to do with the time-scale of the integration. Even in cases where a bidirectional modulation really takes place (thus

discarded and remain on the shadow. No Nobel prize is given to someone that cannot but state that adaptive mechanisms depends on a still not fully understood totality for their functional adjustment. Most biologists only do research in those areas where there is light for the reductionist methodology to succeed; so that science can embrace a progressive research program. There is such a strong methodological pressure for decomposition and localization of functions that only decomposable mechanisms are known. And the dissociation argument takes advantage of these mechanisms, ignoring those which are not well understood precisely because of their holistic nature.

- 24 I must admit that the above formulation of the problem of dissociation is far from satisfactory. It stands on the crossroad of a number of hot topics in current theoretical biology and philosophy of biology. Among them over the tense intersection between internalist and externalist approaches to the sources of order and functional integration in biological systems. However, rather than on its definite formulation it is on the space that it opens that the problem of dissociation comes to find its place in our project.

assuring a higher order holistic organization) the time-scale of readjustment may still be too slow for specific actions to be considered intentional. The second is to which extent can metabolic normativity account for cases of adaptive behaviour other than those involved on the transformation of boundary conditions of direct relevance for metabolic dynamics. A clear example is escape behaviour. It has an appealing taste for us, since we are powerfully and urgently concerned with any perceived threads to our physical integrity and we immediately associate it with a sense of self-preservation. Our phenomenological experience tells us that we do really *intent* to escape from dangerous situations, that we vividly perceive a tension and a deviation from a norm of safety, that when we cannot satisfy that tension rapidly the situation may drive us to a mode of frustration that takes the form of panic (which can even get to paralyse us, contra any kind of biological normativity). In turn, escape behaviour possesses a clear objective functional role for self-maintenance and a strong normative character if an organism is to avoid its physical disintegration (e.g. from a predator or a falling rock). And yet, escape behaviour is completely independent from its participation on a logic of self-production, understood in terms of metabolic circularity and the teleological grounding it provides. *Aglantha digitale*'s circuit of escape behaviour, or its tentacle autotomy (the release of its tentacle upon touch), provided a clear example. Although absolutely crucial for its mode of being, escape behaviour is completely dissociated from any metabolic process. Even more, if escape behaviour mechanisms are to be functionally efficient, this dissociation has to occur *necessarily* for it wouldn't be quick enough otherwise and, additionally, no ontogenetic evaluation is even possible (since failure is irreversibly deadly).

To be fair to the phenomenological insights on escape behaviour, we must admit that it can take two characteristic forms for us. The first, physiologically closer to that of *A. digitale*, is an automatic and unintentional response to specific triggering stimuli (like closing your eyes upon some approaching object or jumping back upon recognition of a snake attack²⁵). The second takes the form of intentional behaviour that I described earlier (compelling, urgent, stressing and vivid). Both cases bear an equivalent functional relationship with our physical integrity and behavioural contribution to self-maintenance. Thus, it cannot be its relationship with metabolic normativity what makes one intentional and preclude the other to be so.

25 When I last when to visit London's Zoo I spent a nice time at the reptile gallery. A black poisonous serpent attracted my attention on one of the showcases. Fetching as it is the challenging glance of serpents I started to look fixedly to its eyes. I was minded to keep the gaze tight to its end, save as I was behind the security glass. Suddenly, the snake launched a furious bite against the glass. Next thing I remember was the echo of my scream and the astonished look of the rest of visitors of the Zoo, my heart pumping adrenaline violently through my body standing in the floor 2 meters away from the viper's showcase. If it were not for the showcase glass (which was keeping my integrity safe without myself having to do anything) the behaviour could well have been said to be highly adaptive, yet it was not accompanied by any kind of intention on my side.

At this point an unwilling question makes room for our next move: If the source of metabolic normativity is dissociated for these cases but they still can appear urgently intentional for us... Where are we to find the source of normativity that gives reason for their intentional nature?

6.3. The mind has a life of its own

There is a different but overlapping angle from which the problem of dissociation can be more clearly exposed. In order to approach it, I would like to come back to the case of adaptive initiation of behaviour and the problem posed by their artificial or maladaptive induction. Recall the case of the MAPK artificially induced activation of chemotaxis in *C. elegans* or a human induction of non-real thirst. It turned out that, for the intentional nature of the human action, it is irrelevant whether the action is initiated as a response to a “really” occurring metabolic deficit or an artificially induced one. On the other hand, a real need does not suffice to create an intentional response either. You may discover through a medical blood analysis that your level of iron is low and that you “need” to eat spinach. But the “need” (understood as a lack that generates a metabolic unbalance) was there, previous to the analysis, and it did not, by itself, create any intention whatsoever. In fact the problem that metabolic normativity came to be superfluous (in addition to been dissociate) arose several times on our previous discussion and the possibility that the dynamics of behaviour generating mechanisms, rather than those of metabolism, be the key for modelling intentionality and normativity came to the surface.

When trying to ground intentional agency in biological organization it turns out that *the normative criteria is defined from outside the dynamic organization of behaviour*. Essential variables are essential for the maintenance of the underlying infrastructure but not directly for the maintenance of the dynamic organization of agency. In relation to behaviour, essential variables appear as externally and contingently given. From the point of view of behavioural dynamics (at least on the simple cases that we have studied in detail) the situation is not really different from that of a Braitenberg vehicle. Regarding its physical constituency an organism has an identity and a norm defined by itself. Yet, regarded as an agent (in relation to its sensorimotor coupling with the environment) its activity appears highly constrained and almost exclusively *subordinated* to an identity and a norm that does not originate and does not pertain to the domain of sensorimotor interactions: behaviour itself is not autonomous, despite the fact that the organism does distinguish itself and creates a norm at the biophysical level.

Finally, I would like to put this issue in connection with the phrase used by Dreyfus to express the minimal structure of intentionality: “to bring the current situation closer to the agent’s *sense* of an optimal gestalt (...) directly

sensing one's current experience as a deviation from a norm". I have highlighted the terms "sense" and "sensing" to which I have dedicated little attention previously. Note that they do not refer to the sensing of an environmental stimulus, but to the sensing of the deviation from the norm. In particular, it seems that it is the mechanism in charge of adaptive behaviour that should both generate sensorimotor interactions *and* sense their relative deviation from a norm (and not a metabolic core that is, at the time-scale at which sensorimotor correlations occur, busy on its own job). And for this mechanism to become a locus of sensation it would need to generate its own autonomy, an identity of its own *as an agent*, and not just as a metabolic self-sustaining and repairing organization.

The other side of the coin is the fact that *cognitive normativity is not coextensive with biological normativity*. Failure to satisfy agential intentionality does not *necessarily* imply failure of biological adaptation. The opposite of cognition is not biological death (organic disintegration) or the running of biologically essential variables outwards viability boundaries but rather, we shall consider, some kind of "madness" or delirium, a loss of behavioural coherence, a "mental death" which does not, ipso facto, imply a biological death. The alternative is to consider that *the mind has a life of its own* and generates a new level of normativity and adaptive autonomy that must be in place in order to explain genuine intentionality and cognitive organization. And this time both sensorimotor interactions and norms pertain to the same regime (they are not dissociated and functionally tinkered as metabolism is with behaviour).

If we think of a second form of autonomous organization inside another (the metabolic one), a form of organization that is generated through ontogenetic time and in continuous interaction with the environment, then the three problems I have highlighted could be solved. There would be no problem of having to justify the intentional character of internal regulatory processes (intentionality would only occur within the domain of a form of identity that is itself generated and regenerated on the domain of coping). The problem of dissociation between behaviour and the source of norms would be transferred to a new arena: that of neurodynamic organization, free from most of the material constraints that made decoupling necessary at the metabolic level and capable of a degree of dynamic integration far from what metabolic-constructive processes can achieve. Finally, we could explain how is it that an action can be intentional independently of its initiation from a real deviation of a biological norm: it is the unbalancing perturbation to the second autonomous system that would initiate a compensatory intentional action (and not the deviation from a biological-metabolic norm).

Before I move to the next chapter of this thesis I would like to prevent three tentative misconclusions: a) that all of the proponents of the C=AB hypothesis were exclusively associating cognition to metabolic normativity (disregarding other sources of normativity), b) that the organic body and internal bio-regulatory functions are not necessary for mentality and c) that the path traversed along the morphophylogeny of agency was in vain. Regarding the first misconception, Maturana and Varela are best known for their theory of autopoiesis but (as we shall see later) they had both defended the autonomy of the NS on its own right. Di Paolo, on the other hand, has developed an insightful view into non-metabolic values that I will closely follow shortly. What needed to be evaluated here was whether metabolic norms were *sufficient* to ground intentional agency. Regarding the second tentative misconception the morphophylogeny of agency permitted to make explicit the mechanisms underlying agency along different scales of biological complexity. And all the bottlenecks and transition I made explicit will remain important components of the process of constructing a model for minds as conditions of possibility of its appearance and ontological support. In particular, I showed how a domain might be created (that of behaviour or comportment) in which dynamics cut across electrochemical neurodynamics, body and world; a domain in which a new level of autonomy can be envisioned. Finally, and most importantly, our bottom-up approach required to model fundamental concepts like functionality, normativity and agency in a domain that is close to empirical scrutiny and emerges from the realm of physical laws: biological (or basic) autonomy. And it is precisely in analogy with this level of biological autonomy that I shall proceed to build a model of *Mental Life* that can overcome the problems that the rooting of adaptive behaviour in biological norms faced to ground intentional agency. Finally, the mode in which biology and cognition get intertwined in mental life will be an important aspect of its modelling, something that in turn may be able to explain why the biological rooting of intentionality appears so appealing for us.

Chapter 8: Mental Life: the autonomy of behaviour

Cognition presupposes the function of an organization for its own conservation and this is a first fundamental analogy with life.

JEAN PIAGET

The question which troubles laymen, and which has long troubled philosophers, even if it is somewhat disguised by today's analytic style of writing philosophy, is this: are we made of matter or soul-stuff? To put it as bluntly as possible, are we just material beings, or are we 'something more'? (...) [T]his whole question rests on false assumptions. (...) The real concern is, I believe, with the autonomy of our mental life.

HILARY PUTNAM

1. SCAPE FROM MIND PRECLUSION: TWO FABLES AND A JAIL

1.1. *C. elegans*, locked inside the jail of adaptive constraints

We left our morphophylogeny of agency at the point where a number of constraints were capable of organizing behaviour adaptively in relation to its biological embodiment (i.e. so as to maintain integrity and essential variables within viability boundaries). The dynamic organization of behaviour appeared constrained by two main factors that are exogenous (i.e. external in relation to the activity of the NS): the cognate architecture of the neural connectivity matrix and the modulatory effect of body signals on it. We concluded that, when the architecture of the NS is highly modular and scarce on feed-back loops, self-organization may be highly contained within local boundaries (modules such as CPGs) and it is difficult to conceive the organization of the

NS as a result of its own activity: i.e. the order found in neural dynamics is constrained by exogenous factors. I am not claiming that such constraints are contained in any kind of genetic program that fully specifies them. The claim is that, as a result of a complex of developmental processes (including genetic, cytoplasmatic, environmental and epigenetic, not excluding interactive aspects) an order is fixated and that this order is a) neither the result of a process of, properly called, behavioural development (in the sense of a prolonged process of interactive, not just proprioceptive, self-modification and increasing assimilation and accommodation of new behavioural capacities), b) nor is it modifiable through behaviour (except for local and predefined variations). Under these conditions, minds, understood as intentional agency, was severely precluded by the dissociation between the sources of norms and the dynamics of behaviour.

In this sense, the neurodynamics of *C. elegans* can possibly be said to be fully specified by exogenous constraints and be incapable of intentional agency. On the one hand, the connectivity matrix of its 302 neurons appears highly stereotyped and the differentiation state of each neuron is highly controlled by autorregulated transcription factors (Hobert 2006). In addition, the connectivity matrix shows considerably few feed-back connections (only about a 5% of synaptic connections—Durbin 1987) and high degrees of modularity and decomposability that may considerably reduce the role of self-organization at the large scale (Reigl et al. 2004). Moreover, synaptic plasticity seems to be local and restricted to specific conditions, and only about a 10% of the core pattern of synapses seems variable between individuals (Durbin 1987). It is reasonable to say that the behaviour of *C. elegans* is fully specified by exogenous constraints (genetic, developmental, metabolic and environmental), and therefore its normative regulation defined from outside the domain of neurodynamic activity. Under these conditions, minds are precluded and behaviour is, so to speak, imprisoned within the constraint of biological adaptation¹.

As the size and connectivity of neural ensembles increases in encephalized animals, adaptive signals and architectural constraints are not enough to instruct the dynamics of the NS so as to produce adaptive behaviour. Neural modulation can be correlated with metabolic needs and evaluate the effect of behavioural interactions on the body dynamics but cannot specify *how* to achieve adaptive behavioural success. In relation to the architectural constraints, as the size of the NS increases, the number of innate constraints play a smaller role on the specification of neural architectures, leaving it open to the recursive activity of the network and its history of interactions with the

1 However these observations need to be taken with care. The data cited here is mostly derived from anatomical patterns of connections from a few samples. Until recordings in vivo start to inform us about the *C. elegans*' neurodynamic organization it is difficult to reach a definitive assessment of this issue.

environment². *A space of freedom is thus created when neurodynamic mediation of adaptive behaviour overcomes the regulatory capacity of exogenous constraints.* Thus, the NS needs to generate its own regulatory principles through the environment and the body, triggering a process of dynamic self-determination that transcends metabolic norms. At this point, a new form of life (understood as an autonomous form of organization) may be said to appear, embedded on biological life but able to generate its own normativity and value, its own distinctive identity and world, a new mode of agency: that resulting from the preservation of an internal coherency of experience, the coherency of the developmental organization of neurodynamic patterns. *Mental Life* appears. But how can this new space of freedom be reconciled with an inescapable subordination to biological (metabolic and evolutionary) adaptive demands? And what does this new level of autonomy consist of? Two fables will let us open the problem space of this chapter. The first will help reconcile subordination with under-determination (so that organic life can be said to be able to leave room for other forms of life). The second will provide a first bite into the type of organization able to constitute Mental Life.

1.2. The dialectics of subordination and under-determination. A fable on academic research

The following story, about a research scholar that has started a new career with a governmental budget, will serve to illustrate how subordination to adaptive conditions does not preclude new degrees of freedom to emerge at a different level. In order to survive on the academic jungle (the struggle for merit) our scholar needs to satisfy a number of constraints; among them the publication (within a period of two or three years) of a minimum number of papers in relatively high impact-factor journals, the attendance to international conferences, the production of patents for the industry (never for society—which is considered of little value), etc. In addition, these constraints need to be satisfied within the margins defined by the academic discipline in which she acquired her research position. If the scholar fails to satisfy these constraints, she will be ceased by selective institutional forces. Leaving aside the so abundant examples of parasitic, scavenger and tyrannous strategies of domination and intellectual slavery, let us suppose that our scholar manages to satisfy these constraints on the basis of her own academic work and effort. Yet, the constraints are general enough so that they do not specify “how” exactly should she proceed and there is room for her to develop her own academic style, identity, preferences and values.

Over the years, through a history of successful academic research, our scholar has been able to standardize new methodologies, to create new jour-

² See section on development for more details.

als according, to define relevant topics for conferences and workshops, etc. As a result, she was capable of generating novel evaluation standards and the methodological rules constituting a new, previously non-existent, level of normativity and, with it, a new world or domain of academic research. Throughout the process, a rather unspecific selective force (institutional evaluation) has made possible the creation of new norms. Analogously, subordination to the global self-maintenance of its biological body (survival) does not necessarily hinder behaviour to generate its own constraints, preferences and values through a history of interactions, till a properly behaviourally grounded level of identity and normativity emerges. Note that reference to adaptation (in the externalist sense of having to satisfy the requirement of the struggle for survival, biological or academic alike) might be of little help in this quest. On the case of a scholar that develops her own intellectual identity and world, nothing (or very little) can be derived from the constraints of governmental/academic evaluation and standards, except for the gross and plain criteria of satisfying a number of publications every year and few more formalisms (e.g. the structure of a review paper, but never its content). And yet, the satisfaction of academic constraints are absolutely necessary for her academic identity to be created and maintained, materially sustaining her life with the monthly salary that permits her to pay for the food at the university's canteen, the dentist or the loan of her apartment. But nothing on the criteria used to evaluate her work and maintain funding for her projects is able to specify *how* was she going to develop her academic career and *what* kind of research domain was she going to open-up. Instead, what becomes an explanatory source to account for the way in which she came to (self-)determine her identity as a researcher is: a) the history of her intellectual development (in which her background of theoretical and methodological assumptions has been continuously shaped by interaction with her objects of study) and b) the internal consistency of discourse that unfolds on this story. Academic standards under-determine research. This is, in fact, the key for scientific progress in a Khunian sense. And the space that is opened by this under-determination is precisely what permits the appearance of regimes of identity and normativity formation within the space of generic subordination to academic (or biological) adaptation.

Yet it is still to be shown how this new level of norms, identity and world can be conceived on the domain of the dynamic organization of behaviour. The case of a scholar is certainly too far away from our focus of interest and we should avoid reference to higher levels of human behaviour where the linguistic and social domains appear highly intertwined on our capacity to create and recreate forms of identity and normativity; far away from the transition that leads from the dynamics of adaptive behaviour to cognitive organization. However, now that subordination and under-determination can be conceived

as coexisting, an additional example may be helpful to specify in more detail the type of normativity that we are looking for (in sharp contrast with that of biological adaptation).

1.3. A bite on cognitive or behavioural autonomy, a fable on canine rage and the freedom-reflex

Phil comes back home every day after work and the first thing to notice is always the joyful approach of Coggy, his dog. Coggy is a friendly pet always ready to play. It never barks to foreigners, specially if they appear accompanied by Phil. And there is someone that Coggy is particularly kind to: Sophia, Phil's girlfriend. Today, like any other day, Phil comes back home and brings with him (like every Tuesday) a special gift for Coggy, the piece of meat that Matt, the butcher, saves from the leavings of the day. There is however something different today, Phil is not coming directly back from work, as usual, but has spent some time with Sophia. A happy coincidence has made it possible today that both Phil and Sophia could meet each other this evening. They can only meet on weekends since Sophia always, unfortunately, enters job two hours before Phil leaves his. But today, due to some refurnishing that is being carried at her office, Sophia could spend a couple of hours with Phil who, exceptionally, could quit his job two hours earlier. Their meeting in such exceptional situation has made a good occasion for an intense love encounter. Two hours latter Phil is now coming back home at the usual time yet with an unusually big smile and intense bodily odour. While opening the fence of his garden, the fresh and meaty bone for Coggy on his bag, suddenly, Coggy jumps violently into him furiously and aggressively barking, biting violently his leg while ignoring completely the bag with the meat³.

It seems difficult (if not impossible) to defend that Coggy's furious barking and biting into Phil is an evolutionary adaptation, or responds to metabolic values of any kind or even that it is a learned or conditioned behaviour. Remember that Coggy is friendly even to strangers (it is certainly not a habit to bite strangers, not to speak of Phil!), that Phil is bringing him a piece of meaty bone (what a biological contradiction to bite the leg that is feeding you while ignoring the food!), that Phil is coming home as every day, at the same time (no need to punish him for being late!) ... On the contrary, Coggy's bark-

3 This example came to me through a television new of the first woman having a face transplant in France in 2005 (if I remember correctly). Amazing stories tend to come in two and the second one in this case had to do with how the woman came to loose her face. Apparently one day, without previous advice, her dog jumped into her biting furiously her face. The incident, the journalist reported, was due to the fact that she had a particularly different smell that day (and I am probably making up what follows) due to a new fragrance she had used. I apologize for the unwholesome origin of the example. Unfortunately, unlike every good philosopher of biology and mind, I have no dog to rely on for my deepest philosophical investigations and need to rely on morbid television news to find biologically rare, yet real, examples to provide myself with anempirical taste for speculation.

ing seems to respond to some kind of cognitive dissonance regarding a breakdown of the habit of identifying Phil in a certain way. A principle of preservation of experiential coherency (non reducible to evolutionary or metabolic values) seems to be in place if we are to explain, in any sense, Coggy's unusual response. Coggy's expectations and anxiety ("Phil must be coming with my weekly gift") was suddenly disrupted by the unusual, uncomfortable, awkward, uneasy, incoherent and distorting presence of a contradictory strong bodily smell of Sophia on Phil, a monstrosity for a dog's recognition routine; a tense unbalance on its otherwise ordered and comfortably predictable life that had to be removed, attacked, solved to regain the ordered stability of its cognitive world.

If we compare Coggy's bite with any imaginable behaviour of *C. Elegans* or *A. digitale* (not to speak of *E. coli*) it soon becomes evident that reducing it to any kind of metabolic value or preference, any kind of biologically-adaptive story is insufficient to account for it. Paradoxically, many will be tempted to say that Coggy became mad, as if madness could come to rescue us from such an explanatory gap, invoking a deep and mysterious source of randomness and irrationality as an explanatory resource. On the contrary, Coggy's behaviour is completely coherent if we interpret it precisely as the resistance to "accept" the impossible, to oppose the violation of orderliness on its world. It could be said that Coggy's behaviour is the ultimate desperate adaptation to maintain an experiential order. It is not an adaptation to the trajectory of any of the essential variables of its metabolic life but to the equally essential stability of its mental life (a source of normativity on its own right).

Merleau-Ponty (1942:134) reports the case of one of Paulov's dogs that repeatedly resisted to play, once more, the conditioned-reflex game, scratching the floor, biting the furniture, continuously salivating... it became unusable for the experiments. Paulov, jailed himself within the conceptual framework of his experimental paradigm, called it the "freedom-reflex". The very idea calls for a revision of his theoretical framework... "What kind of (conditioned) stimuli is the freedom-reflex a response to?" asks Merleau-Ponty. Of course, evolutionary psychologists will be ready to say that preserving cognitive coherency is a highly valuable evolutionary adaptation, that resistance to madness and fight for freedom is essential for survival. That would be to miss completely the point, for what is at stake here is precisely to explain *what* does this coherency exactly consists on, *how* is this particular type of functionality possible and, more interestingly, *how* is it capable of becoming a source of norms on its own right.

2. THE AUTONOMOUS ORGANIZATION OF BEHAVIOUR: UNPACKING A CONCEPTUAL MODEL FOR MENTAL LIFE

At the most abstract level the question we face is that of finding the kind of organization capable of creating a form of identity and normativity at the level of behaviour. At the most concrete level, the question requires to break down the big question so that a form of organization can be built up from elementary components and their relationship: Which are the components of behaviour? How do they relate to each other to form a coherent and integrated whole capable of producing a locus of autonomous agency?

The search for organizational components of behaviour is certainly not new. The notions of drive, representation, habit, reflex-arc, instinct, schema, neural information processing unit or stimulus-response pair are some of the historically proposed candidates. What is newer is to pose the question within the framework of an autonomous organization; and it is in relation to the organizational context that the election of components needs to be made.

2.1. Searching for autonomy in behaviour

On what is probably one of the most central and unfortunate moves of their theory, Maturana and Varela choose neurons and neuronal processes as a departure point to characterize the organization of behaviour, defined by the *operational closure* of the NS:

“Operationally, the nervous system is a closed network of interacting neurons such that a change of activity in a neuron always leads to a change of activity in other neurons, either directly through synaptic action, or indirectly thorough the participation of some physical or chemical intervening element. Therefore, the organization of the nervous system as a finite neuronal network is defined by relations of closeness in the neuronal interactions generated in the network.”
(Maturana and Varela 1980:127)

In a latter work, Varela, Thompson and Rosch (1991) proposed a simulation model, named *Bittorio*, to illustrate this closure. *Bittorio* is a one-dimensional array of binary units (what is commonly known as a one-dimensional cellular automata) that takes a toroidal form (the first and last elements of the array are connected to each other; the whole array forming a circle). Every component unit can take either one of two states (0 or 1, for clarity). Which state each unit rests on depends on the states of its neighbours by an update rule of the kind “if three of the neighbour units are in state 1 change state, otherwise remain” (the particular rule is irrelevant for this explanation). The dynamics of the network are thus fully specified given a set of rules, a number of units and the initial state of the system. What remains important is the use Varela and colleagues make of *Bittorio* to illustrate the operational closure of the NS. They claim that the basic mode of functioning of the NS is that shown by *Bit-*

torio: each change of local activity propagates along the network and can recursively modify its own state until some kind of equilibrium is reached. Crucially, the environment is considered nothing more than a source of perturbations for the system. Varela and colleagues ask us to imagine the circular automata floating on a soup of “0”s and “1”s so that at any given time one of these environmental elements can touch one of *Bittorio*’s units altering its state. What the effect of such encounter is for *Bittorio* remains, they argue, dependent on the state of the perturbed unit (if the state is “0” and touches an environmental “0” it will be none) and the states of all other neurons, which could make the perturbation be rapidly compensated (falling into the same stable configuration and thus reverting the perturbation) or make it move to a different global state that is still defined by the rules and previous state of the array. According to the authors, the important point to highlight is the following: how environmental perturbations will affect the system will depend on its internal dynamics and cannot be understood as an environmental instruction or input that “causes” an internal state, because the effect of the perturbation will always depend on the dynamics of the network specified by its rules of transformation.

The problem comes in when trying to make sense of the closure of the NS in combination with a “structural coupling with the environment” that needs, in turn, to be “subordinated to the autopoiesis of the system”. How can an operationally closed system be additionally *coupled* to its environment (and not merely perturbed by it) is difficult to conceive, particularly if under operational closure the coupling is to be capable of generating finely tuned adaptive interactions. The tension is evidently seen on the treatment of the role (or rather lack of it) that Varela attributes to the environment

“Sensory and effector neurons, as they would be described by an observer who beholds an organism in an environment, are not an exception to this, because all sensory activity in an organism leads to activity on its effector surfaces and all effector activity in it leads to changes in its sensory surfaces. That at this point an observer should see environmental elements intervening between the effector and the sensory surfaces of the organism is *irrelevant*, because the nervous system can be defined as a network of neuronal interactions in terms of the interactions of its component neurons, *regardless of intervening elements*. Therefore, as long as the neuronal network closes on itself, its phenomenology is the phenomenology of a closed system in which neuronal activity always leads to neuronal activity.” (Varela 1979: 242, italics added)

“The changes that the nervous system’s structure can undergo without disintegration (...) are fully specified by its connectivity, and the perturbing agent only constitutes a historical *determinant* for the concurrence of that changes.” (Varela 1979: 242, italics added)

“(...) the domain of the possible states that the nervous system can adopt as a closed system is at any moment a *function of this history of interactions*, and implies it.” (Varela 1979: 245)

“We must recognize that this effect [size constancy] corresponds to a process that takes place completely within the nervous system, *independently* of any feature of the environment, although it may be elicited by interactions of the organisms in its environment.” (Varela 1979: 254, italics added)

The result is an irreconcilable marriage. Either the NS-environment coupling does dynamically integrate environmental features into the activity of the NS or it does not. Either environmental features are “irrelevant”, the activity of the nervous system is “*independent* of any feature of the environment” and the nervous system functions “regardless of intervening elements” or interactions are “determinants” for changes on the NS, these changes are “elicited by interactions” and the domain of states of the NS is “a function of the history of interactions”⁴. You cannot have both. Unless, of course, interactions are independent of features of the environment, which makes no sense.

Maturana and Varela’s resistance to conceive the environment as an integral part of the activity of the NS is partly derived from their rejection of the notions of “input” and “output” that so strongly defines the representationalist approach they oppose. The other reason for their emphasis on closure stems from the strong internalist perspective, so present on their autopoietic conception of life, which conceived the environment only as a source of perturbations and never as a constitutively necessary source of matter and energy. We have followed Ruiz-Mirazo and Moreno’s conception of basic autonomy as constitutively open and the consequences span to the interactive dimension of the system, since interactions become functional requirements for the very self-maintenance of a thermodynamically “hungry” dissipative organization. To conceive the NS primarily as a closed network (that is additionally coupled to the environment) stands in hard tension with the situated approach to adaptive agency we have developed in previous chapters: the activity of the NS is *dynamically coupled* to the environment. But for Maturana and Varela adaptation/cognition means just “behaviour without loss of autopoiesis” (which renders epileptic-attacks perfectly adaptive, even cognitive) recognizing no essential “need” (on the side of metabolic autonomy) that neurally guided interaction has to satisfy. Consequently, a closed view of the operations of the NS appears perfectly consistent with their internalist view. And yet, that the NS is coupled to the environment, that this coupling is precisely the *raison d’être* of the NS, is so evident that they need to postulate it additionally, rendering the whole internalist approach into a conundrum.

At this point a crucial “omission” in the analogy between *Bittorio* and the NS becomes particularly apparent. *Bittorio*, unlike the embodied and situated

4 The conundrum gets even more complicated when Varela tries to make sense of learning, where the environment is taken as a “source of deformations”, for it turns out the environment has to be simultaneously “irrelevant” and a source of deformations. But we shall leave this additional complications aside. It sufficed to note that taking the NS to be operationally closed and coupled to the environment at the same time is a source of irresolute tensions.

activity of the NS, is not endowed with the capacity to modify its environment or its relative situation on it. As a consequence, the “coupling” with the environment appears just as a mere source of perturbations and not as set of features that are selectively integrated on the generation of a dynamic order in the NS (that serves in turn to satisfy adaptive demands). The risk (but not the necessary outcome) of this situated adaptive perspective, as we saw on the preceding chapter, is to lose a sense of autonomy for the activity of the NS; something that would render behaviour heteronomous and devoid of a locus of identity and worldliness on its own right, subordinated (on its functional dimension) to the demands of metabolism. The natural way to solve the problem would be to say that the activity of the NS is closed *through* the environment, so that environmental features (as selected and shaped by the embodied activity of the NS) become an integral part of interactions and the world that the NS thereby enacts. This alternative fits nicely with the *enactive* (i.e. interaction centred) approach that Varela reclaims for cognitive science while opting for *Bittorio* as a conceptual model (Varela et. al 1991).

When, following Maturana and Varela’s framework, roboticists started to couple recurrent dynamical controllers to a sensorimotor environment (e.g. Beer 1991, 1995, 1997) they soon realized that the activity of the controllers could not be analysed in isolation as an operationally closed network (*à la Bittorio*). Unfortunately, despite having drawn inspiration from the autopoietic theory, a strong conception of autonomy was relegated to the infrastructural aspect (the metabolic autopoiesis of the organic level) and roboticists focused instead on modelling adaptive behaviour (as the satisfaction of the system’s hypothetical viability conditions):

Of course, this explicit separation between an animal’s behavioural dynamics and its viability constraints is fundamentally somewhat artificial. (...) 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: 265)

Not everyone succumbed to this assumption, despite it being a very productive one that has permitted to ground a powerful and important research program that has illuminated so many aspects of cognition (and will certainly continue to do so). Tim Smithers is probably one of the earliest situated-roboticists to reclaim a central role for a strong conception of autonomy at the interactive/behavioural level:

“Designing and building autonomous agents thus becomes the problem of designing and building processes that can support and maintain this kind of identity formation through interaction: *processes that, through interaction, are continuously forming the laws of interaction that can sustain and maintain the interaction needed to form them*. In other words, we need interaction processes that can support the self-construction and maintenance of interaction processes through interaction, in essentially the same way that the material and energy in-

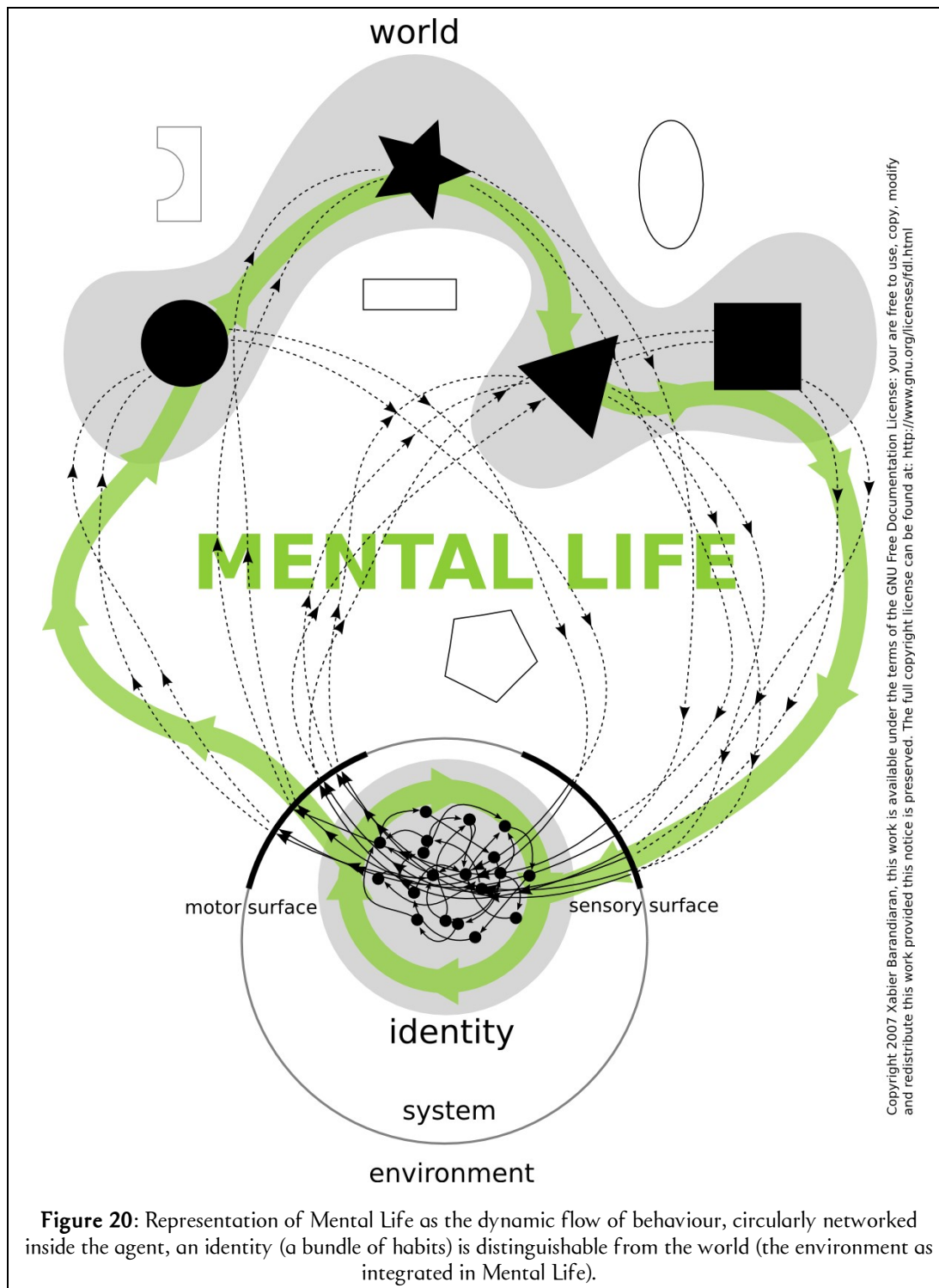
teraction processes of single cells can be understood as being involved in the continual forming of the mechanisms that support this interaction. Such systems will thus be self-law making as well as self-regulating, in essentially the same way as we can understand biological systems and autonomous city states.” (Smithers 1997:102, italics added).

Unfortunately, this formulation of interactive autonomy falls into a circular definition that is difficult to unpack. How can we understand the recursive formation of “laws of interaction through interaction”? As Heidegger pointed out, the real problem of any circular account is not how to escape from it but how to enter successfully into it. Perhaps, an additional circulation between questions of the low level (components of behaviour) and questions of higher level (organization of behaviour) is required to solve the puzzle. Is there any type of behaviour that can itself provide an instance of interactive processes of self-maintenance and law (or structural) formation (and therefore become a building block for a more complex organization that could be called autonomous)?

2.2. A bundle of habits

One entry point into the constitutive circularity of the autonomous organization of behaviour may be provided by the notion of *habits*. A path, that we shall closely follow, inaugurated by Di Paolo for a research program on *Organismically Inspired Robotics* (Di Paolo 2003, Iizuka & Di Paolo 2007, Di Paolo & Iizuka 2008). Needless to say the notion of habits as building blocks of behavioural organization is not new. Aristotle himself developed his ethics (dealing with the pragmatics of action) upon the notion of habit in animals and humans. In turn, the notion of habits became very popular among biologists of the 18th and 19th centuries, particularly in Jean Baptiste Lamarck and Xavier Bichat who stated that “everything in the animal life [contrary to organic or vegetative life] is under the dominion of habit.” (Bichat 1800:127)⁵. Similarly, in his *Principles of Psychology* William James made a similar bet: “When we look at living creatures from an outward point of view, one of the first things that strike us is that they are *bundles of habits*.” (James 1890: 104, italics added). Ivo Kohler (1964) went further when claiming (on the basis of his insightful experiments on visuo-motor habituation) that “habits exist in all areas of human personality” (p. 137) and that only when we undergo a strong process of habituation “do we notice what habit is, and to what extent we consist of many and strong habits” (p. 138). In an even more radical stance on the functional autonomy of habits, Gordon Allport concluded: “The acquired habits seem sufficient to urge one to a frenzied existence, even though reason and health demand the simpler life” (Allport 1937:146).

5 Quote borrowed from Russel (1916:29).



But, what is a habit? We can provide a tentative definition: a **habit** is a *self-sustaining pattern of behaviour* that is formed when the stability of a particular mode of sensorimotor engagement is dynamically coupled with the stability of the mechanisms generating it. The full sense of this definition will unfold as we proceed. We can start by thinking on our everyday experience with habits. Often, for the formation of those habits that we recognize as such (our life is crossed by them but the huge majority appear completely transparent to us) a considerable effort is usually required. Think on the habit of smoking (how difficult it is to smoke the first cigarette), jogging every morning (how tedious it is to confront the physical effort), reading at night (how much resistance it involves for a child) or having a siesta after lunch (this one rarely requires an effort!). Once it is formed, however, the recurrence of the habit is the condition of its own continuation, the very habit “calls” for its exercise and its exercise reinforces its permanence. One way of understanding this circularity is to say that the stability or recurrence of the behaviour that the habit involves (smoking, reading, jogging) both depends-on and reinforces the mechanisms that give rise to it. Like a walking path in the countryside: the more it is used the more it gets clear of grass and rocks, the more clear the more it is used, etc.

Habits bring us the opportunity to penetrate, at least intuitively, into Smithers’ circularity; that of “processes [habits] that, through interaction [through their exercise], are continuously forming [reinforcing] the laws [the mechanisms] of interaction that can sustain and maintain the interaction needed to form them”. I shall come back to Smithers’ statement again later, in order to reconstruct the notion of “autonomy” as an organized bundle of habit. But first I would like to bring attention to the fact that habits, thus defined, possess very interesting properties that make them valuable structures to provide a first intuitive notion for an autonomous organization of behaviour: a) the structure of habits can be traced back to a fully operational/dynamical framework, b) they do not presuppose a distinction or a causal priority between perception and action while integrating both, c) habits are inherently situated or enactive structures cutting across brain, body and environment, d) habits are plastic and malleable (unlike the rigid connotations involved on the notions of reflex-arc or instinct), e) habits provide a concrete sense of self-maintenance (they are both cause and effect of their occurrence, self-perpetuating modes of action) and finally f) habits can be nested or composed at different scales⁶.

6 Obviously the current popular usage of the term “habit” does not fit perfectly with what we probably require at this stage. We consider smoking to be a habit but not to avoid the cars as we cross the road or to drink water when thirsty. Habits seem to us contained within the realm of what is somehow optional and not necessary for survival or to attain a specific goal. Following the every day use of the term, habits form but a small subset of our cognitive life either timing behaviour at high temporal scales (the habit of taking a siesta, waking up early or taking a shower before breakfast), constituting

Note however, that if understood in isolation, habits possess or take over the identity of the agent; it is the very habit that holds a self-sustaining nature not the agent, thus kidnapping, so to speak, the behaviour generating mechanisms of the agent for its own perpetuation (even against the self-maintenance of the agent, as the habit of smoking so clearly illustrates). Notwithstanding, if we take a bundle of habits that cross within the agent (within its behaviour generating mechanisms) a different picture comes out (see figure 20). If habits establish (inside the agent) internal stability dependencies between themselves, an organized web or bundle of sensorimotor structures comes out, defining a set of viability or stability conditions as a whole. In turn, this bundle may be able to sustain itself through the interactions it generates; occasionally “sacrificing” a particular habit for the stability of the whole or requiring specific adaptive interactions to regain stability. In fact, the bundle may be seen as operating in a continuous process of equilibration (to use the Piagetian terminology) whereby it assimilates new situations, accommodating its organized bundle accordingly. It is precisely this adaptive bundle what provides the organizational substrate for a new mode of identity co-defining, in turn, a world of possible interactions on which it depends for its own continuation. The dynamic flow that is thereby created (cutting across brain, body and environment) is asymmetrically laden to the side of the agent as an integrative centre, giving rise to a new mode of identity (causally entangled on the organization of the agent but sustained and extended into its modes of interaction).

On the analogy between life and mind that autonomy affords, what stands as a counterpart for the role of chemical reactions in protocellular systems has been often assigned to the role of neurons or neural connections (as in the notion of operational closure of the NS advocated by Maturana and Varela). One gets a completely different picture when, adopting a radical situated and interactive approach, the role attributed to chemical reactions is attributed to full sensorimotor couplings, understood as habits. The shift from neurons to habits permits to reconstruct the organization of behaviour from an autonomous perspective avoiding the difficulties present on reconciling closure with coupling, since the organizational component units (habits) integrate themselves environmental, bodily and neural processes.

semi-autonomous patterns of behaviour (smoking, etc.) or fixing and reinforcing specific forms of doing something within functionally equivalent alternatives (the habit of heating the milk in a saucepan instead of doing it on the microwave). But in a more general sense, that advocated by pre-Darwinian biologists, habits can be considered as behavioural layers, behavioural schemas that are subject to transformation and concatenation to give rise to complex behavioural repertoires, and more interestingly, able to form modes of organization that sustains and builds itself through interactions. Perhaps deliberating over how to reach the airport in a foreign city can hardly be reformulated as a habit, but crossing the road in front of your house and the sub-tasks involved in the process, like looking right and left, are certainly very important and strong habits. Just remain yourself of the first time you moved to a country where the traffic is inverted in relation to that of your homeland.

Incidentally, the Greek term *ethikos* means “theory of life” and shares origins with the term *ethos* from which the Latin term *habitus* was originated. Most interestingly, the etymology of “*ethos*” is particularly revealing for it shares a profound duality between its meaning “an accustomed place” in which human and animals live or in-habit (a meaning that originated the term “habitat”) and that of “a disposition or character” denoting the personality or character that develops along the lifetime of a human being (or within a theatre play). This is precisely the duality that characterizes Mental Life as a dynamic organization that constitutes both a behavioural identity, a character, and the world it co-defines, a habitat.

We have just outlined an intuitive model of Mental Life. What rests to be done is to make the model more explicit and technically articulated so that it can be naturalized and put in connection with empirically sound models from neuroscience and, through this connection, re-contextualized back into biology. In order to do so, I shall make use of the framework of dynamical systems theory to provide a more detailed formulation of habits, their relation to neural dynamics and sensorimotor interactions and finally reconstruct a conceptual model of self-sustaining behavioural organization.

A last analogy will help clarify the task we are about to confront: Rodney Brooks’ subsumption architecture. (Both similarities and differences between the robotic architecture and the activity of the NS will be crucial.) In the late 80s and early 90s Rodney Brooks proposed the subsumption architecture as a design principle for autonomous/situated robotics (Brooks 1990, 1991). Brooks criticized the traditional AI sense-plan-action architecture (receiving an input, processing it and generating a motor command) and proposed, instead, to build robots on the basis of behavioural layers. A behavioural layer acts as a circuit that, closed through environmental interaction, generates a stable behaviour (such as obstacle avoidance, phototaxis, random search, etc.). The overall behaviour of the robot is the result of internal and environmental interactions between the different behavioural layers: under particular sensorimotor conditions different layers are activated and deactivated, combined or inhibited so that a repertoire of behaviours is generated in continuous embodied interaction with the environment.

The important difference lays on that, unlike habits or the neurodynamic structures that support them (we shall soon see this in detail) Brook’s behavioural layers are built into specific and independent circuits. These circuits are structurally stable in themselves independently of the behaviour they produce. What will characterize Mental Life is a) that behavioural layers are not instantiated on specific circuits (the mapping between behavioural decomposition and circuit decomposition is not one-to-one but many-to-many) and b) that the structure of the behavioural layers depends both on the stability of the generated behaviour (which is the result of the sensorimotor coupling

they sustain) and on the entanglement with the stability of other behavioural layers. As a result, the component structures of the control architecture and the structure of behaviour will depend recursively on each other. A detailed characterization of these emergent and plastic “circuits” and their interactive and internal stability dependencies should occupy us on the following subsections.

2.3. Neurodynamic structures and hyperdescriptions

We have said little about the nature of behaviour generating mechanisms in relation to habits on our previous intuitive sketch. Yet habits, and particularly bundles of habits, require specific properties from behaviour generating mechanisms. These need to be *plastic* (for otherwise issues of stability would make no sense) and able to *integrate* different sensorimotor correlations within the same assemblage, so that habits can be nested to form a bundle. The first thing to note is that specific neural pathways or circuits (like those involved in reflex-arcs) will not do the job; it is rather a network capable of generating different dynamical patterns (and able to modify itself accordingly) what is required. The notion of an emergent pattern constitutes our first building block of a neurodynamic organization. I shall thus introduce this concept in some detail as the general form of which habits may be taken as a particular instance.

Neurodynamic structures may be equated with the more familiar notions of “structures of comportment” that Merleau-Ponty (1942) conceived as emergent forms essential to understand animal and human behaviour or as a kind of emergent “behavioural schemas” within the Piagetian approach to cognition and intelligence (Piaget 1967/1969—see Arbib 2003 for a development of the Piagetian concept of schemas within contemporary cognitive science). More recently, within the dynamical system’s approach to cognition, similar notions have also been proposed (Kelso 1995, Varela 1995, Fuji et al. 1996, Llinás 2001, Freeman 2001, Tsuda 2001, etc.—drawing upon Hebb’s pioneering work on cell assemblies) but we shall pay attention to some of these developments latter. By **neurodynamic structure** I mean the structured activity of a subset of neural variables and their relationships involved in a certain sensorimotor coupling. A neurodynamic structure emerges when (for a given time window) we can systematically reduce the dimensionality of the internal organization of the NS to build a mechanistic-dynamical predictive model of the behaviour of the system. Although they do not necessarily have to take this form, neurodynamic structures may be thought of as attractors within a higher level phase space of a neural network (e.g. when synaptic connections remain stable through the time-span in which the activity of the neurons fall into that attractor).

A neurodynamic structure is dynamic in two senses. First, it is a structure of change, a region of dynamic trajectories in the phase space (e.g. a local chaotic attractor of neural electrochemical activity). Thus, by structure we don't refer here to any physical or anatomical component or ensemble but to a dynamic form that emerges in a particular interactive coupling with the environment. Second, by dynamic we mean that the structure can also change over time: the “shape” of the phase space region is subject to different stability conditions that may alter it. As we shall see, this dialectic relationship between two (or more) temporal and dimensional scales of change is crucial.

Dynamic structures can be operationalized using the notion of hyperdescription developed by McGregor and Fernando (2005) for the analysis of hierarchies of dynamical systems. A **hyperdescription** is a dynamical system which is a higher order and shorter predictive description (S') of a “lower level” dynamical system (S). Three conditions must be met for a dynamical system S' to be a hyperdescription of another dynamical system S : a) *specificity*: there exists a deterministic description function $S \rightarrow S'$ that, for any given time, translates states of S into states of S' ; b) *state-dependency*: the state of S' provides information about the next state of S' and; c) *distinctness*: the state of S' predicts better the next state of S' than it predicts the next state of S . Thus S' is a hyperdescription of S if we can systematically translate S states into S' states (the other way round might not be the case), S' has a predictive value (is close to state determined) and S' is a good predictive model of the higher level regularities, but a bad model for the whole systems in general. Thus information is lost from S to S' but causal organization is maintained within a shorter predictive description. As McGregor and Fernando puts it the distinctness condition “defines a sense in which a higher-level description of a system can *follow a new set of laws* compared to its underlying system” (p. 466, italics in the original). One can think of a hyperdescription as higher order phase space or attractor landscape. Neurodynamic structures can then refer to attractors within this higher order landscape and one can, for a given time window, establish a hyperdescription of the NS and decompose its activity into a set of dynamic structures that predict its behaviour, each corresponding to different couplings with the environment. (I will soon come back to this framework, in order to illustrate it with a clear example.)

A number of dynamic phenomena can be distinguished so far. **Structural formation** is the process of stabilization of a set of relationships between neurodynamic variables and their coupling with the environment so that a hyperdescription becomes available, an attractor landscape is formed and the trajectory of the system falls into an attractor within that landscape. **Structural stability** will then refer to the maintenance of this landscape and the attractors within it. As we shall analyse later, structural stability can depend on a set of sensorimotor correlations *and* on stability dependencies created

between different structures. An important concept is that of **stability conditions**, naming the dynamic constraints (internal or interactive) that bring about the stability of a neurodynamic structure. **Structural transitions** refer to the switching between different structures (the trajectory moving from one attractor to the other within the same landscape). **Structural change** happens when the state space of hyperdescriptions must be accommodated to predict the trajectories of the system through it, i.e. when the landscape is modified (e.g. new sensorimotor correlations are established dividing a basins of attraction into two or more different basins). Finally, **structural disintegration** happens when a structure disappears.

2.4. Interactively dependent stability: a case study in homeostatic plasticity and a single habit

Behavioural stability depends on the way in which neurodynamic structures are coupled (through the body) with the environment. If neurodynamics had no structure at all, undergoing a completely chaotic regime, behaviour would be unstable and unstructured; there is nothing surprising on this. However, to conceive the situation on the other direction may turn out to be more interesting: i.e. to conceive that behaviour sustains dynamic structures too, that the stability of a neurodynamic structure depends on the particular sensorimotor correlations that the coupling it sustains generates. This is precisely what the notion of habit was meant to capture. If this coupling is lost, structural transitions happen (the system moves to a different attractor) or the system undergoes structural change, i.e. it enters a region of structural instabilities until a new structure or set of structures are stabilized/created. Let me bring into stage one of Di Paolo's pioneering simulation models on homeostatic plasticity to illustrate the case of a single habit. This case study shall, in turn, help clarify the conceptualization of neurodynamic structures as hyperdescriptions.

During the 50s and 60s Ivo Kohler (1962, 1964) systematically studied readaptation to visual inversion, through the use of goggles built with mirrors so that the visual field of the subject had a right-left inversion⁷. After a period of two weeks of severe difficulties to coordinate behaviour, subjects wearing inverted goggles started to behave coherently. Soon after they reported that the whole perceptive left-right regularities started to emerge again in their perceptual experience of the world, re-inverting the visual effect of the goggles (although in a fragmentary way, with certain objects swooping left-right inversion while the rest remained inverted!⁸). Interestingly, after goggles

7 Different types of goggles were also used: up-down inversion, coloured glasses, distortions, etc. and despite Kohler's significant contributions to the field he was not, by far, the only partner of this experimental field inaugurated by Stratton during the 19th century.

8 There is no space here to reproduce the insightful details of the whole process of rehabilitation

were removed subjects reported that their visual field appeared upside-down again and they only recovered “normal” vision after a new process of rehabilitation had occurred.

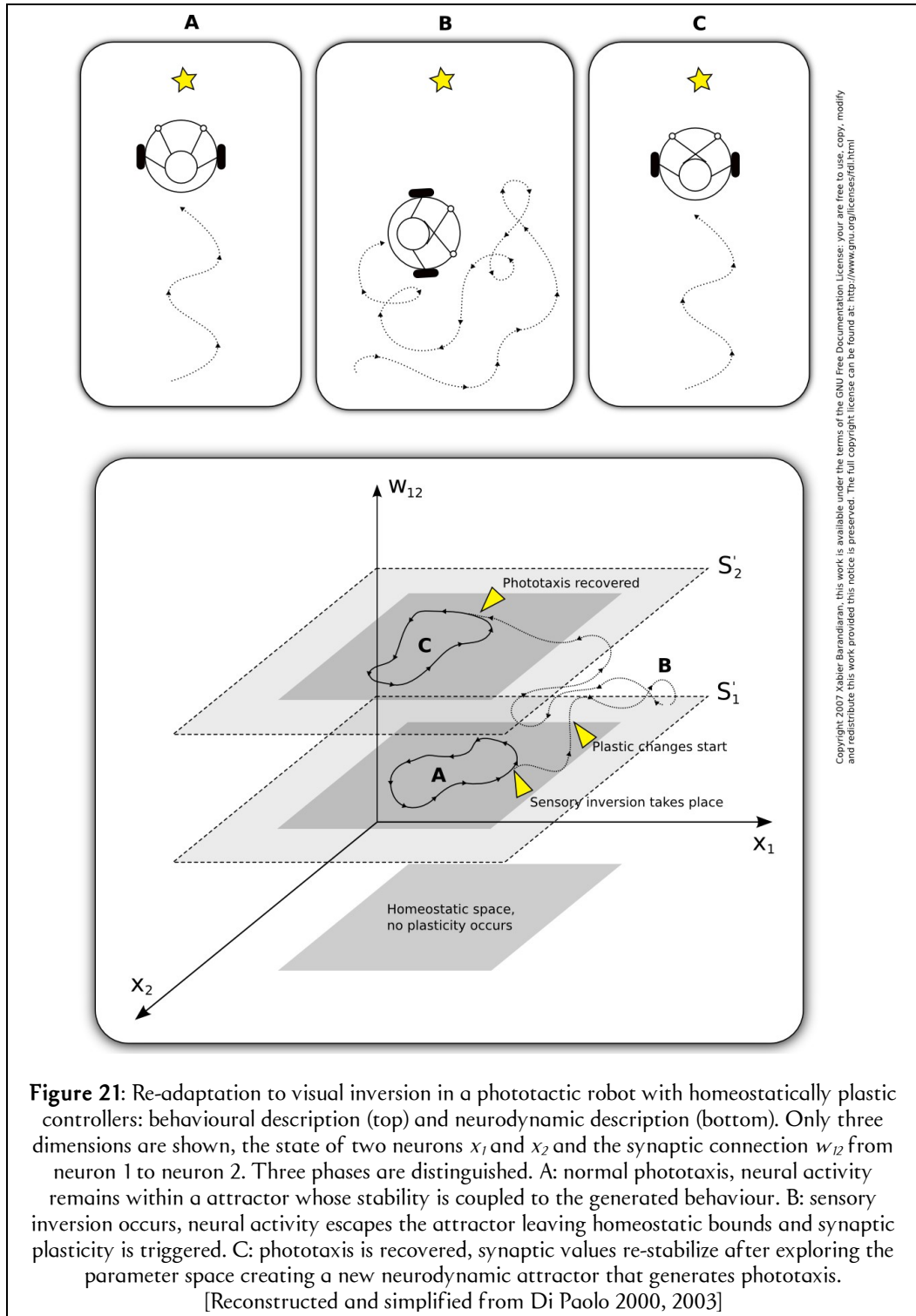
Based on these experimental results in humans, inspired on Ashby's Homeostat (Ashby 1952) and supported by the neuroscientific evidence that synaptic plasticity is regulated homeostatically (Turrigiano 1999) Ezequiel Di Paolo (2000) artificially evolved a robotic simulation model where robotic agents were able to rehabilitate to sensory inversion (without the agents being selected for that task during evolution). The control architecture (a continuous-time recurrent dynamic network with Hebbian plasticity) had two dynamic regimes: one composed of fast neural action potentials and a slower one of synaptic plasticity. The parameters of the synaptic plasticity functions (how much and in relation to what—pre- or post-synaptic correlations, etc.) were subject to evolution. Plasticity was constrained to be activated only when neural activity was over or above certain pre-specified boundaries. In addition, the plastic changes were induced to compensate for the upper or lower violation of the boundaries (e.g. if a neuron's firing rate is too high connection weights to that neuron are reduced, if a neuron is firing low connection weights increase). Hence the name “homeostatic plasticity” since synaptic parameter modification is only activated when neural activity goes out of certain pre-specified bounds, leading to its re-stabilization. The genetic algorithm used for artificial evolution had two fitness functions (i.e. two different optimization criteria were used to “design” the agents). The first was a behavioural measure of how close the agents got to the source of light, thus selecting for phototaxis. The second measured how much plasticity was induced during each trial, selecting for stability, i.e. to diminish the amount of synaptic change during the task which, in turn, made the activity of neurons remain within homeostatic bounds for as long as possible.

After artificial evolution had optimized the parameters for this task, agents were tested for visual inversion. At the beginning of the trial agents performed phototactic behaviour: the agent was able to navigate a 2 dimensional space approaching a light source (condition A at the top side of figure 21). Later on the trial, right and left light sensors were inverted completely, subsequently disrupting phototactic behaviour (condition B at the top of figure 21), until phototactic behaviour was recovered, stabilizing neural activity (condition C at the bottom of figure 21). Agents were not evolved to adapt to sensory inversion ... how could, then, phototactic behaviour be recovered? The experiment demonstrated that, by evolving the agents for phototaxis while selecting for internal synaptic stability, both synaptic stability and behavioural stability became evolutionarily coupled. As a result, ‘normal’ pho-

reported by Kohler, directly reading the original article is a highly recommended and invaluable experience of Heideggerian empiricism.

totactic behaviour was sustained by a stabilized set of synaptic parameters on the agent's control architecture. When the agents' sensors were inverted behavioural coherence was lost and their internal synaptic dynamics entered an unstable region. The instability of synaptic parameters, in turn, produced behavioural instabilities (the agents performed "random" movements) and the synaptic parameter space was explored until phototactic behaviour was recovered again which, in turn, stabilized the values of synaptic parameters.

We can interpret this experiments within the framework of neurodynamic structures and hyperdescriptions to illustrate how interactively dependent stability fits into this framework. I shall repeatedly refer to figure 21 where behavioural descriptions (A, B and C on top) are corresponded with the descriptions at the level of neurodynamics (A, B and C bottom). For easy of explanation pictures have been simplified and idealized from original data (Di Paolo 2000, 2003). Only three dimensions are shown, the state of two neurons x_1 and x_2 and the synaptic connection w_{12} from neuron 1 to neuron 2 (the resulting dynamics can easily be generalized to higher dimensional systems). The agent's control system is composed of two types of variables: synaptic connection values and neural activation values. The former being slower than the latter, the hyperdescription is straight forward: we can create a hyperdescription of S (the dynamical system composed of both neural and synaptic variable) by just removing the synaptic weights as state variables and considering them as fixed parameters (S'_1 in the figure 21). We have thus reduced the dimensionality of the state space (reducing the amount of descriptive information) while maintaining predictability. In addition, since synaptic change is induced only when neural activation values exceed some boundaries (grey box on the x_1 - x_2 plane) we can further assume that while normal phototaxis is taking place the state space trajectory will be reduced to this boundaries, which is, in fact, what occurs in the experiments. This way the new description of the neurodynamic system will have a particular dynamic topology. For simplicity we can assume that when coupled to the environment this topology shows a single basin of attraction for a cyclic attractor that corresponds with phototactic behaviour (condition A in both top, behavioural, and bottom, neurodynamics, of figure 21). This is our dynamic structure: a stable behavioural layer instantiated in a hyperdescribed "circuit". The robot has the habit of approaching the light source.



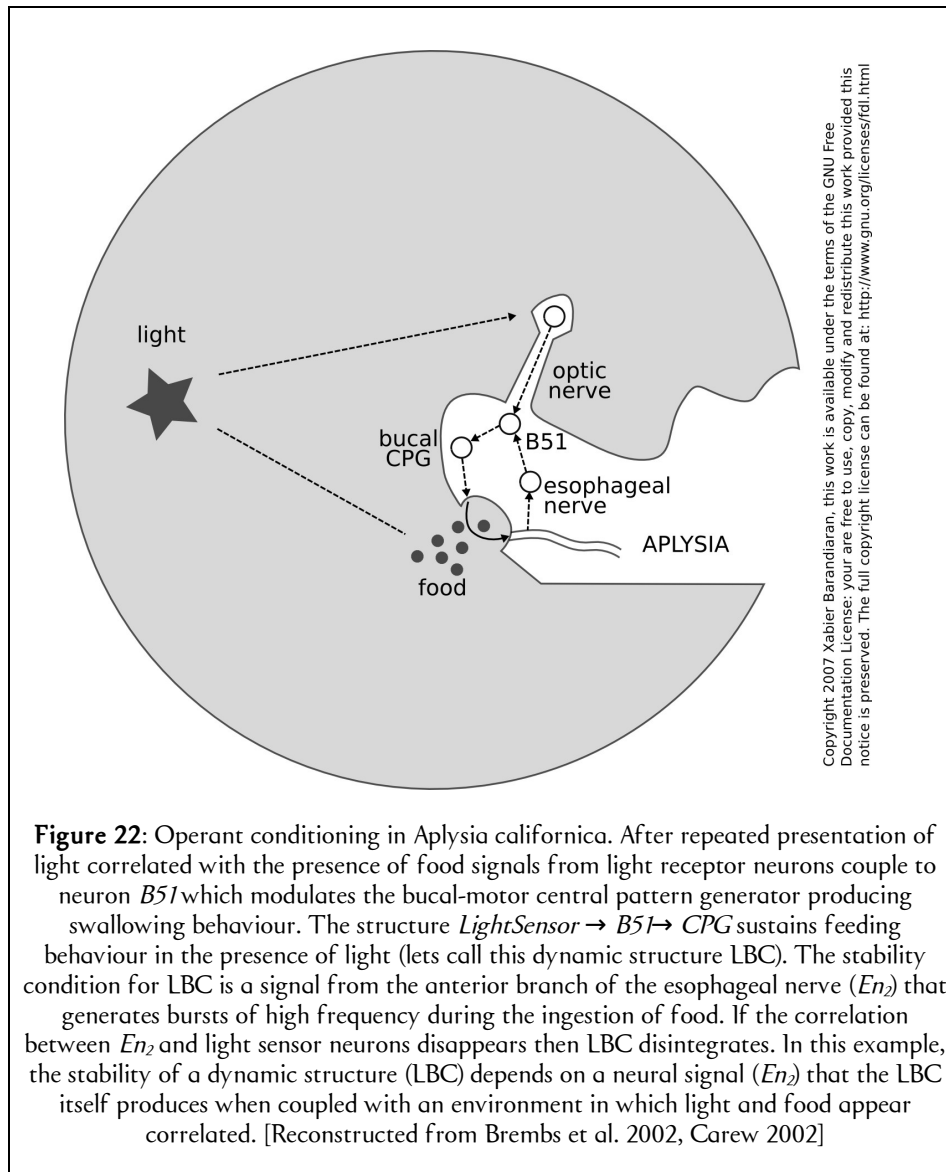
Right after sensory inversion takes place the agent cannot generate the sensorimotor correlations that maintain its neural dynamics within the cyclic attractor: the habit that relied on this coupling is disrupted. As a result, neurodynamic trajectories escape from the dynamic structure **A** within the hyperdescription S'_1 and, at some point, it crosses the homeostatic boundaries and synaptic plasticity is activated. At this stage our hyperdescription does not hold any more and the dynamic structure that previously generated the phototactic habit disintegrates. We need now include synaptic connection values as variables (we need to come back to descriptive framework S). Instability of synaptic values induces instability of neural activation levels. Behaviour becomes erratic and we may have entered a completely chaotic regime (condition **B**). But artificial evolution has done its job, finely tuning the learning rate parameters of the synaptic values so that synaptic stability and phototactic stability have come evolutionarily coupled. Sooner or latter the erratic change of synaptic connections will lead to phototactic behaviour and synaptic stability will be recovered: a new hyperdescription is now available (S'_2 in figure 21) and within it a new dynamic structure **C** appears (a new attractor, behaviourally equivalent to the preceding one).

2.5. Extending the case of single habit

Now it can clearly be seen how the stability of a dynamic structure may depend on the behaviour it sustains and, conversely, how the stability of behaviour depends on the stability of the neurodynamic structure. This is a case of stability dependency characteristic of a single habit that depends on the ongoing coupling between agent and environment. But, of course, habits need not be continuously enacted for their permanence, a system may alternate between different neurodynamic structures and the behavioural coupling they generate and sustain. What remains a requirement for habits is that their recurrence is necessary for the stability of the neurodynamic structures leading to them, for otherwise these structures will gradually disintegrate.

Maintaining the framework of neurodynamic structures, we can now move beyond the case of simple habits which, however inspiring, are severely limited to deliver a picture that makes justice to the variety and richness present in animal behaviour. For instance, early neural development may be evolutionarily canalized to stabilize (under varying environmental conditions) certain key neurodynamic structures. A paradigmatic example can be found in the well known cases of parent-following imprinting in ducks studied by Lorenz. These structures may remain stable for a long period of time and slowly vanish as new structures are formed.

But stability conditions may also appear to depend on future sensorimotor correlations, not only on the continuous ongoing engagement with the environment. This is the case of operand conditioning and reinforcement learning.



We can further develop this idea with a neurobiological example of operant reward learning in the sea snail *Aplysia Californica*, whose neural and molecular details were recently discovered by a number of experiments (Brembs et al. 2002, Carew 2002). Through consecutive presentations of light in the presence of food in its medium, a dynamic structure is created in the *Aplysia*'s NS so that in the presence of signals from light receptor neurons, neuron *B51* modulates the buccal-motor central pattern generator producing swallowing behaviour. The structure *LightSensor* $\rightarrow$ *B51* $\rightarrow$ *CPG* sustains feeding behaviour

in the presence of light (lets call this dynamic structure LBC—see figure 22). But there is a sense in which the behavioural consequence (feeding) sustains the structure too. The stability condition for LBC is a signal from the anterior branch of the esophageal nerve (En_2) that generates bursts of high frequency during the ingestion of food. If the correlation between En_2 and light sensor neurons disappears then LBC disintegrates. In this example, the stability of a dynamic structure (LBC) depends on a neural signal (En_2) that the LBC itself produces when coupled with an environment in which light and food appear correlated⁹. A behavioural scheme depends on a reinforcement signal as its stability conditions.

The above example is drawn from a localized reflex-like neural structure but nothing prevents us to generalize to more complex higher level hyperdescriptions comprising more complex neurodynamic structures. We can now define expectations as dynamic counterfactuals (conditionals): if a certain interactive condition is not met during or after a certain behavioural coupling the dynamic structure involved in the coupling dissolves or its stability diminishes. This way we can say that dynamic structures may depend on the satisfaction of expectations as delayed stability dependencies. Taking neurodynamic structures as component primitives of neurally controlled behaviour, we can still further expand this framework in order to achieve a full fledged organization. For instance, a neurodynamic structure may depend on multiple stability conditions. In turn, different stimuli can bring about the same neurodynamic structure or neurodynamic structures may be concatenated on specific sequences. More interestingly, we can conceive the entanglement or entrenchment between different neurodynamic structures. The most simple case concerns the relationship between two dynamic structures so that the stability of one of the structures depends on the stability of the other and vice-versa. The habit of smoking while having a coffee may provide and intuitive example of this simple case. One can quit having coffee to help stop the habit of smoking. A more complex case may involve a whole set of neurodynamic structures and habits, like those involved on complex visuomotor coordination (as studied on its behavioural and phenomenological aspects by Kohler), so that inversion of the visual field requires a long process of reorganization of habits.

Given an integrated neural architecture and the non-localized nature of some dynamic structures, it is highly probable that the stability of certain dynamic structures be dependent on the coordinated activity of the whole sys-

9 This presentation of the facts, following the explanatory template of the authors, probably makes the notion of “conditions of stability” more clear. Note, however, that a Piagetian inversion of the processes involved is in fact more appropriate to explain *Aplysia*’s behaviour. The departure point should be the $En_2 \rightarrow B51 \rightarrow CPG$ circuit, closed through the action of swallowing, constituting a kind of self-reinforcing habit (... *swallow* $\rightarrow$ *food* $\rightarrow$ $En_2 \rightarrow B51 \rightarrow CPG \rightarrow$ *swallow* ...). And it is to this pre-existing habit that the presence of light is assimilated.

tem at different time-scales. The stability of a given dynamic structure may depend on the existence of a network of other dynamic structures which are in turn dependent on the “expectancies” or stability conditions that are recurrently satisfied through behaviour. The instability of a particular dynamic structure could generate a cascade of instabilities and transformations propagating along the web of dynamic structures and reconfiguring it. Thus, from the point of view of transitions and stability, *webs of dynamic stability-dependencies* can be created between dynamic structures too (in addition to behavioural stability dependencies) so that a global structural interdependency emerges (similar to those that are found between different reactions in self-catalytic networks).

Michael Arbib reaches a similar conclusion when considering a higher level organization of behavioural schemas:

“Through learning, a complex schema network arises that can mediate first the child's, and then the adult's, reality. Through being rooted in such a network, schemas are interdependent, so that each finds meaning only in relation to others. (...) Each schema enriches and is defined by the others (...). Though processes of schema change may affect only a few schemas at any time, such changes may “cohere” to yield dramatic changes in the overall pattern of mental organization. There is change yet continuity, with many schemas held in common, yet changed because they must now be used in the context of the new network.” (Arbib 2003:998)

The organization of behaviour can thus be pictured out as a nested web of neurodynamic structures: a interdependent bundle of habits that continuously maintains itself through interactions.

2.6. The autonomy of behaviour revisited

We have seen how dynamic structures can be operationally defined, what the relationship between neurodynamic structure stability and behaviour can be and how this provides a first sense of self-maintenance in the domain of neurodynamic organization of behaviour. In addition, these structures can be nested, i.e. organized, at different scales and by different means (concatenation, integration, etc.). Now, given the far-from-equilibrium condition of this web (it needs to be maintained through interactions), we could hypothesize that a self-maintaining organization emerges when the web as a whole is homeostatic and behaviour is directed towards the adaptive maintenance of the global stability conditions of the web. It is not any more an adaptation to material or biological self-preservation but higher order regulation for the stability of neurodynamic organization. From a set of initial conditions of great developmental plasticity triggered by biological adaptive signals and channelled by architectural and body constraints, the NS generates more and more internal constraints and interdependencies between behaviourally emergent self-organized patterns (neurodynamic structures). Gradually, the preserva-

tion of the internal coherency of these nested structures takes over the regulation of situated and embodied brain dynamics. Of course, throughout its development this neurodynamic organization must have internalized biological-adaptive requirements in the form of internal stability dependencies. But just like the case of research scholar, the mode in which it has done so, and the resulting organization is not any more something reducible-to or deducible-from an evolutionary adaptationist story, nor from the adaptive requirements of metabolism. **Mental Life** comes into existence when the *adaptive conservation of the internal organization of neural dynamics becomes the main principle of sensorimotor regulation*.

We can now come back again to Smithers' definition: "identity formation through interaction: processes that, through interaction, are continuously forming the laws of interaction that can sustain and maintain the interaction needed to form them." The organization of behaviour is supported and generated by neurodynamic organization. In turn, neurodynamic organization, on its most abstract conception, consists on a dynamic space S able to *generate*, through interaction, a structural web S' of neurodynamic patterns in which a new set of "laws" (i.e. predictive hyperdescriptions) hold, able to *maintain* the stability of this web and *adaptively regulate* it in a continuous process of re-equilibration. A new level of identity (*auto*) and normativity (*nomos*) is created in the process beyond the biological one.

If the preservation of internal global stability in continuous equilibration is the governing principle of neural organization, a model for minds demands explanatory resources beyond those provided by exogenous constraints. It is the developmental history, the internal organization and coherency and the world that is created what needs to be brought into explanation to account for Mental Life. Like in the case of Di Paolo homeostatic robots, it is the coupling between internal and behavioural stability what becomes an explanatory source, not evolutionary adaptation to a task that never required, in evolutionary time, adaptation to visual inversion. The adaptive anatomy of behaviour, thus understood, becomes a new source of normative functionality on its own right (not a mere appendix of biological adaptation). Coggy's behaviour is not a random malfunction as if an evolutionary or optimality *ceteris paribus* clause was fortuitously violated driving us into an explanatory *limbo*. Coggy's behaviour can now become a proper explanandum as a *cognitively-adaptive* behaviour that acquires both functional and explanatory sense in terms of preservation of internal coherency and stability

Up to now we have sketched a conceptual model for the autonomous organization of behaviour, a model we have labelled Mental Life. The very idea certainly requires a deeper analysis since its consequences extend far beyond this schematic formulation and it deeply touches on the characterization of cognition. But, as it stands now, Mental Life is nothing more (nor less) than a

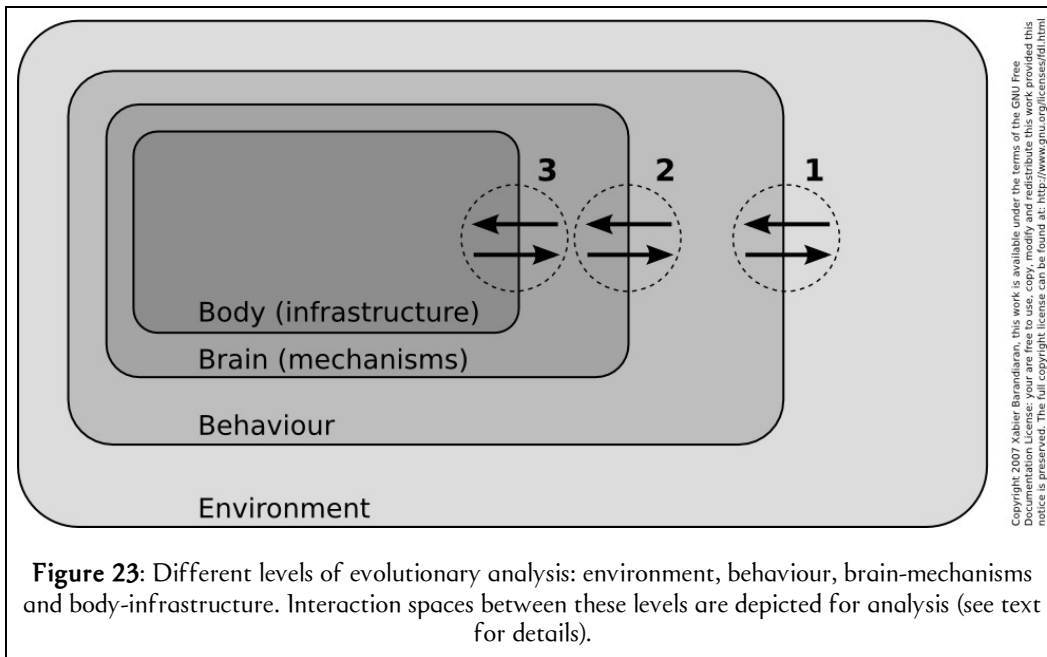
conjecture model; a regulatory concept that defines the horizon of the path we shall follow. This preliminary approximation was necessary to provide a conceptual framework to interpret and assimilate different experimental and theoretical work. It still remains to be explained how it relates to the underlying biological organization, under which conditions it could have appeared on the evolution of adaptive behaviour and how it fits with some of the more encompassing and currently available models of brain dynamics. The conjecture model just sketched will unfold on some of its more revealing details as we go into three key areas of research in depth: the evolution of cognition, cognitive developmental neuroscience and large-scale models of brain dynamics.

3. THE EVOLUTIONARY ORIGINS OF MENTAL LIFE: ENVIRONMENT, BEHAVIOUR, BRAIN AND BODY

We could expect Mental Life to appear, somewhere higher on the morpho-phylogeny of agency than where we last left it, within the evolutionary pathways of increasingly complex adaptive behaviour. From a rough analysis of our conjecture conceptual model of Mental Life we can abstract three main requirements of the mode of complexity that Mental Life requires: a) diversity of multiple neurodynamic structures, b) integration of this diversity on a unified whole, and c) plasticity as the capacity to create and modify such structures, particularly in an organizational context of equilibration.

How likely is the type of complexity to evolve? Which are the evolutionary pathways that lead to it? Is there any empirical evidence for that? What kind of bodyplan transitions will require and make possible such an evolutionary pathway? How will life be transformed in the process? The evolution of complex behaviour is the result of multiple organisms-environment feedbacks that occur at multiple scales: genetic, developmental, behavioural, social and ecological. The entanglement between all these factors is crucial to understand some of the basic features of the evolutionary process and treating them in isolation risks delivering misleading conceptions about causal priorities and driving forces in evolution. An alternative is to choose some of the most significant areas of interaction and analyse them with caution in order to distil a number of loose but valuable principles to understand some of the basic features of the evolutionary process that lead to Mental Life.

I have chosen to divide the picture in four main levels of evolutionary factors (see figure 23): a) the environmental level, b) the behavioural level, c) the level of behaviour generating mechanisms (brains) and d) the underlying infrastructural aspects (body organization, genetics, etc.). Although it might be possible, to a limited extent, to draw necessary conditions at each level (or rather at each interaction space between levels), probably none of these con-



ditions, taken in isolation, would be sufficient to sustain an evolutionary process that leads to Mental Life. I could not pretend to settle such a difficult question here. But throughout this section I shall focus on the interaction between these levels and try to extract principles that may help understand some features of the evolution towards Mental Life, its plausibility and implications. The first subsection will touch upon general considerations of the behaviour-environment relationship in the evolution of complex behaviour (interaction space 1 in figure 23). The second will provide some evidence for the evolution of brain architectures able to generate such behaviours (interaction space 2 in figure 23) and finally, I will abstract some consequences and requirements that stir from the coevolution of embrained bodies and embodied brains (interaction space 3 in figure 23). In turn, I shall try to nest all three blocks at each transition and some cross-considerations will slip into each subsection rendering a more integral picture. (I shall, however, leave some important developmental aspects for the next section.)

3.1. How evolution might open up a complex behavioural and environmental space for Mental Life

Although it may seem at first sight that increasingly complex behavioural or agential capacities should provide a considerable adaptive advantage (and subsequent evolutionary stability) it is not at all clear that it does. It is a fact that evolution does not always (not even that often) lead to increasingly com-

plex behaviour. On what concerns plasticity as an important aspect of behavioural complexity, for instance, within the lineage of *Opisthobranchia* the latter species have been shown to have lost the capacity of sentization on a withdrawal reflex due to a loss of serotonin modulation of sensory neurons (Wright 2000). This loss is correlated with inhabiting niches where the risk for predation is very low. Obviously it is difficult, if not impossible, to know a priori whether *Phylaplyia* and *Doladrifera* latter species first lost their capacity for sentization and only survived those that migrated to safer niches, or whether, on the contrary they first moved to those niches and then lost their sentization. The fact is that, at least in one (empirically well documented) case, evolution did not lead to an increase in complexity. Trivial and expectable as it may be, the path toward higher levels of behavioural complexity is not inevitable.

Even if a single case suffices to make the point, there are some important evolutionary reasons behind this lack of inevitability. On the one hand, behavioural plasticity is very expensive in terms of the required organismic resources and it will tend to be drawn off when possible. In Jerison's words:

"Adaptation to one's niche can be accomplished in many ways, and to adapt behaviorally by brain enlargement is expensive energetically. The brain is profligate in its use of energy, and almost any other solution to an adaptational problem is less costly" (Jerison 2002:267).

But it is not just a question of energy cost. Some kind of behavioural adaptive economy must also be considered. Various simulation models have shown that even when behavioural phenotypes are plastic, if specific environmental cues are stable and reliable to achieve adapted responses, plasticity will tend to get lost and canalized on quick and effective reactions (if developmentally available) as predicted by the Baldwin effect (Baldwin 1896, Hinton & Nowland 1987, Ackley & Littman 1992). Finally, learning involves the risk of potentially catastrophic mistakes and, additionally, a reliable phylogenetic pathway must exist that leads to mechanisms capable, in ontogenetic time, of creating new (or increasingly rich) and adaptively significant sensorimotor correlations.

Thus, Godfrey-Smith (1996, 2002) concludes that in general, when considering an increase in behavioural complexity as a means for higher adaptivity, other strategies will likely be favoured: e.g. generalist dump strategies (like building a shell) or cover-all behavioural responses (such as hiding or escaping under most kinds of unexpected perturbations). So, one may ask, which are the evolutionary opportunities that lead to an increase and conservation in behavioural complexification? Godfrey-Smith has defended that cognition (loosely understood as behavioural plasticity¹⁰) will evolve as a "response" to

¹⁰ Anil Seth (2002) has drawn attention to the importance of clarifying the distinction between the terms behaviour, mechanisms and cognition. usage of the term cognition by Godfrey-Smith ambiguously covering both behavioural and mechanistic aspects depending on the context. When explicit reference is given for the content of the term "cognition" so central for the thesis, it oscillated between opposing

types of environmental complexity (heterogeneity) that are relevant for the organism, if: a) there are environmental cues that permit a reliable detection of states of the environment to form adaptive responses, b) the chances that a mismatch between those responses and the state of the environment are not catastrophic¹¹, and c) pay-offs for flexibility are smaller than the advantages gained by sustaining flexible sensorimotor correlations. Given the appropriate conditions, then, behaviourally flexible strategies may be favoured; otherwise buffering mechanisms will most probably be selected (e.g. a shell, increase in size, etc.)¹². Sterelny (2001) adds a number of additional requirements. He notes that there are degrees of behavioural flexibility and that the important point is to distinguish which kind of flexible behaviour will evolve with which kind of environmental properties. One may think of the most basic type of “flexible” behaviour as what Sterelny calls an “associationist engine” able to assign different responses to different stimuli on the basis of their measured adaptive reward (what we have conceptualized as highly constrained adaptive behaviour). But this type of cue-specific sensorimotor coupling, he argues, is fragile (cue-bound sensorimotor couplings are deemed to failure in noisy and variable environments) and not sufficient for genuine cognitive capacities¹³.

poles, from a fragmentary “tool-kit” (including memory, decision making, representational equipment, learning capabilities, etc.) to the unified property of behavioural plasticity and integration. This conceptual crossing between different levels of description and modelling (behavioural, mechanistic and functional-computational) is very common along the literature of cognitive and brain evolution. Although I have tried to be as clear as possible on distinguishing them it was not on my hands to assure the distinction between levels on the quotes from other authors.

- 11 As Sterelny puts it: “[T]hough it’s necessary that doing the right thing at the right time is better than doing the same thing all the time, that is not sufficient for plasticity to be of value. Information will never be perfect, so trying to do the right thing at the right time risks doing the wrong thing.” (Sterelny 2001: 182)
- 12 To avoid potential misunderstandings it is convenient to note that Godfrey-Smith is not here arguing for *sufficient* conditions but for *necessary* ones. All this conditions may be met and yet no evolution occur in the direction of increase on behavioural complexity.
- 13 In a somewhat parallel quest to ours, what Sterelny is trying to achieve on *The evolution of agency and other essays* is a clear cut distinction between two types of systems. On the one hand he finds “associationist engines”, “skinnerian” or “nike” organisms (they *just do it*), understood as adaptive agents whose behavioural capacities are limited to “associated” sensorimotor correlations to adaptive rewards and appear bound to the demands of biology. On the other end (the final destination) are genuine “cognitive systems” able to detach their behavioural organization from a purely adaptive economy and enter into the domain of reasons (beyond immediate adaptive reward and instinctive motivation). An essential aspect of this task, to which Sterelny devotes most of his efforts, is to provide a “good” definition of “representations”, good enough to leave adaptive behaviour out, so that cognitive systems can stand in peace with their own distinctive mark (the use of representations) saving, altogether, representationalist philosophy of the mind from the uneasy situation of seen its much appreciated and exclusive object of enquiry spilled down to the level of bacteria. (The end of the plot, for those interested, is that genuine representation users will not appear until a complex social domain is in place, somewhere in the phylogeny of primates, where behavioural requirements include strategic deliberation, delusion, etc. i.e. a theory of mind.) Interestingly, Sterelny’s strategy involves the division of agency into the domain of behaviourists and cognitivists without conceiving a framework (e.g. dynamicism) able to provide a more “natural” and gradual hill-climbing strategy capable of integrating more diverse and richer aspects of evolution (development, body, neural organization, etc.)

The crucial type of behavioural flexibility that increasingly complex forms of cognition require is *robust* flexibility. Sterelny equates this robustness with a notion of representation that is not just a response to specific environmental cues but capable of robust detection and tracking of distant functional features of the environment through different channels and conditions¹⁴. The main “reason” for evolving robust flexible behaviour is a *translucent* heterogeneous environment where a variety of proximal cues can relate to the same distal functional (adaptively relevant) aspects of the environment (with different degrees of reliability) and, conversely, the same proximal cue can relate to different functional aspects. In other words: the mapping or correlation between current sensorimotor conditions and distant adaptive relevance is variable. In such environments the organism’s behaviour requires a wide perceptive and active capacity to engage on the appropriate sensorimotor coupling (and not just in the single-cue-single-response fashion, characteristic of a reflex-like architecture). Sterelny argues that *translucent* environments will be typical of: a) organisms that inhabit *different niches* (ecological generalists) that will be less likely to rely on specific cues; b) environments that “may be variable over short periods in ways that generate no simple and reliable physical signals” (p.266); and c) the hostile world of predators, preys, parasites and competitors that will try to exploit specific stimuli for their own advantage and make errors more costly, driving a co-evolutionary positive feedback effect for robust and complex perception-action cycles¹⁵. We could expand these considerations in two different directions (not considered by Sterelny). The first is that not only robustness (as a perceptual category¹⁶) will be re-

than what the rigid externalist requirements of representation and reason can encompass. The problem, no doubt, is that the dynamical approach seems to provide poor scaffoldings to achieve a properly cognitive normativity or autonomy of cognition that could escape from the tyranny of biology. It is here where the notion of Mental Life may offer an attractive middle ground, despite its difficulty for providing thick bricks to separate the departments of biology, psychology and phenomenology. All this considerations however do not render Sterelny’s picture a sterile land for us since we can take advantage of some of his distinctions and most generic conclusions without a strong commitment to his methodological and ontological bet on representations.

- 14 We can however adhere to his notion of robustness without an explicit commitment to a representational language, after all, without the inclusion of further requirements, this type of robustness involves nothing but a sensorimotor scheme, or a combination of them, able to assimilate a wide range of environmental variations within it.
- 15 “Predation is not just a danger to life and limb; predation results in epistemic pollution. Prey, too, pollute the epistemic environment of their predators. Hiding, camouflage, and mimicry all complicate an animal’s epistemic problems.” (Sterelny 2001: 247). It seems, at first sight, that one could override this argument appealing to the Red Queen hypothesis: predator-prey strategies will continuously catch up one another moving on circles of phenotypic-behavioural variability without accumulation of increasingly complex forms. However, under an ecological network it would be much more difficult for changes to synchronize and, one could argue, behavioural flexibility will be favoured over rigid behavioural strategies.
- 16 Sterelny splits robustness into the poles of robust perception and a wide “menu” of behavioural options.

quired but also integration of different sensorimotor correlations to lead to different concatenation of behavioural schemas to “navigate” deep/translucent heterogeneous adaptive environments. This will be required if correlations are deep in the sense that the adaptive reward requires to extract a sequence of correlations (e.g. finding the burrow of a prey through a searching process). Second, that *cooperation* (not just hostility) is also an important source of sophistication of behavioural contexts: matting behavioural coordination driven by sexual selection (such as bird-singing), child-parent dependencies which are specially relevant for behavioural plasticity (in terms of behavioural development) or inter-species cooperation (such as the coevolution of the visual system of bees, benefiting for nectar, and the shape and colour of plants, benefiting from cross-pollination).

However, resolving evolutionary questions as problems posited by external conditions is a controversial strategy (see Lewontin 2000 for a recent review), risking a number of speculative misfortunes. Environmental niches do not pre-exist to species, waiting to be occupied, selectively pressuring organisms to adapt to them. Converging with the line of enquiry we have followed till now (in terms of agent-environment co-determination) Sterelny himself acknowledges that:

“The *prima facie* problem for externalism is that both external and internal changes are necessary for an evolutionary response to complexity, but neither are sufficient. (...) Internal factors are structuring causes. (...) Internal features help determine whether the environment is relevantly heterogeneous, and whether there is a right option in response to that heterogeneity. (...) So the array of internal factors that are necessary for the evolution of adaptive flexibility are not mere general “fuels for evolution”. They are specific, not general causes of evolutionary response. They are causes of structure, not just generalise support for evolutionary change. (...) If properties of the organism make some aspects of the environment causally relevant in explaining the organic system’s response, we cannot predict evolutionary response just from our knowledge of the environment and a few general features of the organic system. We need to understand the organic system in enough detail to understand which features of the environment will be relevant to it.” (Sterelny 2001:184—185)

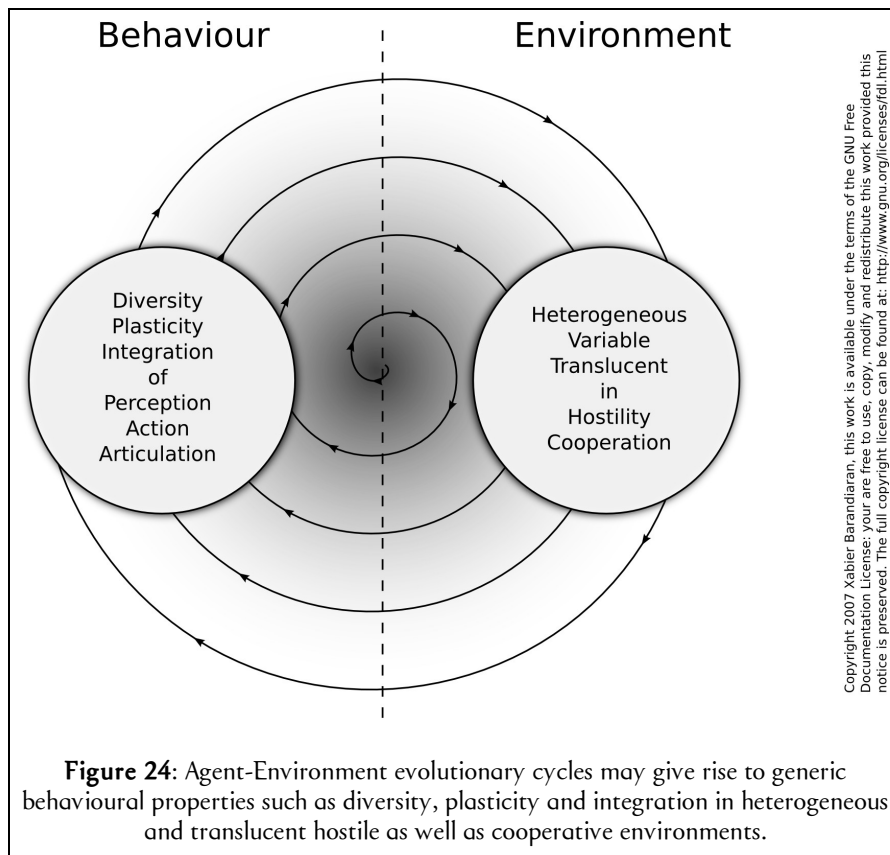
Should we give up on specifying behaviour-environment principles (however loose or generalist) for this reason? An alternative would be to skip altogether this level of analysis and focus exclusively on the kind of bodily transformations that would make possible the appearance of more complex behaviour generating mechanisms. This strategy will, unfortunately, miss an important point: the need of some type of selective pressure for these transformations to proliferate, for otherwise we will be putting the appearance of Mental Life into the hands of pure chance (not at the level of generation of variation but at the level of the retention of this variation). We know that more complex behaviour will not be selected inevitably, even less if accompanied by more energy and resource consuming mechanisms and bodies. Therefore, there is a

central explanatory role for principles that favour an increase in complexity *and* principles by which phylogenetic paths get trapped into system-environment behavioural relationships (not just environmental pressures) that do not permit a move back into previous or less complex modes of adaptive behaviour.

We can, therefore, sum-up and briefly elaborate onto some of Godfrey-Smith's and Sterelny's best guesses (adding some additional insights from Christensen—2007) to distinguish some of these principles. The first is straight forward. An increase in the *diversity* of sensorimotor correlations available for the organisms will be most likely favoured since it would open-up wider environments and evolutionary “routes” (provided that the ecological context provides sufficient diversity, something that we could take for granted once life has started to proliferate and diversify). Second, *plasticity* will be favoured if pay-offs are not too high and the environment is considerably variable but reliable and gradual on the correlations required to adaptively stabilize new sensorimotor correlations. Third, *robustness and integration* will be favoured when environments are heterogeneously rich, translucent and deep on their correlations (proximal cues are variously correlated with distant adaptive opportunities/dangers). Finally, there are important cross interactions between these factors. On the one hand, increase in behavioural diversity will create a tension to achieve global coherency, leading to an integration pressure (Christensen 2007): the more diverse and robust the repertoire of sensorimotor strategies the harder its global coherent concatenation will become, but also the more adaptively advantageous to achieve it. On the other hand, plasticity will make available diversity wider and can also potentially facilitate integration.

The traps come from ecological contexts that are both cause and effect of the above principles. These contexts must be relevant for organisms in highly-rewarding and cooperative manner requiring increasing coordination so that lost of achieved complexity will imply an almost direct lost in adaptation (e.g. parent-child or sexual matting dynamics) and/or in a hostile and bad-punishing form, creating a complex environmental context (in terms of ecological behavioural relationships) that the organisms can simply not ignore (e.g. predator-prey dynamics)¹⁷.

¹⁷ Hostility, this is the main point that Sterelny makes, is not like the metabolic “choice” or co-determination between system and environment (e.g. sucrose as a nutrient), escaping from predators is simply not a choice: an hostile environment with predators and competitors will actively seek your destruction.



We can recapitulate by saying that it is reasonable to expect, given the appropriate conditions, a positive feedback effect between the two poles of the behaviour-environment dialectic relationship that converges on an evolutionary tendency toward increasingly complex behavioural organization (see figure 24):

- *On the behavioural pole:* a) diversity (multiple specialization), b) plasticity (capacity for change at different time-scales) and c) integration (coherent coordination) at three different levels of behaviour: i) perception (robust tracking), ii) action (contextually-dependent behavioural alternatives) and iii) articulation (structured concatenation of perception-action cycles).
- *On the environmental pole:* heterogeneous, variable and translucent in correlations, at different time-scales in both highly cooperative-rewarding and competitive-hostile manner.

A recent conceptual simulation model of the evolution of controllers for situated competing agents gave rise to a similar conclusion that provides important insights to move into the next subsection (i.e. relating behavioural

complexity with brain organization). Using a platform called *Polyworld*¹⁸, Yaeger and Sporns (2006) simulated evolution in an artificial ecology populated by agents consisting in neural network controllers with Hebbian plasticity. Agent's genomes encoded phenotypes on a probabilistic generative manner, with the number of nodes, plasticity parameters, and connectivity architectures subject to evolutionary change. The environment was feed with different food sources but agents could also kill and eat each other allowing for hostility to co-evolve. The results reveal an increase through evolutionary time of four significant magnitudes on the controllers: plasticity, entropy, mutual information and complexity. Entropy is a measure of information and loss of mutual information is a measure of functional specialization; both combined are interpreted as functional diversity. Results showed that integration raised up quickly on the simulation and was latter maintained with minor fluctuations leading to an overall increase on complexity (understood as a compromise between functional diversity and integration). Yaeger and Sporns conclude:

Driven by a broadly defined fitness function that promotes a variety of behavioral traits and increased population density, neural systems in *Polyworld* are exposed to progressively higher rates of sensory input and must process this information to generate coherent behaviors. These challenges are met by the emergence of more structurally elaborate and more plastic networks, whose activity exhibits more differentiated as well as integrated dynamics as measured by complexity. (Yaeger & Sporns 2006:335).

This simulation model should be taken as a limited, yet promising, proof of concept for a generalist trend in evolution towards higher plasticity, diversity and integration¹⁹.

3.2. Trends in neocorticalization: evidence for the evolution of complex behaviour generating mechanism

There is empirical evidence of phylogenetic trends leading to more intelligent behaviour, and not in terms of a diverse tool-kit for specific adaptive couplings (as many evolutionary psychologist suggest) but in terms of behavioural diversity, plasticity and integration as outlined above. In fact, intelligence,

¹⁸ <http://sourceforge.net/projects/polyworld>

¹⁹ One important remark is made by the authors and is deserves some attention: "This particular simulation environment can probably be thought of as a single ecological niche, in which we have now observed a growth in complexity up until such time as the niche is fully exploited—until the population has reached a maximum and the individuals' expressed behaviors fully satisfy the only demands placed on them. We speculate that a well-formed measure of complexity applied to a biological species first occupying a new niche may exhibit similar growth. Then, given that all niches are not created equal, and that in more complex environments agent behaviors may result in additional niche creation, it is not difficult to imagine the observed growth in complexity extending to multiple niches and ecologies as a whole." (Yaeger & Sporns 2006:335)

defined generically as flexible behaviour; “has apparently evolved independently in different classes of vertebrates (e.g. birds and mammals), and in different orders of the same class (e.g. cetaceans and primates), as well as in different families of the same order.” (Roth & Dicke 2005: 250). A rather extended agreement on the brain evolution literature is that neocorticalization is an important evolutionary trend, that it correlates fairly well with intelligence and that it may even drive to encephalization²⁰ more generally (another gross indicator of mechanisms capable of increasing behavioural complexity). According to Jerison (2002) earlier species tend to have less neocortex than later mammal species, and species of progressive orders (those that evolved more) have more neocortex than species from archaic orders. The “neostriatum” in birds and reptiles is analogous to the neocortex in mammals and it also shows an increase through evolution. In addition, “cortical volume increases faster than brain volume as a power function with an exponent of 1.13” (Roth & Dicke 2005: 253) which makes animals with a high degree of encephalization have proportionately larger neocortex.

The significance of neocorticalization is twofold. On the one hand, the neocortex (and anatomically and evolutionarily equivalent parts of non-mammalian animals) is associated precisely with the requirements of diversity, plasticity and integration we outlined above. In La Cerra and Bingham’s words:

An extensive literature underscores the enormous functional plasticity of the neocortex, a distinguishing characteristic of mammals. This evidence supports the position that cortical representational features are systematically constructed by the dynamic interaction between environmentally derived neural activity and intrinsic neural growth mechanisms. *The information-processing capacities of the neocortex are largely constructed by the problem domains confronting the individual throughout development, and remain modifiable throughout the life history.* (La Cerra & Bingham 1998: 11290, italics added)

On the other hand, the neocortex is, interestingly, the last brain region in developmental order which sheds light both on the development of cognitive capacities and its evolution. As Streidter reviews (2006), the most likely explanation for the 1.13 power scaling function of neocortex growth through evolution is Finlay & Darlington’s hypothesis (1995):

[T]hat evolution generally enlarges brains by prolonging brain development (letting all precursor cells divide more frequently) but conserves the “birth order” of the various brain regions. This birth order constraint would cause late-born regions, such as the neocortex, to become disproportionately large as absolute brain size goes up. (Streidter 2006: 4—5).

There is, therefore, a major developmentally available mechanism that the evolution of behaviour can exploit: heterochronic changes of brain growth (which are under heritable control). “These highly linked regularities in the

20 Used to term the increase in brain size compared to the body size.

development and evolution of mammalian brains”, Quartz concludes, “make evolutionary proposals that depend on region-specific, modular adaptations problematic. Rather, it suggests the evolution of robust, flexible developmental strategies in which the determination of mature function in the cortex depends heavily on epigenetic mechanisms, including progressive specialization by afferent input through constructive development.” (Quartz 1999).

In sum, relational (organism-environment), phylogenetic and developmental factors converge presumably on the fact that there is a set of evolutionary trends towards more diverse, plastic and integrated behavioural agency. On the side of organism-environment relationships, this type of behavioural agency was both required and enabled by ecological environments in which high-energy resources were available to agents through increasingly sophisticated sensorimotor correlations and ecological and social co-evolutionary traps through hostility and cooperation (that block or severely limit a return to path previous or less complex behavioural modes of life). On the phylogenetic side, there is evidence for the progressive increase (in different lineages) of brain areas associated with integration and developmental plasticity of behavioural agency.

This evidence, of course, does not force the conclusion that an increasing tendency towards behavioural complexification will *necessarily* occur (for a recent review of the controversial issue of “progress” in evolution see Rosslenbroich 2006). But there are important reasons (other than the facts of evolution in planet earth) to believe that the Gouldian left wall of the evolution of behavioural agency will be shifted to the right as evolution favours increasingly complex agency. Some are due to behaviour-environment relationships as we have previously seen. But the displacement of the left wall is also due to other organism-environment knots that preclude evolution in the opposite direction. One of them, as reviewed by Roth & Dicke (2005), concerns the diet. For instance, the encephalization required to achieve behavioural flexibility will permit a much richer diet but it will also require it. This is due, on the one hand, to the energy requirements of the brain²¹ and, on the other hand, to the hypothesis that the increase on brain energy usage may induce a reduction of other tissues, including the gut, thus rendering the dietary requirements even more complex. We thus enter into another evolutionary positive feedback effect: to opt for high-value energy benefits you need to invest more energy and the more energy you invest the higher your need to find difficult but highly rewarding energy sources. The diet, however, is but one of the set of the bodily organizational transformations of life that the evolution of behavioural agency brought with it. The organization of multicellularity

21 “The human brain occupies only 2% of body mass, but consumes about 20% of total metabolism” (Roth & Dicke 2005: 255).

(the agential body) will crucially be transformed in the process in ways that shall become equally crucial for Mental Life.

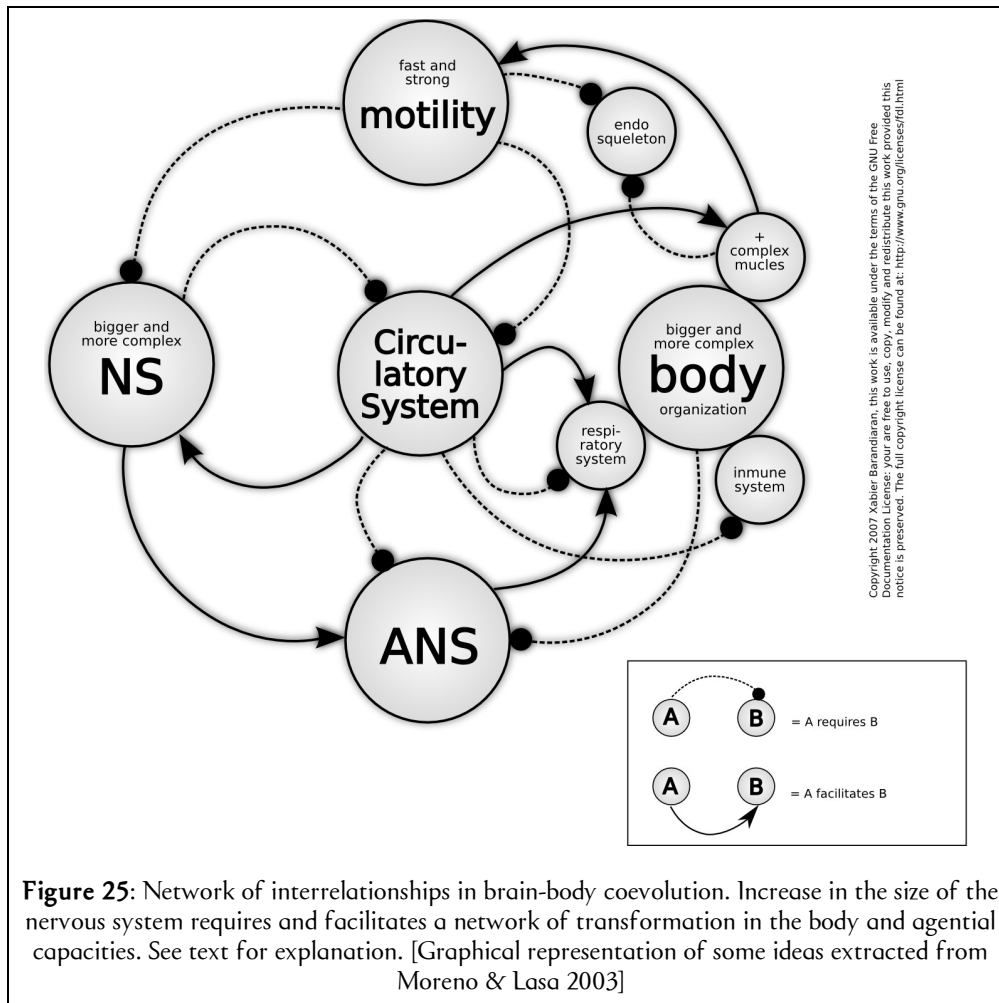
3.3. Encephalized bodyplans: co-evolution of embodied brains and embrained bodies

The evolutionary trend that lead to more complex behaviour had to surmount a number of organizational or bodily factors that impose difficulties onto the increase in brain size and fast body movements. We shall mainly follow Moreno & Lasa (2003) throughout the set of transformation that came to overcome these problems leading to the vertebrate bodyplan. The most immediate problem is that a significant increase in the volume of neurons in a specific area poses several infrastructural problems due to the need for adequately feeding and maintaining them. To solve this problem, encephalization requires a *closed circulatory system* finely regulating blood pressure and flow depending on different internal and environmental circumstances. In turn, the directional management of this system can only be achieved by a system of regulators distributed along the blood vessels, to control blood pressure, oxygen concentration, and acidity level (through the modulations of the rhythms and strength of the heart beating and contraction/dilatation of the walls of the vessels) able to manage the differential transportation of nutrients and the oxygen to tissues far away from the external surfaces of the animal. This also requires a more complex immune system, because the required circulatory system permits a quick spreading of pathogens along the whole body. In turn, for analogous reasons to those that required the appearance of NS, the coordination of these regulators cannot be conceived without direct control of a neural network, which requires even greater neural resources, that in turn require a more finely tuned circulatory system, and so on.

The progressive specialization of the NS on internal bioregulatory functions lead to a split between two major divisions of neural organization: a) the sensorimotor NS (or SM-NS hereafter) and b) the NS of the interior (the NS-I) comprised by the autonomic nervous system, the neuroendocrine system, etc. With a neural subsystem taking care of metabolic control the interaction between the NS and metabolism becomes more intimate, overcoming some severe limitations of earlier interaction systems mostly based on diffusion processes of hormones. In fact, the effect of this body-NS interaction from the point of view of biological organization is crucial since the NS will now subsume basic metabolic normative regulation. The ANS and the neuroendocrine system makes the NS an essential part of metabolic organization, to the extent that there is no possible life in vertebrates without neural control. The other way round is also important to consider. For the increase in the size of the NS to translate into more complex adaptive behaviour an increase

in the complexity of the coordination of bodily organs is required. Once again the NS-I resolves this problem acting as a neural subsystem able to regulate and coordinate the homeostatic dynamics of different organs in response to behavioural needs (accelerating blood circulation to muscles, halting less important functions, like digestion, etc.). The NS-I (particularly the autonomic NS) is not a mere reflex-type homeostatic control system but an allodynamic regulator (Berntson & Cacioppo 2000) constantly interacting with integrated neural processes and capable of flexibly and adaptively transforming metabolic-body dynamics to face existing or anticipated bioregulatory demands (both in terms of metabolic-constructive and behavioural requirements). This flexibility of the ANS will in turn become crucial for the interaction between the NS-I and the SM-NS with enormous consequences for neurodynamic organization (in terms of development and behaviour), as we shall see later. In turn, these two different functions of the NS (sensorimotor coordination and body regulation) would require, given a certain level of complexity, the formation of specialized structures that couple both subsystems for coordination: the hypothalamus and the limbic system (particularly the amygdala).

There are a whole set of network-like interdependent relationships of organizational requirements and possibilities that leads to the appearance of vertebrate encephalized bodyplans (see figure 25 for a simplified diagram). All this makes manifest that there is a close relationship between body and NS complexification. The potential of the NS to generate increasingly complex behaviour cannot be developed independently of changes in the general organization of the body and its development; only in very specific bodyplans does an open/sustained process of encephalization become possible. Body organizations and NS of animals co-evolve, imposing limits and opportunities to one another; i.e. co-determining each other on their capacity to generate complex behaviour and transforming the very organization of life in the process. All these changes in the internal organization of body-brain allows as a consequence the possibility of bigger animals able to perform fast, plastic and strong forms of movements because now there is available an efficient cardiorespiratory system controlled by the ANS capable of coordinating different organs in order to recruit and channel the necessary energy for kinetic-mechanic work. This, in turn, allows the creation of new ecological niches based on big size and efficient quick motility, from sophisticated predator strategies to migratory capacities.



Additional changes were also required at developmental and genetic scales. The appearance of the vertebrate bodyplan, with particular reference to the NS, depends upon increase on developmental degrees of freedom, something that Lenny Moss has directly attributed to the appearance of the neural crest: a set of cells that detach freely from the neural tube during development and give rise the central and autonomic nervous system, hormone producing cells and its cell derivatives leading to all bones and cartilages. In turn, as in every major transition, the material/genomic conditions of possibility for the neural crest are made possible thanks to the “conversational complexity” of the genome (its capacity to include and exploit a regulatory syntax) more than on a mere increase of genes or genome size (which do not relate well with organismic complexity, not to speak of brain or behavioural complexity).

The ability [within developmental contexts] to respond rapidly to contingent and transient features of an environment entails several characteristics. It entails a capacity for cells to step back or 'detach' from the manifold of available ambient stimuli and make salient, or privilege, some kinds over others, thereby already exercising an emergent level of autonomy. It entails a capacity to make nuanced distinctions between closely related signals and an ability to internalize these signals (and these distinctions) resulting in rapid and fine-tuned adjustments to the internal state of the respective cells. The kind of enterprise that vertebrate existence brought into being was one characterized by two-way dialogues between the sensory surface of the organism and its ambient surround, and between the sensory surface of the organism and its viscera, its interior. It is the advent of the neural crest that makes this transition to new levels of contingent responsiveness possible. (Moss 2006:932—934)

Other developmental aspects were also crucial. For instance, in invertebrates neural concentrations are just accumulated one over the other with less room for nourishment structures. Even with a closed, efficient circulatory system, a fine energetic maintenance of this kind of increasingly big neural concentrations would be difficult (Montalcini 1999). The increasing complexity of the NS in vertebrate evolution is also facilitated because it is embryogenically developed around the walls of a cylindrical cavity and is therefore favoured by nourishment from the inside, as well as from the outside limits. Other important transformations include the appearance of a whole set of support cells for the NS like myelinated cells that permit action potentials to travel faster and further (thus allowing the formation of bigger brains) or glial cells that feed neurons among other functions.²²

A whole set of evolutionary interactions between environmental factors, behavioural organization, brain architecture and bodily and developmental infrastructure come to explain the phylogenetic trends that lead to bigger, more plastic, diverse and integrated NSs. Somewhere in the vertebrate phylogenetic pathway, if not earlier, epigenetic factors were ready to give rise to Mental Life. The complexity involved required that adaptive behaviour be the result of a developmental process that exceeds the constraining capacity of genes, their interactions, developmental self-organization or environmental factors: Mental Life needs to be born and become through a behavioural developmental process.

4. NEUROCOGNITIVE DEVELOPMENT: BECOMING MENTAL LIFE

Mental Life would be precluded if the organization of behaviour would be completely or strongly constrained by innate or exogenous constraints (i.e. fully specified by a genetic program or developmentally channelled by pro-

²² Interestingly a subclass of glial cell, the astrocytes, are currently believed to have assumed neurodynamic functions as well (Volterra & Meldolesi 2005).

cesses outside the domain of behaviour). We have provided some evidence of an increasing evolutionary trend towards plastic and integrative centres in brain evolution, particularly those that develop late and leave room for post-natal and interactive organization. The goal of this section is to provide evidence at the developmental/ontogenetic scale for the growing of an organized bundle of neurodynamic structures through interactions (as required by our conjecture model for Mental Life)²³.

It is perhaps useful, at this point, to present the central thesis of one of the major approaches to cognitive organization for it almost systematically tends to leave aside developmental aspects, and defends nevertheless a picture of the mind in clear opposition to that of Mental Life. What has been called the Swiss-knife metaphor conceives cognitive organization as be a *toolkit* of problem specific modules that evolved under specific selective pressures. It can be exemplified by the following programmatic quote

The mind is what the brain does; specifically, the brain processes information, and thinking is a kind of computation. The mind is organized into modules or mental organs, each with a specialized design that makes it an expert in one area of interaction with the world. The module's basic logic is specified by our genetic program. Their operation was shaped by natural selection to solve problems of the hunting and gathering life led by our ancestors in most of our evolutionary history. The various problems for our ancestors were subtasks of one big problem for their genes, maximizing the number of copies that made it into the next generation. (Pinker 1997:21).

As we shall see, current trends in neurocognitive and psychological development have challenged this view providing insights into the cognitive and behavioural organization of the brain that come to support our conceptual model of Mental Life.

4.1. Some general genetic and developmental considerations

A bottleneck certainly exists on how much of the brains circuitry can be genetically specified. As Elman et al. (1996) have noted, in humans (and vertebrates more generally), only global architectural and chronotopic constraints participate on the development of the SM-NS. Chronotopic constraints affect the timing of certain developmental processes whereas global architectural constraints specify gross neural pathways, types of synaptic properties, types of neurons, etc. But none of these constraints can specify the circuitry that produces functional behaviour in adult brains. The gross numbers speak for

23 Unfortunately, most of the studies that emphasize the open, plastic and interactive character of behavioural development are focused on humans, motivated by its strong philosophical implications (nature/nurture dilemma, the nature of human language, its difference with other species, etc.). Inevitably, part of the literature on this topic will be drawn from studies of human cognitive development but most of the conclusions that I will extract can be extended to vertebrates in general and particularly to mammals and birds.

themselves: in the human case, if neural circuits were genetically pre-specified, this would require that 10^6 genes (of which only 20-30% are expressed during the construction of the NS) would have to store enough information to codify for the 10^{11} neuron's properties and those of up to 10^{14} synapses, which, in turn, may take on quantitative values from a fairly wide-ranging spectrum. Moreover there is increasing evidence for the generalist character of the genes participating in the development of the NS (Kovas & Plomin 2006): the same genes affect the development of most cognitive abilities (they are pleiotropic and polygenic) and there is little (if any) direct mapping from differentiated genes to differentiated anatomical divisions of the brain (not to speak of local connectivity patterns). The reduction in variability, from the set of possible parameters and connections of the NS to those which are functional for the organism, forces a developmental and interactive selection, construction and stabilization of neural organization.

However, this type of genetically loosely constrained view of development is not a specific feature of the brain. Complex organisms interactively construct the context and content of subsequent stages of development. Evolution does not shape or select genes or genetic programs but shapes developmental processes in which a distributed number of resources (genetic, cytoplasmatic, environmental, etc.) interact at different scales configuring the developmental processes (Gilbert 2003). Although common to most developmental processes, the significance of this phenomenon does not rest importance for the understanding of how cognitive capacities are generated and shaped by evolution, particularly in opposition to some influential strong nativist conceptions (Griffiths & Stotz 2000). What is probably more characteristic and distinctive of brain development is the extent to which it relies on the available plasticity, its dependency on endogenous neural activity, the duration of the developmental process and, particularly as we shall see, the fact that it is strongly behaviour driven: from prenatal stages, sensorimotor coordination generates the regularities and organization that leads to subsequent developmental stages incorporating environmental-contextual features that are, in turn, the result of its own activity.

4.2. Embryonic (prenatal) brain-body development

In the case of vertebrate embryogenesis, as it is well known, during the third week after fertilization, the epithelium folds into itself to form the neural tube, the precursor of the NS. (There is a beautiful recapitulation here between the evolutionary origin of the NS as epithelial conduction and its ontogenetic formation as an internalized folding of the epithelium.) Soon after cells detach from the neural tube to form a "free-floating" ensemble of neurons called the neural crest. According to Striedter (2003), two major modes of development govern the formation of the brain as an organ. The first is the

compartmentalization mode, in which cells self-aggregate by affinity into developmental compartments called neuromeres. By the time neurons start to be functionally connected compartmentalization is almost complete forming progressively smaller adult cell groups. This mode of development is highly conservative through evolution and it defines certain gross anatomical structures, particularly the hindbrain. The second form of development Striedter calls the *dynamic network mode* is “characterized by axon-mediated developmental interactions and involves both trophic and activity-dependent mechanisms (...) [it] provides for functional integration of disparate brain regions and, thereby, promotes evolutionary changes in brain organization”. During the dynamic network mode, compartments are rearranged, interconnected and cells migrate and establish large-scale connections. At an early stage, axon growth and innervation (independently of action potential and synaptic activity) lead to a kind of compensatory game in which groups of cells tend to match connections to other groups and settle down into some equilibrium of afferent and efferent proportion to target groups. At a latter stage, activity and sensorimotor dependent changes start to selectively sculpt this gross anatomy and grow new connections.

During this dynamic network mode of development, still at the embryonic stage, body and NS start shape each other through movement. As Müller (2003) and Robinson (2005) review, there is evidence for the role of embryonic motility in morphological and body development which turns to be particularly relevant for the embodied conception of cognitive capacities. Movement dependent mechanical stimulation influences the formation of skeletal structures and joints and, in turn, environmental factors (temperature, light intensity etc.) affect the rate of embryonic movements: a sensorimotor system is already in place shaping and enabling the formation of its own body. Inhibition of such movements has severe effects on different bones and cartilages:

“Less widely appreciated are the teratogen-like effects that result from a simple absence of embryonic motor activity. Foetuses that experience a period of akinesia (loss of movement through drug exposure or myopathy) exhibit a suite of morphological effects, including microstomia (small mouth), retarded lung development, skin and facial abnormalities, immobilized joints and altered bone growth, short umbilical cord, and long-term movement disabilities. Fetal Akinesia Deformation Sequence is a stark demonstration that fetal movement is an important contributor to prenatal morphological and behavioral development.” (Robinson 2005:470)

These are thus important cases of influence of the NS activity driving motility driving body formation. But there are also crucial influences on the other direction: the myogenic activity of the muscles (their synchronized electrical spontaneous movements) guides their innervation and organizes the early neurodynamic activity of the spinal cord canalizing the formation of loc-

al rhythms that shall become essential for the control of behaviour (particularly to create motor synergies and channel the enormous amount of degrees of freedom and muscular coordination that behaviour requires):

“Thus, there is a close impedance matching (having the same dynamic properties) of neurogenic movement to the properties of the muscle. The upshot of this is that the external properties of the animal have begun to be internalized in the brain. The motoneurons stay electrotonically coupled until the upper part of the system, the brain stem (also at this point electrotonically coupled), starts making its synaptic connections with the motor neurons. At that point, the motor neurons become electrotonically decoupled, but the upper part of the system remains coupled. In addition to becoming electrotonically decoupled at this stage of development, motor neurons also begin to receive synaptic inputs from other parts of the nervous system that do not specifically relate to the activation of given muscle groups. These additional inputs relate more to the global movement of the total mass of the animal, and involve the vestibular system, the organ of equilibrium that informs the motor neuronal network (thus the musculoskeletal system) about holistic properties of motricity.” (Llinás 2001: 60).

Thus, early embryonic movements provide the first steps for a functional coupling between neural and body dynamics. Neurodynamic structures start to form matching global motile relationships integrating limb distribution, frictional and joint synergies, and gross environmental invariances such as gravity. But the embryo and foetus are severely limited on their exploration of the world before birth, the maternal environment soon reaches a bottleneck on its capacity to afford complex sensorimotor experiences. Some sensorimotor structures are ready at birth (particularly those of subcortical areas) but great part of the organization of the brain (and behaviour) will have to wait for postnatal development.

4.3. Embodied and situated behavioural development

Piaget’s theory of cognitive development through an early stage of sensorimotor exploration stands as a major reference framework for cognitive development. The cognitive organization of the newborn, far from waiting for environmental triggers to unlock innate structures or undergo a passive maturational process, bootstraps itself through self-sustained sensorimotor interactions. Early cognate schemas provide the basis for an increasingly complex process of assimilation of new sensorimotor correlations and equilibration of the internal cognitive organization in a positive feed-back loop that subsequently creates the organization required for the next developmental stage and transforms itself:

“At multiple levels of analysis at multiple time-scales, many components open to influence from the external world interact and in so doing yield coherent higher-order behavioural forms that then feedback on the system, and change that system.” (Smith & Thelen 2003:347).

Dynamical embodied perspectives on human development (Thelen & Smith 1994, 2003) are becoming increasingly widespread in developmental psychology focusing on the way in which the very physical-morphological embodiment of the sensorimotor system becomes a crucial aspect of developmental dynamics. From the very beginning the body is in continuous movement, constrained by the physical parameters of body morphology, generating the very experiences (change of visual field, grasping, sucking, etc.) that permit to extract relevant sensorimotor correlations, that in turn will allow the organism to master the sensorimotor skills that will in turn bring about new experiences, etc.

Pioneering work by Held and Hein (1963) on cat's visuo-motor development showed that paralysed kittens could not develop vision. Furthermore, kittens paralysed but attached on a trolley to other kitten (and thus subject to the same visual input as the other, non-paralysed, kittens) also failed to develop visuo-motor coordination. These experiments suggest that the cognitive development depends critically on the capacity to interact "freely" with the environment. Therefore, it is not about getting the right input from the environment to trigger or activated maturational processes, it is the very context of action, the sensorimotor involvement and the accommodation and internalre-organization of habits and sensorimotor skills what becomes central for development.

Particularly revealing are the results of almost two decades of work on neuroconstructivist development by Karmiloff-Smith (Karmiloff-Smith 2005, Karmiloff-Smith & Thomas 2005). She has paid special attention to William syndrome caused by the lack of some 28 contiguous genes on one copy of chromosome 7. This syndrome is claimed to provide empirical evidence for the genetically pre-specified and functionally separated modules on the basis of a very specific cognitive effects of this syndrome: impairment of spatial and numerical cognition and very low IQ but high proficiency with language and face processing. Nevertheless, Karmiloff-Smith has shown that these symptoms are accompanied by subtle effects and impairments distributed along very different cognitive functions (not just isolated into the malformation of a couple of modules). In turn, this distributed effect can be explained by developmental cascading abnormalities (e.g. inability for referential pointing and very late language acquisition). Interestingly, these abnormalities were shown to be associated with early sensorimotor deficits such as atypical eye movements and difficulties on visual tracking or on the timing of certain body coordination. These sensorimotor deficits created further difficulties to develop higher cognitive functions or skills under "normal" social and behavioural contexts. Karmiloff-Smith's work is particularly revealing for it shows how adult brain specialization depends on early sensorimotor abilities even for higher cognitive functions such as spatial and numerical abilities.

4.4. Neural constructivism

Some of the neural details of the Piagetian framework to psychological development are slowly becoming available to empirical scrutiny (Quartz 1999, Johnson 2001). Developmental cognitive neuroscience (Johnson 1997, Munakata et al. 2004) is a growing field trying to map the increasingly complex organization of the brain and behaviour through development. Evidence from plastic reorganization of functional sensory and motor mapping in the brain during development is nowadays huge. Projections from sensory areas (visual, tactile or auditive) to the sensory cortex are highly experience dependent (even during adulthood). For instance, partial deprivation of senses (e.g. amputation of facial vibrissae in young mice and rats or blindness in humans) has been shown to produce re-routing of axonal projections of other sensory modalities to those areas of the sensory cortex “normally” associated with the deprived sensory source. But the challenge lies on understanding the mechanisms that underlay full sensorimotor skills and higher cognitive functions.

Two major paradigms compete to provide a large-scale theory of developmental brain organization: *neuroconstructivism* (Quartz and Sejnowski 1997, Quartz 1999) and *neural Darwinism* (Edelman 1987). Whereas the first highlights the experience dependent growth of neural structures through development, the second puts the emphasis on neural degeneracy and selection of connectivity patterns (after synaptogenesis—stages of overproduction of cortical structures). But from a more abstract perspective, both can be interpreted as the two sides of the same coin. Whether neurocognitive development relies on pruning and selection along redundant and unspecific neural structures or it works by directing dendritic growth and strengthening is a secondary issue for our framework. In fact, there is evidence for both types of processes during development and they are both complementary. What is central to both approaches is that the neural architecture of newborns is highly unspecified (underspecified and/or overspecified) and that empirical evidence shows that there is a considerable amount of behaviourally dependent specification of connectivity patterns through development (particularly in cortical regions).

How could this remarkable growth and delicate tuning of neural structures through behaviour be accomplished (particularly if heritable constraints are loose)? An often invoked metaphor is that the process can be understood in analogy with geological formation of rivers where, under many available routes towards the sea, it is the flow of water itself that creates and reinforces progressively the river bend. Small initial differences (like the presence of a rock) can be amplified and result in progressively stronger river bends. In the long run the river can create sophisticated routes like meanders. Similarly neurodynamic structures are created/selected through experience by self-organization. Yet, it rests to be explained how does this neural development

achieve the remarkable behavioural and adaptive capacities that it does for, in this case, we do not confront a river that arbitrarily shapes any possible route but adaptively and coherently tuned ones (as if the river were to draw precisely one among the possible, yet intricate, picture on the landscape required for perfectly watering the whole valley). The end result needs to be a coherent adaptive agent, not any kind of self-organized structure or river bend. Some kind of guidance is required to reach it.

Three main factors may be considered essential to channel and guide the adaptive and cognitive formation of neurodynamic structures (although specific details are still largely unknown). First, body morphology and the innate architecture of some subcortical circuits. Whereas cortical (higher level) structures remain unspecific, the brain-stem and hypothalamic circuits are highly constrained at birth as illustrated by their early myelination (from birth) whereas cortical areas myelinate much later (up to adolescence in humans), this being a feature that spans across all vertebrates (Gibson 1991). Second, the strong modulatory capacity of the NS-I (in charge of internal bioregulatory functions and able to strength unspecified cortical neural pathways) together with activity dependent (self-organized) plasticity mechanisms (first proposed by Hebb—1949). The third factor is behaviour itself and the structure of the environment, social and physical, in which it takes place (for a recent review of the effect of enriched environment on cognitive development in rats see Nithianantharajah & Hannan 2006).

Early sensorimotor schemas (like sucking or moving limbs) specified by subcortical circuits constrained by body morphology and coupled to the environment trigger the first sensorimotor experiences. As the newborn starts to meet new situations the consequences of its behaviour are “emotionally evaluated” by the NS-I that releases synaptic modulators to prune or strengthen, in an activity dependent manner, those (still loosely specified) pathways that may have participated on the new behavioural sequence (Edelman 1987, Damasio 1994, Damasio 2001). Edelman has called this mechanisms of reinforcement “value system” to highlight the mode in which sensorimotor interactions acquire a value or significance through their embodiment on a set of modulatory constraints that channel development and learning towards adaptive behaviour. The more a given pathway is “positively evaluated” the more it crystallizes, the more it crystallizes the more it is used, just like a habit. This process is not only restricted to specific connections or alternative switches (like the case of bacterial odour preference learning in *C. elegans*) but spans across most of cortical organization, through subsequent stages of development. Different stages of development are marked by overproduction of synapses (probably timed genetic chronotopic constraints) and become subject to shaping through the behaviour generated by previously stabilized structures. Self-organized plastic changes can also occur without need for

evaluation (or error-correction), like sensory mappings by means of competitive (winner-takes-all) learning mechanisms. As the sensorimotor organization becomes more complex, it will become capable of guiding its own development recruiting “value systems” for its own normative regulation (I will try to explain how this might be possible in the next section).

As a result, as shown by two independent comparative EGG studies in adults, infants and newborns (Meyer-Lindenberg 1996, Anokhin et al. 1996), the correlation dimension of human brain activity increases with development: i.e. new neurodynamic structures are created. Unlike having two eyes and one nose, the adult configuration of the cortex achieves a unique configuration for each individual (Lewis 2005) and highly organized in the sense the shape of every local configuration is precisely tuned with distant configurations to render sophisticated cognitive and behavioural capacities. To say it with Gerald Edelman, “each brain has uniquely marked in it the consequences of a developmental history and an experiential history” (Edelman 1999:71).

The result (always open to further transformation) is a highly specialized distribution of brain organization, but equally integrated. The existence of specialized locations in the brain (like functional maps of sensory and motor cortex) is an undeniable fact. But specialization is the result of a developmental process that starts early in the foetus and is highly sensitive to developmental behaviour. That a given area is highly activated when an animal observes triangles or navigates a maze does not mean that it be specialized on “triangle processing” or “spatial navigation” by a mere fact of correlation with an environmental feature that is otherwise inaccessible to the system. It is only within the contextual totality of neurodynamic organization, body and environment, finely orchestrated through development, that the activity of a certain area of the brain may be said to be specialized, as it is differentially integrated on a neurodynamic structure that couples to the triangle or the spatial context in an organizationally relevant way. How this integration occurs in “real time”, and how Mental Life is continuously actualized in the process, is the object of the next section.

5. MENTAL LIFE IN ACTION (OR 2001 A BRAIN ODYSSEY INTO THE COSMIC DANCE)

At some point during the 20th century it seemed clear that the turn of the 21st century was hiding a new keystone for human knowledge, somewhere in the confines of the universe. By the time when collective imagination had expected humanity to be exploring the mysteries of space, science was discovering a much closer and intimate, yet vastly fascinating, galaxy whose premonitory

vision was not advanced by Sir Arthur C. Clark but by Sir Charles S. Sherrington:

“In the great head end which has been mostly darkness spring up myriads of twinkling stationary lights and myriads of trains of moving lights of many different directions. (...) The brain is waking and with it the mind is returning. It is as if the Milky Way entered upon some cosmic dance. Swiftly the head-mass becomes an enchanted loom where millions of flashing shuttles weave a dissolving pattern though never an abiding one; a shifting harmony of subpatterns.” (Sherrington 1940: 177-178)²⁴

2001 was not the year of the space odyssey. Despite public promises (academic and literary alike), and 50 years of work on Artificial Intelligence, HALL²⁵ was far from ready to bring us beyond planet earth. Something in living minds was resisting the combinatorial power of computational AI. The cosmic dance of the neural Milky Way hold a secret that became progressively discovered on the computer readings of neurodynamic imaging techniques. 2001 became the year of neuroscience; far away from the lenses of the telescope of astronomers. Despite some pioneering early studies of mesoscopic brain dynamics (Freeman 1975) it was right at the beginning of the century that a major breakthrough of nonlinear neurodynamic accounts of large-scale brain activity was presented in mainstream journals (Edelman & Tononi 2000, Friston 2000, Freeman 2000, Varela et. al. 2001, Bressler and Kelso 2001, Llinás 2001 and Tsuda 2001, Engel et al. 2001). This section is drawn mainly from those studies and some latter developments.

We have sketched a conceptual model for the autonomous organization of behaviour as a web of situated neurodynamic structures. We have seen that evolution offers supportive evidence and theoretical consistency for trends toward a form of adaptive behaviour that makes room for plastic, diverse and integrated behaviour generating mechanisms, as required for Mental Life. We have then reviewed how this organization may be interactively constructed through development. It is time now to contrast our conceptual model with how real embodied brains generate complex behavioural tasks. The goal is not to provide a direct neurodynamic implementation of Mental Life as a fully operational model nor to state that current neurodynamic approaches necessarily come to support it. My aim is, rather, to suggest some possible connections and resonances between Mental Life and a simplified reconstruction of some large-scale approaches to brain functioning.

As our model of Mental Life suggests, the essential features of cognitive agency (identity and normativity) stir from global organizational properties. Therefore, it is at the scale of the integrated brain that we need to find evid-

24 Quote borrowed from Freeman (1997).

25 In Arthur C. Clark's book and Stanley Kubrick's celebrated film *2001 A space odyssey* HALL is the name of the computer system that governs the space-ship and leads humans to their next transition on the evolution of knowledge.

ence, if any, of the plausibility of our conjecture model. But the dynamic complexity of the vertebrate brain is overwhelming. Electric activity travels in just a hundredth of a second along thousands of millions of dendrites in many directions, affecting each other reciprocally through multiple feedbacks. The Cartesian decompositional method of breaking it down to isolated functional components and to put them back linearly together to reconstruct their collective functionality will not do for the purpose of understanding how brain dynamics are organized at the large-scale. If anywhere in the natural realm it is in brain dynamics where the presence of non-linearities (starting from the single neuron) and multiple feed-back loops (spanning across different temporal and spatial scales) is central to the generation of behaviour. Even the oscillatory dynamics of two reciprocally connected neurons challenges any decomposition, their behaviour is the result of their coupled intrinsic dynamics and completely transforms the behaviour of each neuron studied in isolation. Therefore some current mechanistic accounts of brain functioning take the form of systemic models that conceptualized brain activity as that of a massive network of oscillators whose collective activity yields to the formation of synchronized patterns of activity distributed over large areas of the nervous system and operating at different frequencies. The result is a many-to-many mapping between anatomy and function that is continuously being reconfigured.

Until very recently, however, there was no available picture of measurable neurodynamic activity able to inform an holistic model (both mathematical and measuring techniques were not mature enough). Investigations relating mind and brain were necessarily limited either to study the behaviour of isolated local circuits (reflex arcs, feature detectors in anaesthetised brains, etc.) or to anatomical lesion studies. The higher level observations of brain dynamics were limited to a descriptive status of observational statements (like Sherrington's quote above) or to the establishment of very abstract correlations whose causal status was difficult to assess. But the situation has changed recently:

"What allows us a fresh start now is our ability to image brain activity during normal behavior and to model our findings with the tools of nonlinear dynamics. However, these new data are being acquired under preconceptions embodied in old experimental designs, and we have to reinterpret them as they bring new concepts to light. It is hard for nonspecialists to grasp the elementary properties of neural activity, but it is even harder for specialists to unlearn old points of view to make way for new ones." (Freeman 2000: 12)

As Freeman suggests, the new data that functional brain imaging generates is, to some extent, being partially assimilated into the received view, particularly within those based on localizationist assumptions and on modular computationalist framework (not without its problems—Uttal 2001). But, to a large extent, a new paradigm based on non-linear dynamical systems (Haken 1977)

and network theory (Strogatz 2001) has proven to be a powerful and fruitful framework to give new sense to these data revealing important properties of brains which would otherwise remain out of reach. In particular it is proving to generate routes to discovery along a wide range of research areas: theoretical neuroanatomy (Sporns et al. 2000, Sporns et al. 2004), perceptual binding and attention (Varela et al. 2001, Herrmann et al. 2004), neuroscience of consciousness (Edelman & Tononi 2000), intentionality (Freeman 1997a), cognitive and emotional integration (Lewis 2005, Pessoa 2008) and even long-studied psychological pathologies (such as schizophrenia—Loh et al. 2007, Ford et al. 2007) among others. Most importantly for our purposes, nonlinear neurodynamics is changing our view of how the brain *works* and what the mind *is* and provides a methodological mechanistic continuity with current dynamical system approaches to cognition (as we have integrated them on our morphophylogeny of agency). The goal of this section is to relate our conjecture model to some of the principles that the new sciences of brain dynamics are bringing forth.

5.1. Micro-Meso-Macro: the emergence of integrative scales

Neurons were long hypothesized to respond integrating post-synaptic potentials over a period of time and firing if the sum is higher than a threshold value (i.e. the firing of a neuron depends on the total incoming input within a time window). The implicit or explicit assumption of the received view (consistent with the integrate-and-fire model) is, in general terms, that “a neuron sends its message (...) down its axons to all neurons to which it is anatomically connected. Those receiving neurons combine (e.g. sum and threshold) all the different inputs that they receive from all neurons to which they have connections.” (Fries 2005:474). Therefore, inputs to sensory neurons and subsequent anatomical connections specify information-processing routes. The traditional way of conceiving the architecture of the brain, and the mode in which it specifies cognitive functionality, is to think on a hierarchical organization composed of ascending (sensory) and descending (motor) pathways. Sensory input “moves” bottom-up on the control hierarchy through which increasingly abstract features are extracted (Barlow 1969). The ascending pathway reaches associative or central headquarters at the top of the hierarchy from which an action-plan is delivered to the descending pathway, where it is decomposed and processed down the motor hierarchy to reach the muscles²⁶.

26 Although this is a very simplified picture many contemporary textbooks and reference books use similar descriptions of the functional organization of the cortex: “There are several different sensory systems localized in the cerebral cortex, including the somato-sensory, visual, auditory, vestibular, taste, and olfactory systems. Within each of these systems there is a hierarchical organization for information processing such that input from sensory receptors in the periphery is relayed through the thalamus, first to the primary sensory cortex, then the secondary sensory cortex, and finally to the cortical association areas. The primary sensory cortices are involved in detecting, localizing, and discriminating

This view has progressively lead to a different perspective. In short, the emerging view conceives the brain in a permanent ongoing activity, waves of electrical activity travelling in many directions, resonating and synchronizing at different temporal and spatial scales and modulating local activity as higher order patterns of oscillations emerge (Buzsáki & Draguhn 2004). I will highlight four significant aspects that have contributed to this shift providing the basic conceptual toolkit to reconstruct some current views in neurodynamic research. The first is the view that neurons act as resonators rather than integrators, the second is the massive connectivity of individual neurons and their collective organization in clusters, third the bi-directional architecture of cluster connectivity in the brain and finally the prevalence of endogenous activity

As opposed to the integrate-and-fire model of rate coding, the resonator model of neuronal dynamics defends that neurons fire when the incoming inputs have a frequency that resonates with the neuron's subthreshold oscillatory frequency, which is transformed by incoming input (Izhikevich 2007). Therefore, resonators are highly sensitive to incoming signals that are *synchronized* with their subthreshold oscillations. What matters is *when* does a spike train get to the postsynaptic neuron not how much gets there²⁷. This is why they are also called "coincidence detectors" and this mode of communication named "temporal code".

At the microscopic connectivity level, in vertebrate nervous systems a neuron is, on average, connected to over a thousand other neurons and receives inputs from the same number of them. The activity of a single neuron can hardly make a distinction in such hyperconnected networks, its contribution to trigger or inhibit the activity of a neighbour neuron (not to speak of functional clusters of neurons) is but one over a thousand. In addition, neurons spike very irregularly and their individual activity is often considered extremely noisy (in terms of predictability and functional reliability). Only if a neuron can spike in synchrony with its neighbouring neurons can it produce "significant" effects; i.e. can *they* have a significant effect (Freeman 2000).

the different properties of a stimulus, be it tactile, visual, auditory, etc. The secondary sensory cortices receive this information and integrate it with previous memories of the stimulus to help identify it. The sensory association areas, in turn, receive and integrate information from different sensory modalities to provide conscious perception of the stimulus and initiate plans for behavioral action in response to it. (...) Three areas, the prefrontal, parietal, and temporal cortex, integrate and interpret sensory stimuli of all modalities and plan and execute behaviors in response to the stimuli." (Cechetto & Topolovec 2002: 671—675)

27 Although some important consequences follow if one adopts either one of such models (particularly for large-scale model of neurodynamic organization) the resonator and integrator models are not mutually exclusive, both models can be interpreted as different phases of the same neuron (for a detailed account of this debate see Izhikevich 2007).

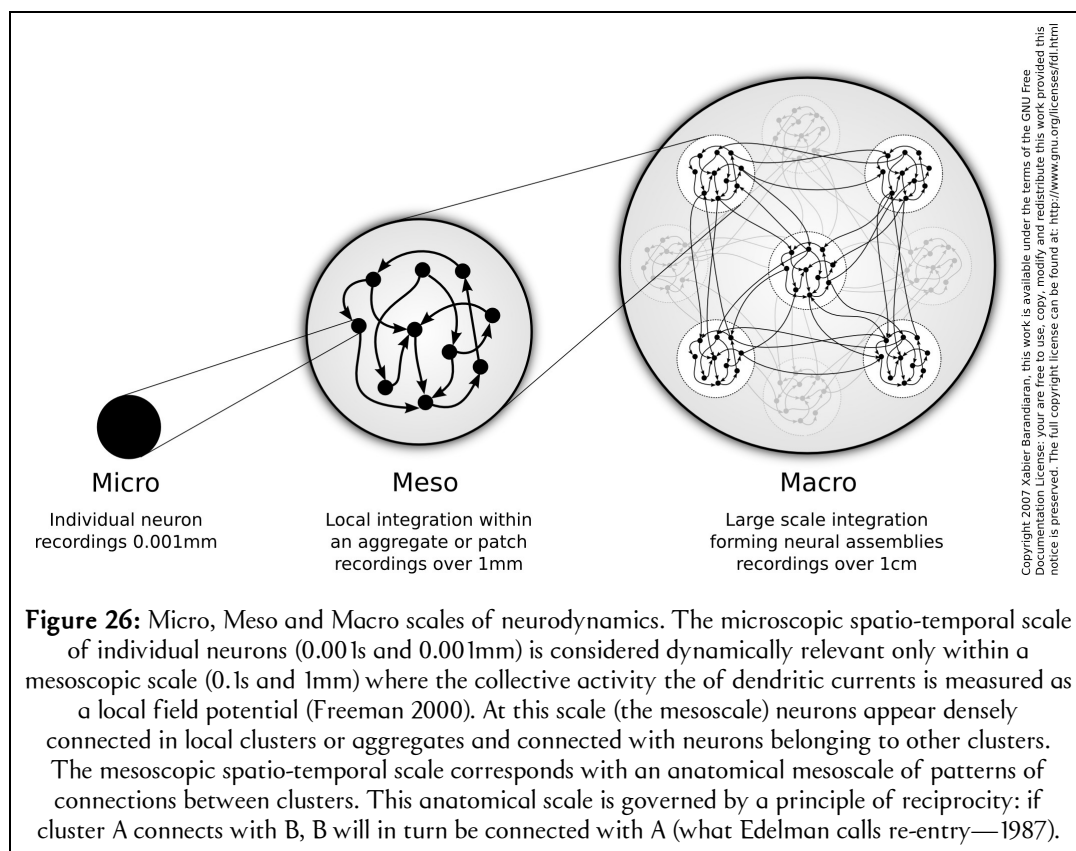


Figure 26: Micro, Meso and Macro scales of neurodynamics. The microscopic spatio-temporal scale of individual neurons (0.001s and 0.001mm) is considered dynamically relevant only within a mesoscopic scale (0.1s and 1mm) where the collective activity of dendritic currents is measured as a local field potential (Freeman 2000). At this scale (the mesoscale) neurons appear densely connected in local clusters or aggregates and connected with neurons belonging to other clusters. The mesoscopic spatio-temporal scale corresponds with an anatomical mesoscale of patterns of connections between clusters. This anatomical scale is governed by a principle of reciprocity: if cluster A connects with B, B will in turn be connected with A (what Edelman calls *re-entry*—1987).

The microscopic spatio-temporal scale of individual neurons (0.001s and 0.001mm) is thus considered dynamically relevant only within a mesoscopic scale (0.1s and 1mm) where the collective activity of dendritic currents is measured as a local field potential (Freeman 2000). Interestingly at this scale (the mesoscale) neurons appear densely connected in local clusters or aggregates²⁸ and connected with neurons belonging to other clusters. In other words, the mesoscopic spatio-temporal scale corresponds with an anatomical mesoscale of patterns of reciprocal connections between clusters (see figure 26).

The mesoscale pattern of connections between local clusters is governed, with a few exceptions, by a principle of reciprocity; if cluster A connects with B, B will in turn be connected with A (what Edelman calls *re-entry*—1987). Modular information processing decomposition or classical hierarchical control decompositions have difficulties providing an accurate view of mesoscop-

²⁸ Different terminology is used to term these aggregates: cluster (Sporns 2004), local populations (Freeman 2000), pools (Fuji et al. 1996) etc. I will use the term *cluster* (which has become the technically most widespread term in network analysis) to name these populations of neurons with high number of connections between them and selective connections with other clusters.

ic brain dynamics. For every ascending pathway on the sensory hierarchy there is a descending pathway (and similarly for the motor descending pathways) so that there is a continuous top-down bottom-up feedback blurring the explanatory power of the hierarchy (see figure 28 top, page 331). In addition, there are “horizontal” reciprocal connections between the different layers of the “hierarchy” (e.g. between early visual and late motor layers) so that, again, the hierarchy is bypassed and blurred²⁹.

Finally, each cluster is, to some extent, dynamically ‘autonomous’: when neurons within a cluster are interconnected above a certain threshold the cluster is able to generate and sustain its own collective activity without necessity of input from adjacent or sensory areas (Freeman 2000)³⁰. Instead of stimulus driven, the activity of the central NS comprises a high level of endogenously generated and maintained activity (Llinás 1988). Sensory input does not reach an architecture that is quiet and ready to process the incoming stimulus but falls into an active dynamical context.

Roughly speaking, then, rather than the individual-neuron-based serial processing of incoming stimuli, it is the temporal synchronization of reciprocally connected clusters with endogenous activity what constitutes the pillar of current neurodynamic approaches to brain functioning. Within this framework the problem is not so much the establishment of the processing of incoming stimuli but the coordination of the activity of diverse and specialized local clusters to produce a global coherent behaviour. How does the brain integrate the continuous activity along different areas so as to produce a flow of complex behaviour in a sensorimotor context?

5.2. Neurodynamic cortical structures in the gamma-band

The solution to the problem of neurodynamic coordination is currently believed to be achieved through *synchronized oscillations* (Bressler & Kelso 2001, Varela et. al. 2001, Fries 2005)³¹. Assuming that neurons operate as resonators, spiking probability will depend on a neuron’s subthreshold oscillations. Therefore, for two neuronal groups to “communicate” (i.e. to be mutually dynamically sensitive) they must oscillate together, their frequency and phase needs to coincide. This is what is called coherence (Fries 2005). Experimental evidence suggests that on top of the *anatomical connectivity*, syn-

29 For a detailed review of the problems of the hierarchical perspective on perception see Vidal Miranda (2005).

30 The facts that clusters are capable of generating internal rhythms and, therefore, considered ‘autonomous’ by Freeman should not be confused with the view that such clusters constitute information processing ‘modules’ because reciprocally connected and recurrently interacting dynamic clusters do not map into linearly decomposable functional modules.

31 In addition, oscillation based synchrony is the energetically most efficient form of coordination (Buzsáki & Draguhn 2004).

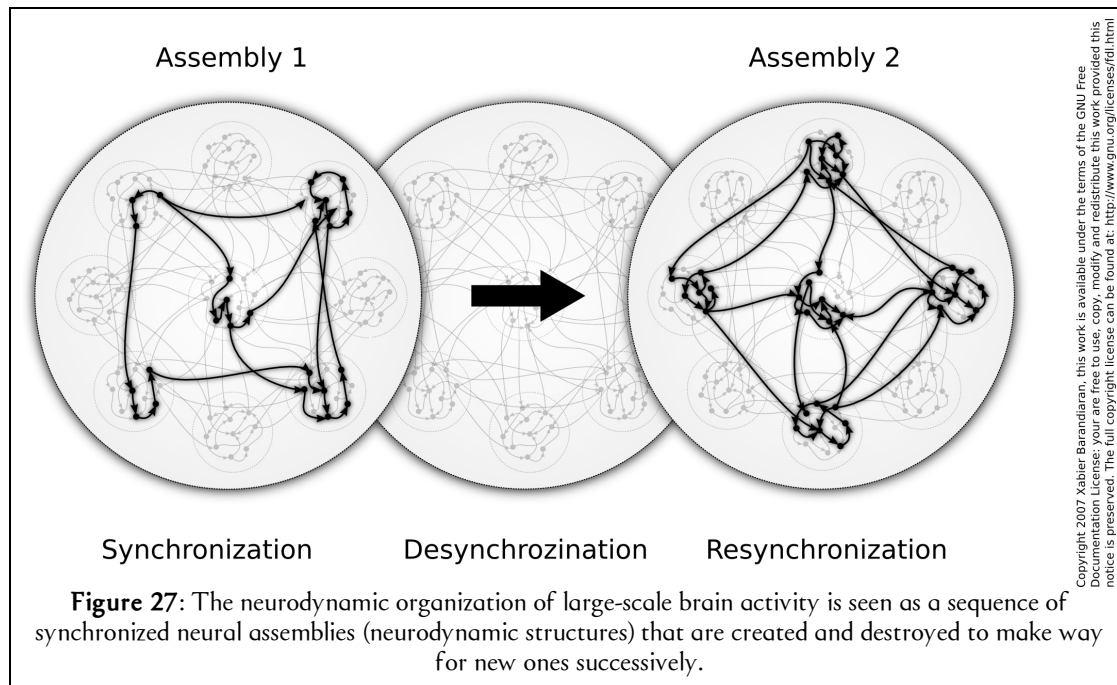


Figure 27: The neurodynamic organization of large-scale brain activity is seen as a sequence of synchronized neural assemblies (neurodynamic structures) that are created and destroyed to make way for new ones successively.

chronized oscillations create forms of transient *effective or functional connectivity* among different neural clusters yielding brain dynamics into a coordinated pattern that integrates sensorimotor activity into a coherent pattern of amplitude modulation that spans across the entire brain, constituting a “cognitive moment” (see figure 27). Llinás summarizes the picture in the following manner:

A purely hierarchical connectivity alone is simply too slow and unwieldy to keep pace with the ever-changing aspects of the external world. There must be another mechanism at work. (...) This mechanism is most likely temporal coherence. (...) Mapping connectedness in the time domain, superimposed on top of the limited possibilities of spatial connectedness, creates a vastly larger set of possible representations through the almost infinite possibilities of combination. This is the concept of perceptual unity based on spatial and temporal conjunction. Building on physical connectivity, the nerve cells of the brain have created an “interlocking” solution: the synchronous binding in the time domain of those individual neuronal activities. (...) This time-interlocking phenomenon is temporal coherence. (Llinás 2001:120—121)

Different names have been used to term the dissipative neurodynamic patterns that are formed through the distributed synchronization of different neural clusters: *cell assemblies* (Hebb 1949, Varela 1995), *dynamic cell assemblies* (Fuji et al. 1996), *dynamic core* (Edelman and Tononi 2000), *global attractors* (Freeman 2001), *dissipative dynamic structure* (Llinás 2001), *chaotic*

attractors (Tsuda 2001), *effective connectivity* (Fries 2005), etc. Based on Fuji et al.'s formulation of the hypothesis (1996: 1333s) we can integrate what these terms have in common and provide a neuroscientifically grounded notion of neurodynamic structures: *A neurodynamic structure is a 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 situation, spontaneously form a cross-correlated pattern through a coherent synchronous integration of local clusters* (and where different clusters can participate on the formation of different neurodynamic structures).

The gamma frequency band (30-150Hz) has attracted the most attention as a privileged frequency for the formation of neurodynamic structures (Varela 1995, Varela et al. 2001, Engel & Singer 2001, Jensen et al. 2007). This is so for a number of reasons. First, the gamma-cycle (10-30ms) matches the duration of the time window of cortical neuron's temporal integration (the excitatory postsynaptic potential is approximately ~ 10 ms). Second, oscillations in the gamma band are typical of inhibitory-excitatory connections between clusters of neurons and, as reviewed by Fries et al. (2007) gamma oscillations have been found in many diverse areas (hippocampus, somato-sensory cortex, parietal cortex, etc.) in different animals, and have been found to correlate with different cognitive functions such as attention, perceptual binding, working memory maintenance, and awareness more generally³². Finally, gamma-band permits coordination at the large scale in the brain. This is so because gamma-band oscillations fit nicely with the time span required for whole brain integration. Varela (1995) made the following calculation: at a spike travelling speed of 10m/s a spike wave would take about 40ms to make a return trip between both hemispheres (25cm travel). One such cycle will thus involved a frequency of $1000/40 = 25$ Hz. The gamma band is just above the minimum frequency required to synchronize the activity of the full brain (or, at least, the cortex):

[T]he main idea is that fast-oscillations in the gamma-beta range serve as carriers for a phase synchronization of neuronal activity, thus allowing for a process of selection by resonance into a transient coherent ensemble that underlies the unity of cognitive act in a fraction of a second. (Varela 1995:83)

On the one hand the selective formation of neurodynamic structures in the gamma band permit to render sensory flow into its neurodynamic context of relevance within the task at hand. For instance, recognition of ambiguous figures (with different possible interpretations) has been shown to correlate with gamma-band oscillations synchronizing into a unitary percept (Rodríguez et

32 In addition the gamma band has been shown to integrate precisely the top-down and bottom-up processes along the classical anatomical "hierarchy-pyramid" of sensorimotor processes (see Herrmann et al. 2004 for a review), so that even the activity of early sensory layers of the cortex appear modulated by "top-down" influences (Vidal Miranda 2005).

al. 1999, Varela et al. 2001—see also Engel & Singer 2001 for a review of visual sensory integration through synchronization patterns). On the motor side, so to speak, neurodynamic structures permit to shape motor degrees of freedom into a coherent and constrained coordination:

[A]s a complex movement proceeds, the control system must be able to reconfigure itself dynamically so that these collectives are cast temporarily, quickly dissolved, and rearranged as required. Because the central nervous system has many possible solutions for a given motor task, it follows that any given functional synergy organized by the brain must be a fleeting, dissipative construct. (Llinás 2001:36).

Thus, neurodynamic structures may appear integrating sensory, motor and/or associative neural clusters into a coherent pattern. Although synchronized at the large-scale (comprising distant areas into the same neurodynamic structure), the weight of the integration may be shifted to sensory areas (e.g. if the task is the recognition of an ambiguous figure) or to motor areas (e.g. if focus is required on finger movement coordination). What remains important is that all three major cortical divisions (sensory, motor, and association) and their endless specialized ramifications can, at any given time, influence or directly participate on the formation of the dominant neurodynamic structure. It is thus hypothesized that what a complex neurodynamic structure permits is a kind of “remembered present” (Edelman 1989): the reconstruction of a complex of neurodynamic relationships integrating traces of past experiences within the current situation, extending it over future expectancies.

Within the plastic, integrative and diverse space opened by the evolution of behavioural agency, when the developmental process builds up a complex web of behavioural organization, neurodynamic structures are the expression of a complexity that moves beyond the activation of a simple sensorimotor pathway. A neurodynamic structure is formed when a coherent integration is reached in brain activity expressing a high level of complexity: a cross-correlated activity of a huge number of distributed and specialized local patterns of activity. This way neurodynamic structures can be thought of as the outcome of a dynamic interplay between functional segregation and integration (Tononi et al. 1998, Edelman & Tononi 2000, Sporns 2004):

Brain dynamics is shaped by the continual interplay between segregation and integration, which manifests itself as complexity and metastability across multiple spatial and temporal scales. Segregation and integration are evident in the anatomical organization of brain networks, as well as their functional connectivity recorded in the context of perceptual or cognitive processing. Segregation reflects the need to optimally extract and generate local information, while integration is necessary in order to create coherent brain states. Together, they provide the means by which the brain can (optimally?) balance the simultaneous demands of information extraction and binding, responding to the moment-by-moment challenges of the external world by selecting transient and globally coherent internal states. (Sporns 2004: 213)

Interestingly, this transient integration of distributed neuronal activity has been proposed to be the mark of consciousness (in the sense of primary awareness, i.e. not necessarily self-reflexive) since it correlates with states of wakefulness and awareness while it disappears in states of deep sleep, anaesthesia, and epileptic seizures (Tononi & Edelman 1998, Edelman & Tononi 2000). It has also been proposed that the formation of a neurodynamic structure along thalamocortical areas could occur at the gamma-band (Llinás et al 1998), or that gamma-band synchronization in visual areas is the correlate of conscious visual experience (Crick & Koch 1990) or perceptual awareness (Engel & Singer 2001). The formation of an integrated and dominant neurodynamic structure (what Edelman and Tononi call a “dynamic core”) would account for the unity of conscious experience and its temporal constraints (we can only perceive and decide above a few hundred milliseconds—coincident with a few gamma cycles). And it could also account for the unique differentiation and richness of each conscious process, thanks to the always different participation of specialized neuronal units on each integration process:

The dynamic core is a process, since it is characterized in terms of time-varying neural interactions, not as a thing or a location. It is unified and private, because its integration must be high at the same time as its mutual information with what surrounds is low, thus creating a functional boundary between what is part of it and what is not. The requirement for high complexity means that the dynamic core must be highly differentiated—it must be able to select, based on its intrinsic interactions, among a large repertoire of different activity patterns. Finally, the selection among integrated states must be achieved within hundreds of milliseconds, thus reflecting the time course of conscious experience. (Tononi & Edelman 1998:1850)

Despite the theoretical consistency and the observational evidence across differing cognitive tasks, brain areas and animal species, the hypothesis that neurodynamic structures are the building block of neurodynamic organization, faces an important challenge. The presence of a mere correlation between neurodynamic structures and cognitive processes does not necessarily imply the existence of direct causal link between observed coordinated oscillations and behaviour. Recently, however, experimental evidence by Meyer-Lindenberg and colleagues (2002) has demonstrated that the presence of such neural spatio-temporal correlations is not a mere epiphenomenon. Previous work by Kelso (1984, 1995) had shown that repetitive metronome-paced movement of the index fingers of both hands showed two stable patterns at different frequencies (in-phase and out-of-phase movement). The dynamics of this bi-stable behaviour could be captured by a relatively simple model of differential equations. Based on this work Meyer-Lindenberg and colleagues found two neurodynamic structures corresponding to both stable movements (and other parameters of the model) using neuroimaging techniques. Then, transcranial magnetic stimulation was used to perturb one of the

neurodynamic structures (corresponding to out-of-phase bimanual coordination) provoking a transition to the other neurodynamic structure (in phase coordination) which, in turn, produced the subsequent behavioural change. The direct manipulation of the neurodynamic trajectory around the first attractor (i.e. the neurodynamic structure corresponding to the out-of-phase coordination) showed that the transition to the second was the cause of a subsequent behavioural transition.

At the scale of full behavioural scenes, the neurodynamic organization of large-scale brain activity is seen as a sequence of dissipative dynamic structures that are created and destroyed to make way for new ones successively. An agent, as it behaves, transits from one, very specific, neurodynamic structure to another one, moving from a highly integrated activity to an equally integrated but distinct one. It has been suggested that positive and negative feedback between three or more clusters of neurons suffices to create a background chaotic activity that renders neural populations into a metastable but highly sensitive dynamic regime (Freeman 2000). This makes possible that interconnected clusters form a dissipative dynamic structure in a very short time. Located at a second order state transition point, global brain dynamics has access to a high number of metastable emergent configurations, i.e. neurodynamic structures, that it can reliably access in a quick and efficient manner (Chialvo 2004, Werner 2006). This characteristic behaviour has also been conceptualized as *chaotic itinerancy* (Tsuda 2001): large scale integration transits strange attractors that are quickly formed and dissolved. Chaotic itinerancy does not permit the brain to get locked in a particular attractor (characteristic of epileptic seizures) and maintains a permanent state of sensitivity to incoming stimuli and endogenous variability. This way, neuronal activity not directly involved on a synchronized assembly serves as a background activity providing the context for the transition to a new neurodynamic structure, always sensitive to new internal and sensory fluctuations.

The centrality of gamma oscillations should however be taken with care³³. The gamma-band is very effective to manage communication in small time windows and appears specially well suited to produce quick modulations of neuronal activity to engage the brain in ongoing fluid behaviour with the environment (precisely at the temporal scale of animal sensorimotor coping). But the gamma-band should not be considered in isolation. The power spectrum of brain recordings is much richer in frequency bands and modulatory scales; from very low frequencies of 0.05Hz to ultra fast ones of 500Hz. It has also been suggested that higher frequencies are characteristic of local coordination, whereas slower frequencies are typical of larger brain coordination (Buzsáki & Draguhn 2004). Synchronization occurring at different scales

33 Thanks to Juan Vidal for pointing to me some early oversimplified claims regarding the gamma-cycle and for providing a wiser perspective on neuronal dynamics at the large-scale.

plays an important role on modulating and forming different local neurodynamic structures and a global neural coordination of the entire brain may require a higher order synchronization subsuming local patterns (as suggested by Engel & Singer 2001).

5.3. Embodiment, emotions and appraisals: the integration of the sensorimotor and the interior nervous system

What we have just described could be called “horizontal integration in the gamma band” in which the neurodynamic domain appears almost purely decoupled from the living body, mediating sensorimotor loops mostly through cortical structures (sensory, motor and associative). Lewis (2005a) has recently developed a comprehensive review of what he has called “vertical integration” (the expression is taken from Tucker et al. 2000) that is hypothesized to operate mainly through the theta-band (4-8Hz)³⁴. This vertical axis should not be confused with the bottom-up and top-down processing of the classical hierarchical view along the cortical horizontal axis (see figure 28). The vertical axis comprises bidirectional interactions between the body (through the autonomic NS), the brain stem, the hypothalamus and the limbic system up to cortical areas. The intersection between the vertical and the horizontal axis can be interpreted, to a large extent, as a body-NS interaction system or interior vs. sensorimotor NS integration. As we shall see, this is what many neuroscientists take to be the biological basis of cognitive-emotional integration.

The top-down influence along the vertical axis involves the triggering of body-responses mediated by the autonomic NS that controls body arousal as a form of action readiness. It also involves influences of the activity of the horizontal axis into hypothalamic and limbic structures controlling the endocrine system with multiple bioregulatory effects. This is a necessary requirement at big body sizes in which body regulation must be tuned to sensorimotor requirements, the SM-NS will affect the NS-I so that adequate body responses precede interactive needs (e.g. more oxygen and nutrients for muscles). This type of bodily regulatory reaction to specific environmental and internal situations has long been considered as the basic constituent of emotions (Damasio 1994). The bottom-up influence on our vertical axis involves a bodily evaluation of sensorimotor engagements as a modulation of neurodynamic structures. As we have previously seen when analysing neurocognitive development, at an early stage this type of neural modulation is carried on the basis of the effect of the behavioural consequences on the body and its internal

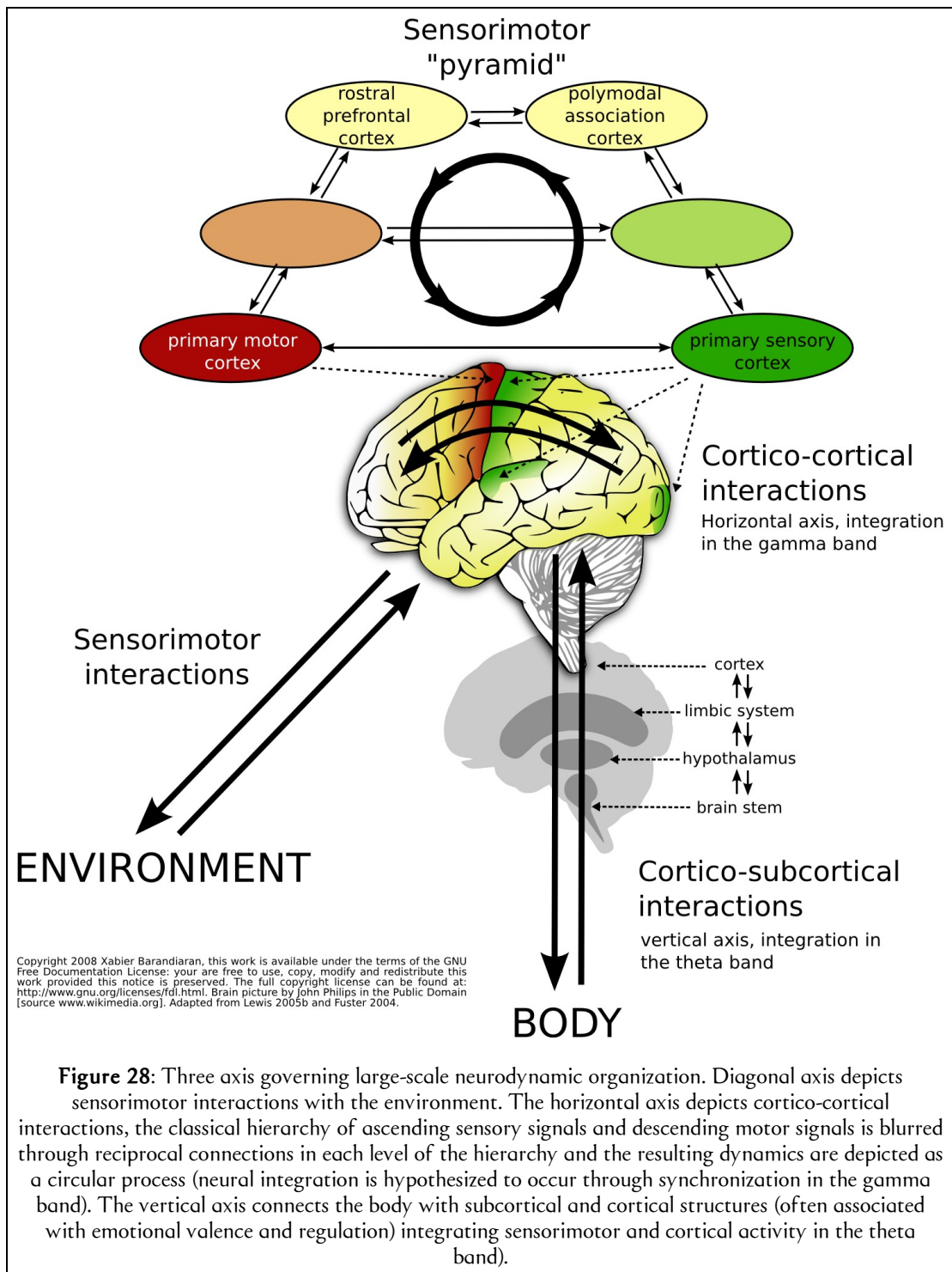
³⁴ The distinction between the vertical-theta and horizontal-gamma dimensions is clearly an oversimplification of many important details but it serves as a helpful heuristic to integrate different neuroscientific data.

bioregulatory dynamics, serving as an evaluative or “value” system that facilitates the formation of connectivity patterns and the organization of brain dynamics through development (Edelman 1987, Damasio 1994, Friston et al. 1994).

The interactions at the vertical axis could have stopped here: evaluation of interactions by the NS-I and modulation of NS-I by SM-NS for interactive “readiness”. But emotions and affects are not primitive evolutionarily predefined boxes ready to be activated only when the body is in danger, pain or great pleasure. They involve continuous reciprocal feed back with the sensorimotor integrative axis and are moulded as a result of the history of such interactions. In fact, the vertical axis is not just a channel for top-down and bottom-up separate and unidirectional influences, but an integrative axis on its own right whereby different feed-back loops open up the space for a rich entanglement. First and foremost, emotional body states (interactive readiness) will feed back to cortical areas through the vertical axis. It has been suggested that this feed-back constitutes a form of internal perception of the emotional body states triggered by the sensorimotor system. This way the sensorimotor cycle integrates a kind of “feeling” or emotional valence of its ongoing activity. This feed-back between body and sensorimotor NS is involved on what Edelman calls *primary consciousness* (Edelman 1987, Edelman 1992, Tononi & Edelman 2000) and Damasio calls *feelings based on secondary emotions* (Damasio 1994) or *core consciousness* (Damasio 1999). These authors suggest that the coupling with the body’s bioregulatory dynamics provides a sense of background bodily self, an internal environment that is in continuous feed-back with what is going on through the sensorimotor environment. It is this way that the body becomes an integral part of neurodynamic organization.

Far from cognition and emotion constituting two separate processes, they appear as the two poles of an integrated vertical axis, and it is precisely through the intersection of the vertical and the horizontal axis (often associated with the prefrontal cortex—Damasio 1994, Pessoa 2008) that this integration occurs. This integration through the vertical axis is hypothesized to be achieved through a higher level synchronization in the theta band and involves the mediation of subcortical areas and the effect of diffuse neuromodulation (Lewis 2005a, 2005b). On the one hand limbic and hypothalamic structures (as integral part of the vertical axis) include important hubs³⁵ mediating cortical areas and become privileged zones to drive higher level cortical neurodynamic stability. On the other hand, the activity of diffuse neuromodulatory pathways from the brain stem to cortical regions has been shown, as Lewis reviews, to operate in the theta band. Neuromodulation actively contributes to the “on-line” formation of large-scale neurodynamic structures through amplification (driving a positive feedback effect) or inhibition (stabil-

³⁵ Hubs are, within the language of network analysis, nodes that are highly connected.



izing through negative feedback). This permits to focus attention into emotionally salient events (what Lewis calls emotional interpretation) by integrating sensorimotor and emotional dynamics into a coherent pattern. It may permit action monitoring for complex tasks by coordinating the transitions between different gamma-band neurodynamic structures through higher-order theta-band synchronization. This hypothesis may be supported by recent empirical studies showing how the slower theta frequency band directly modulates the faster gamma-band in various cognitive tasks (Canolty et al. 2006, Jensen & Colgin 2007)³⁶.

This second order integration is also in charge of error assessment and learning. Based on Holroyd & Coles (2002) Lewis relates the vertical-emotional integrative axis to error awareness (particularly through the anterior cingulate cortex and the ventral tegmental area) through increasing theta band synchrony. Luu et al. (2004) have also shown that when subjects realize that they made an erroneous action (forced by a rapid decision task) limbic theta band modulation participates on error awareness and subsequent learning. Lewis also stresses the involvement of vertical integration in *learning* (understood in terms of potentiating or debilitating synaptic connections). It facilitates long term potentiation of synapses (LTP) through the release of neuromodulators and neuropeptides and by sustaining arousal and attention long enough so that LTP can take place. Moreover, theta synchronization has also been correlated as facilitating LTP. Thus, from the huge set of sensorimotor contingencies that impinge upon neural dynamics, only those particularly relevant for theta integration as emotionally interpreted will lead to changes in connectivity:

According to the present model, the self-stabilization phase of an emotional interpretation is the necessary precondition for this learning. Once appraisals have stabilized, interpretations, action plans, and expectancies endure for some period of time, as mediated by coupled cognitive and emotional elements. These enduring couplings seem necessary to strengthen the connections responsible for learning. (Lewis 2005a: 177)

Interestingly we are not any more in a stage of early development or highly constrained adaptive behaviour. As development proceeds, the modulatory or regulatory capacities involved in attentional focus, action monitoring, error assessment and learning will not depend any more on highly constrained emotional responses (corresponding to biological or evolutionarily adaptive norms) but will be triggered by a “cognitive” evaluation. Thus, a qualitative but gradual change occurs as the vertical axis achieves higher levels of integration with the horizontal axis. Antonio Damasio and collaborators have studied the role of emotions in decision making and higher cognit-

36 Unfortunately these studies have only focused on cortical measurements without references to the origin or source of theta oscillations.

ive tasks (Damasio 1994, Bechara 2004). As the interaction between the NS-I and the SM-NS becomes stronger, and after emotional body responses are internalized through the vertical axis, the role of the NS-I and the body is progressively reduced and their modulatory capacity taken over by higher-level cortical areas (particularly those situated at the intersection between vertical and horizontal axis). An “as if loop” is thus created bypassing direct body and NS-I reactions, so that horizontal integration (“cognition”) can recruit vertical modulatory capacity (“emotion”) for its own appraisal; *combining* both vertical and horizontal integration in the same process. This is particularly clear when cognitive error or dissonance triggers the activation of areas correlated with bodily pain without any bodily or adaptive thread being present. For instance, the anterior cingulate cortex (that I previously mentioned to be involved in error assessment) is also “activated mostly when you are in pain of the intractable, long-term type, such as that from cancer” (Llinás 2001:158). It is tempting to suggest that, at some point of evolution, the mechanisms in charge of the adaptive regulation of behaviour (in correspondence with biological normativity; e.g. the pain caused by a thread to the body) became, through a developmental process, recruited to regulate the internal consistency of neurodynamic organization (e.g. cognitive error awareness).

To sum up, the emerging simplified picture is that of a behavioural organization through two integrated axis. At the horizontal axis sensorimotor coupling is sustained by the rapid formation of an assembly of distributed neural processes coordinated through phase-locked synchronization in the gamma-band: a rapid collective (but selective) emergent pattern of neural spikes mediating sensorimotor interactions. One could interpret this axis as constituting a web of neurodynamic structures (instantiated on the synaptic architecture of the cortex). The intersection between this axis of sensorimotor integration through cortical regions and a vertical axis of emotional integration with the former provides a form of self-regulation: a) the channelling or modulation of sensorimotor neurodynamic structures (attention and appraisal) and b) the transformation of their anatomical basis through synaptic potentiation and depression (learning) when some “cognitive” dissonance takes place (error assessment).

5.4. Autonomy and normative regulation in neurodynamic organization

The oversimplified picture I have drawn spans in complexity and important details at different levels (even if we ignore socially structured environments that may involve a qualitative change of neurodynamic organization). On the one hand, integration and modulation occurs at different scales (not just at the gamma and theta bands) and each integrative scale may be thought of as

corresponding to different cognitive-emotional aspects like moods, attention, involvement and as eliciting different forms of learning and memory (like working-memory, episodic-memory, skill-acquisition, trauma, etc.). On the other hand, the precise neuroanatomical details simplified into horizontal (cortical) and vertical (subcortical) axis appear fragmented on multiple distributed pathways, connectivity hubs and specialized areas with varying properties. In addition, although often postulated as a single coherent dynamic core, neurodynamic structures or synchronized cell assemblies are more generally shown to appear in local spatial scales forming different and overlapping reverberating circuits nested into higher level structures. Finally, real life behaviour does rarely involve a single sensorimotor coupling and a more realistic conception would require to consider the subsumption of sensorimotor schemes at different scales. For instance, visuomotor tracking (in itself is a sensorimotor loop) may be nested within spatial navigation, which may in turn be subsumed under a more general scheme of escaping from a predator. However, a simplified abstract model permits to see the big picture and to establish valuable bridges between a model of behavioural autonomy and current large-scale modelling of brain dynamics.

I proposed a conjecture model of Mental Life as a self-maintaining web of sensorimotor structures and its adaptive regulation. The simplified picture of neurodynamic organization just sketched provides some insights into how Mental Life could be realized in the brain. First, the notion of neurodynamic structure as the building block of behavioural organization seems to be supported by the fact that brain dynamics appear condensed or governed, moment by moment, by a transient dissipative assembly of synchronized and phase-locked clusters of neurons. At every “cognitive moment” the dimensionality of the brain is reduced, as a result of an spontaneous self-organized process, to a lower dimensional effective space (constituting a hyperdescription of the system). This emerging causal structure may be in turn composed of different local synchronized patterns. A form of “on-line” self-maintenance and coherency can already be pictured at this level. The dominating pattern is stabilized as the result of a mesoscopic negotiation of internal stability dependencies (gamma-band synchronization), modulated by macroscopic order parameters (theta-band amplification through neuromodulators) and the bodily and environmental context of behaviour. This self-maintained pattern, however, quickly dissolves to make way for a new one. What remains throughout successive neurodynamic structures is the capacity to generate integrated coherent patterns:

My hypothesis is that brain dynamics is governed by an adaptive order parameter that regulates everywhere neocortical mean neural firing rates at the microscopic level, and which finds expression in maintenance of a global state of self-organized criticality. Under perturbation by environmental input (including that from the body), brain dynamics moves away from its basal attractor and gener-

ates repeated state transitions in its attempt to regain balance. These local states form chaotic itinerant trajectories that constitute a search for a course of action that can be predicted to restore balance. (Freeman 2005: 205)

At a longer time-scale the brain maintains its capacity to keep “making sense” of the situation, to be able to create integrated coherent neurodynamic structures mediating behaviour), through the transformation of its underlying synaptic architecture. This way, a new sense of self-maintenance may be said to emerge for the long-term organization of behaviour through anatomical re-configuration. The first thing to highlight is that the anatomical basis of large-scale integration permits the coexistence of many overlapping accessible neurodynamic structures. As behaviour generates and explores new or altered situations, the anatomical basis of neurodynamic structures is reinforced or transformed through long term potentiation and depression of synaptic connections. The space of possible overlapping neurodynamic structures is thereby transformed. This transformation occurs as a result of vertical integration capable to induce anatomical changes in the face of emotional valence of an interaction, like the error assessment after certain expectancies are not met. Yet, this adaptive reconfiguration of the web of potentially available neurodynamic structures is not driven by a dissociated or decoupled system, imposing constraints or a direct control over lower level dynamics. Rather, through vertical integration, it is the autonomy of sensorimotor integration that modulates and modifies itself. It is the large-scale brain dynamics, coupled to the environment, that will specify *when* and *which* synaptic connections will be strengthened or modified maintaining, in the long-run, its own capacity to keep sustaining coherent sensorimotor interactions, preserving its internal organization, its ability to make sense of the situation, of its world. Since the same neuronal clusters may take part on different overlapping neurodynamic structures, changes occurring as a result of learning might propagate along the full organization of behaviour, equilibrating it through the successive transformations induced as a result of the course of an interactive history

Thus, whereas horizontal integration provides a form of complex self-organization that may be considered autonomous (both at the scale of ongoing behaviour and at the scale of long-term organization), vertical integration may be said to permit a second order modulation or adaptive regulation of this autonomy originated as a result of the very activity of cortical integration³⁷. A new level of normativity can be said to appear as a result. It is not

37 To some extent this distinction between two different “levels” (horizontal-vertical, gamma-theta) might be understood as forming a hierarchy. However this sense of “hierarchy” is very different to the classically postulated cortical hierarchy or certain forms of control hierarchy in which a physically distinguishable part exerts a direct control over other identifiable parts. The same clusters of neurons may be synchronized at the gamma *and* theta bands, and still theta frequency be modulating the gamma, without implying that cluster A induces or modifies the activity of cluster B.

any more the satisfaction of biologically adaptive norms that renders behavioural phenomena functional, but their ongoing contribution to the maintenance of a behavioural identity, that resulting from a history of interactions, instantiated on a web of neurodynamic structures and regulated by its emotional embodiment. In fact, the major role of emotional dynamics (vertical integration) can be interpreted as the regulation of sensorimotor organization towards global reinforcement and coherency: attention (amplification of sensory stimuli in order to satisfy a certain goal or stability condition), stress (the generation of global micro-instabilities for dynamic rearrangement—like in simulated annealing models), satisfaction (reinforcement of the dynamic structure whose produced behaviour satisfies a global stability condition) and, in general, shaping the neural dynamics in the direction of its own far-from-equilibrium conservation, preserving the organization that has been created through the history of the agent. Functionality and normative regulation are not any more confined to the realm of metabolic biophysical self-maintenance but *become*, in the domain of behaviour, autonomous (originally defined by loose exogenous constraints but progressively self-determined through recursive environmental interactions and its capacity for self-modification).

This way, Mental life brings-forth the emergence of a new mode of agency. This mode of agency overcomes the limitations and problems we have previously seen adaptive behaviour suffered to satisfy the minimal structure of intentional agency. This is so because the new level of agency is constituted directly within the behavioural domain, so that norms are both cause and effect of sensorimotor interactions (not derivatives of an underlying metabolic infrastructure). As a result there is no dissociation between the source of the norms and the regulation according to these norms, vertical integration permits a continuous monitoring and reconfiguration of action according to the self-organized and holistic activity of (potentially) the entire brain.

6. MENTAL DEATH: THE EFFECTS OF SENSORIMOTOR DEPRIVATION AND SOLITARY CONFINEMENT

What is notably missing from large-scale models of neurodynamic organization is the role of recurrent environmental interactions on the shaping of neural dynamics and its architectural transformations: the world itself, as that through which Mental Life exists. Despite the pragmaticist bias of some neuroscientists (e.g. Freeman) the environment coupling plays a very impoverished role on current large-scale neurodynamic models. Those few studies that move beyond mere perception or finger coordination focus on classical conditioning, rarely is goal-oriented operand conditioning neurodynamically analysed at the large-scale, not to speak of tasks requiring a selective coupling



Figure 29: Sensorimotor deprivation torturing techniques applied at Guantanamo Bay. From top to bottom on the right picture: the subject bears silent earphones that isolates him from any surrounding sound, the glasses and the mask avoids any visual and smell/taste sensory stimulation. The thick globes on the hands impede touch sensing, while the hands tied and sitting over the legs crossed precludes any full body movement. As a result proprioception remains reduced to a minimum and prolonged maintenance of such a position will soon lead to numbness. Prolonged exposure to this (and similar) conditions is hypothesized to lead to Mental Death: the loss of core beliefs and the sense of self.

with an environment that is rich in opportunities and features. This may be partly due to the constraints imposed by brain imaging techniques, that often require the subject to be quiescent and limit experiments to the presentation of different images on a screen and the pushing of a single button. Another limitation of current studies is the time scale at which measurements can be done and analysed (often just for a few seconds), whereas a full account of a self-maintaining and cohesive neurodynamic organization would require a much longer period of observation (from hours to weeks, up to months).

It is therefore difficult to find empirical support for the claim that Mental Life is an interactively self-maintained form of organization. According to the proposed model not only during development (which has been widely demonstrated as reviewed in previous sections) but also in adulthood, neurodynamic organization depends on a set of recursively satisfied sensorimotor interactions that integrate the world as a constitutive side of its own existence. A testable hypothesis will follow from this feature of Mental Life: that the isolation from a world of regular interactions would destroy or severely disrupt the organization of behaviour, the identity of the cognitive subject. Interestingly (and unfortunately for many) this is what occurs. Sensorimotor deprivation is a widely used torture technique in criminal and milit-

ary prisons all around the globe and, at least temporary, it can show devastating effects on the core identity and cognitive capacities of the subjects³⁸.

After the second World War, during the 50s and 60s, the CIA and the Pentagon initiated an intense research program of the effects of sensory deprivation, motivated by the impact of Communist “brain washing” techniques. Major psychiatric and neuro-psychology researchers were involved directly or indirectly on this research program across Harvard, Princeton and Cornell Universities (among many other medical and scholar institutions) including Donald Hebb (who latter described it as an “atrocious procedure”). He concluded:

Once development is complete, does the organism then become less dependent psychologically on sensory stimulation? When a man’s or a woman’s character is formed, his or her motivations and personality pattern established, is character or personality an entity that exists so to speak in its own right, no matter where or in what circumstances (assuming physical health and reasonable bodily welfare)? In the Korean war the Chinese Communists gave us a shocking answer: in the form of brainwashing. The answer is No. Without physical pain, without drugs, the personality can be badly deformed simply by modifying the perceptual environment. It becomes evident that the adult is still a function of his sensory environment in a very general sense, as the child is. (Hebb 1958:111)

By the end of the 70’s the research program was complete and in 1983 the CIA’s *Human Resource Exploitation Training Manual* included detailed techniques and reports on its psychological effects. The Manual begins with the following declaration of purposes:

The purpose of all coercive techniques is to induce psychological regression in the subject by bringing a superior outside force to bear on his will to resist. *Regression is basically a loss of autonomy*, a reversion to an earlier behavioral level. As the subject regresses, his learned personality traits fall away in reverse chronological order. He begins to lose the capacity to carry out the highest creative activities, to deal with complex situations, to cope with stressful interpersonal relationships, or to cope with repeated frustrations. (CIA 1983, p. K-1, italics added).³⁹

According to McCoy (2006) these torture techniques are now applied worldwide by the USA “intelligence” agencies. Figure 29 shows a number of prisoners in Guantanamo Bay. The right side picture shows in detail the conditions of sensorimotor deprivation. From top to bottom: the subject bears silent earphones that isolates him from any surrounding sound, the glasses and the mask avoids any visual and smell/taste sensory stimulation. The thick globes on the hands impede touch sensing, while the hands tied and sitting over the

38 The first steps of my research on this topic were facilitated by an anonymous publication on the internet, it is meant to be an early draft of paper for the American Psychological Association, but I could never found it published as such. The anonymous version can be found at: <http://valtinsblog.blogspot.com/2007/08/my-apa-paper-on-isolation-sensory.html>

39 Quoted in <http://valtinsblog.blogspot.com/2007/08/my-apa-paper-on-isolation-sensory.html>

legs crossed precludes any full body movement. As a result proprioception remains reduced to a minimum and prolonged maintenance of such a position will soon lead to numbness. Similar forms of torture have been described without directly immobilizing the subject, just by transforming and reducing the environment (e.g. use of isolation chambers in which meals are presented irregularly and lights are switched on and off at irregular intervals to avoid any sensory regularity for the subject to rely on).

The effects that these regimes and torture techniques have on behavioural and cognitive organization have also been reported in extreme cases of burn patients (Lasagna & Germoglio 2002) forced to sensorimotor deprivation due to burn injuries (to avoid infection and pain) and also in solitary confinement in USA Supermax prisons (Grassian & Friedman 1986, Grassian 1993, Haney 2003) and other regimes of solitary or sensory deprivation:

In addition, a pattern of psychiatric disturbances similar to those I found at Walpole [a supermax security prison where solitary confinement was used] have been seen in a variety of other—non-prison—settings, all of which, however, share in common features of restricted environmental stimulation: These latter have included observations of prisoners of war, of hostages, of patients with impairment of their sensory apparatus (for example, hearing or visually impaired patients), of patients confined in the intensive care unit, of patients undergoing long term immobilization in hospital (e.g. spinal traction patients), of observations of psychiatric difficulties suffered by explorers (for example, Arctic and Antarctic exploration by individuals and small groups) and of observations of difficulties encountered by pilots during solo jet flight. (Grassian 1993: 6)

But what are exactly the effect of these types of regimes? According to John Zubek, one of the leading researcher on sensory deprivation during the 60s: “[B]oth simple and complex measures of visual and motor coordination are *adversely affected* by sensory and perceptual deprivation (...) and considerable impairment on unstructured behaviors” (Zubek 1969:165, italics added)⁴⁰. Psychiatrist Stuart Grassian (expert on pathological effects of solitary confinement) reports that, in extreme cases, solitary confinement can lead to psychotic disorganization and a rare complex of symptoms with no parallel (other than an Acute Organic Brain Syndrome): a *delirium*. Components of this complex of symptoms are: hyperresponsivity to external stimuli, perceptual distortions, illusions and hallucinations, panic attacks, rage, problems with impulse control, difficulties with thinking, concentration and memory, intrusive obsessional thoughts and overt paranoia (Grassian & Friedman 1986, Grassian 1993, see also Haney 2003 for a recent review). Although some symptoms are described in high level psychological terms (like “intrusive obsessional thoughts”) which are difficult to apply to non-human or minimal models of Mental Life, extreme symptoms include difficulties on motor coordination (Zubek 1969), sensory disorganization (like distortions, illusions

40 Quoted in: <http://valtinsblog.blogspot.com/2007/08/my-apa-paper-on-isolation-sensory.html>

and hallucinations) and the dominance of modes of sensorimotor engagement generated by subcortical areas: like “problems with impulse control” and rage⁴¹. In sum, it may be said that prolonged exposure to regimes of sensorimotor deprivation can lead to a severe disorganization of behaviour.

Interestingly, these symptoms may be grouped and labeled under the psychological notion of “Mental Death” (Ebert & Dyck 2004). Similar conditions to those described above can “destroy a person’s identity, the essence of mental death, is well documented in the torture literature” (Ebert & Dyck 2004: 618). Ebert and Dyck reference more than 10 different reviews and include an illustrating testimony of a torture survivor: “the individual loses his/her ability to react normally; and they especially lose their ability to process, understand and articulate their experiences, ... resulting in numbing or a paralysis, and mental death” (Ebert & Dyck 2004: 618). Mental Death reveals the core organization of psychological or cognitive identity in two respects: the conditions in which Mental Death is caused and the psychiatric definition and symptoms associated with Mental Death. Conditions in which it is caused involve: psychological torture (unpredictable environment, change of roles, impossible choices), solitary confinement and sensory deprivation. According to Ebert & Dyck (2004) Mental Death is considered an extreme and complex case of PTSD (Post Traumatic Stress Disorders). The main hypothesis is that totalitarian control of the environment, torture and isolation destroy core beliefs, assumptions and constructs that permit the person to cope with her world. Regarding its psychological characterization Ebert & Dyck provide the following definition of Mental Death:

[T]he essence of mental death is the loss of identity, defined as the perception of sameness and continuity of the self—and the self in relation to others—based on the relative constancy of one’s assumptions, beliefs, values, attitudes, and behavior. (Ebert & Dyck 2004: 621).

Conversely, by a single change of substitution of “death” by “life” and “loss” by “maintenance” we get a appealing definition of Mental Life: *The essence of mental life is the maintenance of identity, defined as the perception of sameness and continuity of the self—and the self in relation to others—based on the relative constancy of one’s assumptions, beliefs, values, attitudes, and behavior.* Needless to say, this characterization of Mental Life extends beyond the limits of a minimal characterization but remains illustrative of how the life-death dichotomy might be extended beyond its organic dimension. A more naturalized definition would require to reveal the anatomical and neurodynamic basis of assumptions, beliefs and attitudes and the process of its progressive disintegration.

41 Shame rage can be triggered even when basal ganglia and hypothalamus (evolutionarily highly conserved and developmentally constrained structures) are surgically separated from cortical areas in cats (Brown 1911).

It seems reasonable to assume, given the best available picture of brain organization, that Mental Life involves the orchestrated activity and architecture of the brain as is coupled to the environment it enacts and in which it depends. To say it with Donald Hebb, the issue of sensory(motor) deprivation “raises the whole question of the relation of man to his sensory environment” (Hebb 1958:111). As is occurs with the far-from-equilibrium organization of organic life, always dependent on a self-regulated flow of matter and energy for its continuing existence, psychic or mental life also depends on its sensorimotor coupling with the environment in order to sustain its own far-from-equilibrium existence. The mode in which identity and world are intertwined in the mode of organization that characterizes the autonomy of behaviour reveals the needful significance of the behavioural with the world: over and above the mere satisfaction of biological/metabolic or evolutionary constraints (remember that subject may be perfectly feed while deprived of an environment) and the referential-semantic relationship between internal representations and externalstates of affairs.

Chapter 9: Recapitulation and conclusion

1. EXTENDED SCHEMATIC SUMMARY: FROM CONCEPTS TO COMPLEX GENERATIVE MODELS, FROM LIFE TO MIND.

On what follows I offer an extended summary of the previous chapters in the form of a schematic recapitulation of the main claims and concepts¹:

Thesis: A concept for minds can be build as a generative mechanistic model of Mental Life: the self-maintaining adaptively autonomous neurodynamic organization that appears in the domain of behaviour. The domain of behaviour emerges as a dynamical system that is hierarchically 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.

PART I: Methodological considerations. The concept of mind requires a generative model of a complex mechanism, that takes the form of a naturalizable, universalizable and minimalist model of agency.

→ Chapter 1: On the categories of mind and mechanism

- The Cartesian mindset permeates our concept of minds assuming a divide between mind and mechanisms. Even when rejecting the Cartesian metaphysical dualism a category mistake pervades whereby the mind (as a negation of the mechanical) belongs to the same category as the mechanical.
- As a result, mental vocabulary is introduced as an intermediary cause of intelligent behaviour and minds are considered intellectual para-mechanical processes. The “intellectualist myth” posits an internal domain of reasons and mentalistic causes of intelligence, disregarding the every-day behavioural *know-how* characteristic of minds.

¹ For clarity references have been omitted, central concepts and chapter topics are marked with bold and auxiliary concepts underlined, and central claims with italics.

- The category mistake is solved by considering minds as learned dispositions to behaviour. But dispositions can be operationally linked back to organized mechanisms.
- The concept of mind requires to account for everyday behaviour through mechanisms capable to generate learned dispositions to behave in certain characteristic manner.

→ **Chapter 2: On types of explanations and the epistemology of modelling complex mechanisms**

- Descartes' influence on our mode of conceptualizing the mind stirs fundamentally from his influential mechanistic methodology.
- Mechanicism, as an explanatory strategy, subsumes both behaviourist and functionalist explanatory modes while providing a stronger epistemological adequacy by permitting to relate abstract (and often indeterminate) functional decompositions into structures upon which we can operate.
- Traditional mechanicism (as conceived by Descartes) has been methodologically forced to remain within the limits of atomistic reductionism (assuming a one-to-one mapping between structural and functional decomposition).
- Complex systems could not be properly modelled assuming atomistic reductionism because they show some properties that defeat a one-to-one mapping between structure and function: internal emergent functionality, interactive emergence, chaos and hierarchical organization.
- Complex systems have many-to-many(-to-all) emergent generative mappings between structures and functions. Biological and cognitive systems are of this type.
- Complex systems can be modelled using a bottom-up synthetic mechanistic methodology using computer simulations that can assemble together data from model organisms and numerically reproduce their behaviour.
- Simulation models can have four epistemic uses: mechanistic-empirical, functional, generic and conceptual. This last use is of particular interest to define a concept for minds while a circulation between all four levels will become an integral part of the construction of such concept.

→ **Chapter 3: On how to model a concept for minds starting with the analysis of a generic notion of agency**

- Models are epistemic artefacts that involve a set of assumptions and interpretative frameworks. How models relate or connect with phenomena is a complex process in itself, in which the construction of the model bears a significant weight on its use as an epistemic artefact. A philosophical investigation about modelling minds requires to make explicit this constructive process.

- Theories are families of models that share construction assumptions and interpretative frameworks and that appear hierarchically structured. Within this hierarchy some concepts occupy a privileged status as basic level models, whose construction involves a circularity between empirically connected models and higher level theoretical principles. The concept of mind belongs to this class.
- The task of developing a **concept for minds** involves the making explicit of the process of construction of a conceptual model within a theory of complex systems.
- The **goal of the thesis** is stated: *making explicit the construction of a model for minds as a complex generative organization: i.e. as a class of hypothetical systems in which a set of component processes relate to each other in a specific interdependent manner capable to generate a set of characteristic features that distinguish mindful systems*. More specifically a concept for minds requires making explicit the domain in which minds occur [domain specificity] and the type of organization, how parts relate to each other to generate a set of characteristic properties [organizational specificity].
- Under the difficulties of current computational and dynamicist theories in cognitive science to demarcate the specificity of the mind as a distinct phenomenon, agency is chosen as a bottom-level template to depart the construction of a model for minds.
- The most generic notion of agency involves *a system doing something by itself according to certain goals or norms within a specific environment*. From this early characterization, three different, although interrelated, conditions can already be distinguished: (i) there is a system as a *distinguishable entity on its environment*, [individuality conditions] (ii) this system is *doing something by itself* on that environment [causal asymmetry condition] and (iii) it does so according to a certain goal or *norm* [normativity condition].
- Nature has produced increasingly complex forms of agency until the appearance of characteristically cognitive or mindful ones. We can take advantage of this evolutionary process and adopt a morphophylogenetic approach (the analysis of evolutionary transitions in the organization of agency) as a scaffolding for the construction process.

Part II: The morphophylogeny of agency. A conceptual trip from basic autonomous systems to adaptively behaving multicellular organisms uncovers a number of crucial transitions: from the constitution of a physico-chemical identity and normative functionality to the requirement of a decoupled, situated and potentially high-dimensional subsystem in charge of controlling the interactive requirements of its biological embodiment.

→ Chapter 4: On the minimal organization of life as autonomous agency.

The metabolic dimension of life, understood as a network of self-sustaining chemical reactions endowed with a self-produced membrane and capable of controlling its boundary conditions, already constitutes a minimal model of agency.

- The self-organized dissipative structures formed as macroscopic patterns of the non-linear interactions between multiple components, in far-from-equilibrium thermodynamic conditions, provides a preliminary notion of self.
- The formation of a self-organized chemical component production networks (a precursor of metabolism) permits the initiation of a process of complexity increase through variation and selection (the alternative foundation of agency in replicators is argued to be insufficient).
- When this network is capable to produce a self-encapsulating membrane that actively constraints its boundary conditions **basic autonomous systems** appear.
- Basic autonomous systems can be said to instantiate a minimal model of agency:
 - Their self-produced and controlled system-environment distinction (through a selective membrane) provides a first sense of individuality.
 - The differential contribution of molecular reactions and membrane components to the self-maintenance of the whole provides a first sense of functional normativity: processes *must* occur in a certain manner in order for the system to assure its continued existence.
 - When energy flows are coupled and recruited to produce functional constraints for its self-maintenance the system provides a first sense of causal asymmetry on system-environment relationships, it is the subject of actions.

→ Chapter 5: On the appearance and limitations of decoupled regulatory subsystems and their sensorimotor coupling with the environment, leading to adaptive motile agency.

- Basic autonomous systems are severely limited on their robustness by pure structural stability. The solution to this bottleneck requires the appearance of a decoupled subsystem capable to regulate the system in relation to its boundaries of viability.
- The central idea of decoupling is that the system generates a set of mechanism that do not participate directly on the metabolic cycle and become thus free to deal with the environment to change some internal or external conditions for the benefit of the system. Adaptive agency is the normative regulation of interactive processes by a decoupled subsystem.

- **Motility** is the capacity of an autonomous system to adaptively *direct* the cycle of sensorimotor correlations through self-generated displacements in the environment. Detection-of and functional response-to environmental changes becomes, in the case of adaptive motility, a sensorimotor cycle that permits the enaction (active generation) of an environment out of sensorimotor correlations.
- Adaptive agency is severely limited on its chemical implementation in unicellular systems due to the potential interferences (that may arise between chemical pathways mediating sensorimotor coupling and metabolic pathways) and also due to the difficulty of diffusion processes alone to achieve full body coordination for unitary motile displacements as size increases.
- The appearance of eukaryotes permits an increase in size but appears still bounded by organizational bottlenecks. Big unicellular eukaryotes rely on epithelial conduction to coordinate motility (not on diffusion process) but the homogeneous propagation of action potentials along their membrane is very limited to achieve complex sensorimotor correlations.
- The genomic toolkit of eukaryotes permitted the formation of multicellular organisms with the creation of a stable internal milieu and cellular and tissue specialization.

→ **Chapter 6: On the origin, characterization and modelling of the embodied and situated nervous system leading to adaptive behaviour**

- Since the early evolution of multicellularity the electric conductivity that cells are able to sustain made possible an extended network of dynamic variability capable of coordinating distant cells [epithelial conduction], generating endogenous rhythms that could further be affected by sensory signal transduction mechanisms producing different sensorimotor couplings between organism and environment.
- However, epithelial conduction alone was not able to exploit the implicit potential that electric conductivity affords for agency.
- Cnidarian agency, exploits a more specific and agentially relevant use of electric conduction: that sustained on the activity of networks of differentiated cells, the neurons (specialized on the selective, integrable and modifiable transmission of electric potentials).
- Cnidarian agency appears, however, limited by a radially symmetric bodyplan and its primitive developmental mechanisms. It was the Bilaterian bodyplan, together with the appearance of the developmental regulatory capacity of the Hox gene family, what permitted a much wider exploration of what neurons (and bodies) could do for agency, as is manifest on the significant complexification of agency that the Cambrian explosion

permitted. It is along Bilaterian evolution that the full potentiality of the nervous system will be exploited.

- The significance of the **nervous system** (NS) relies on the fact that it can generate a high dimensional dynamical organization of behaviour which is plastic and stable at the same time and free from a high number of limiting constraints that other modes of intercellular coordination present.
- In particular the NS is hierarchically decoupled: its dynamics are under-determined by the continuous process of metabolic self-maintenance, reproduction and morphological transformations that the organism undergoes by means of digestion, cell replication and growth. Its modelling permits, and requires, the abstraction from local mechanistic-molecular details and the specification of the NS as a high dimensional dynamical system, composed of electrochemical signals between neurons, their modulation and connectivity matrix².
- The sensorimotor and biological embodiment of the NS and its situatedness mark essential characteristics of its mode of functioning.
 - Sensorimotor embodiment. On the one hand, the NS is embodied on specific tissues capable of channelling metabolic energy into efficient mechanical energy by means of muscle contraction and a system of fixation (along with the mechanical cohesion of the body) that transforms these contractions into articulated, directional and reversible movements. As a result, the dynamics of the NS is a function of the body's morphological and mechanical properties that shape possible interactions and relative positions through enabling biomechanical constraints. On the other hand, embodied sensory surfaces appear limited and specialized on specific ranges, transformations and filtering of sensory perturbations exploiting physical and relational features to transform environmental perturbations into functionally pre-structured signals. In addition, both sensory and motor embodied surfaces appear highly intertwined due to the circular and recursive nature of sensorimotor interactions that have evolved and developed together.
 - Biological embodiment: The NS is embedded on a living body some of whose internal bioregulatory functions are subsumed under the NS itself. Through this coupling the NS can modulate sensorimotor processes in coordination with internal regulatory demands by means of body signals (pain, pleasure, hunger, satiation, etc.) with a high neuromodulatory capacity.

² Once the necessary mechanistic details are available to disambiguate between available functional models it is the dynamic organization of behaviour of that model that matters and, conversely, only when a dynamic model is available can all the molecular, physical and environmental factors be functionally integrated in a mechanistic model.

- **Situatedness:** The situatedness of action within a geometrical environment permits to achieve functional behaviour exploiting relative position and orientation and, in more abstract and general terms, exploiting the recurrent effect of motor transformations into sensory states through the properties of the environment.
- **Adaptive behaviour** appears as a result of the hierarchical decoupling from metabolism, the sensorimotor and biological embodiment and the situatedness of agency.
- The simplest picture of adaptive behaviour may be seen as that of the embodied and situated dynamics of the NS constrained by its architecture and modulated by body signals to generate, when coupled to the environment, a behaviour that maps into the maintenance of a number of essential variables of the organisms within viability boundaries.

Part III: The distinctive organization of mind as form of life. Minds require the emergence of a new form of individuality and normativity distinct from generic biology: a form of adaptive self-maintenance of the organization of behaviour that *becomes* autonomous through a developmental/historical process of self-determination.

→ **Chapter 7: On the problems of interpretation for the sufficiency of adaptive behaviour for mindfulness (understood as intentional agency).**

- A research tradition that roots cognitive capacities in biological organization has defended that metabolic normativity provides the foundations for intentional agency in the sense that sensorimotor interactions become intentional by means of their correspondence or congruence with the norms defined by their metabolic embodiment.
- A minimal structure of intentional agency might be phenomenologically depicted without appealing to propositional structures: *the interactive compensation of a deviation from a norm*. The structure of intentionality-in-action and our own phenomenological experience can serve to evaluate the sufficiency of candidate mechanisms of adaptive behaviour for intentionality.
- Detailed analysis of failure of behaviour generating mechanisms in model organisms to operate according to the norms dictated by metabolism reveals the attribution of intentionality problematic. An in principle reason for to ground intentional agency in biological normativity is that the source of the norms (metabolism) is dissociated from the mechanisms generating behaviour (whose operations in accordance with the norms may not always be assured by the organismic totality but highly specified by inherited constraints).

- A new level of normativity, that is generated in the very domain of behaviour, may be envisioned as an alternative approach to ground intentionality.

→ **Chapter 8: On Mental Life as the adaptively autonomous organization of behaviour and its grounding for a new level of identity, normativity and agency.**

- From a set of initial conditions of great developmental plasticity, triggered by biological adaptive signals and channelled by early architectural and body constraints, the NS generates more and more internal constraints and interdependencies between behaviourally emergent self-organized patterns (neurodynamic structures or habits).
- Gradually, the preservation of the internal coherency of these nested structures takes over the regulation of situated and embodied brain dynamics. Throughout its development this neurodynamic organization must have internalized biological-adaptive requirements in the form of internal stability dependencies. But the mode in which it has done so, and the resulting organization is not any more something reducible-to or deducible-from an evolutionary adaptationist story, nor from the adaptive requirements of metabolism.
- **Mental Life** becomes into existence when the *adaptive conservation of the internal organization of neural dynamics becomes the main principle of sensorimotor regulation*.
- Three research fields provide feasibility to the previous model of Mental Life: the evolution of cognition, cognitive developmental neuroscience and large-scale models of brain dynamics.
- The evolution of cognition suggests that relational (organism-environment), phylogenetic and developmental factors converge presumably on the fact that there is a set of evolutionary trends towards increasingly more diverse, plastic and integrated behavioural agency.
 - On the side of organism-environment relationships, this type of behavioural agency was both required and enabled by ecological environments in which high-energy resources were available to agents through increasingly sophisticated sensorimotor correlations and ecological and social co-evolutionary traps, through hostility and cooperation, that preclude a return path to previous or less complex behavioural modes of life.
 - On the phylogenetic side, empirical evidence is reviewed for the progressive increase (in different lineages) of brain areas associated with integration and developmental plasticity of behavioural agency. In turn a whole set of bodyplan transformations came to be facilitated and required by an increase in behavioural complexification, including the

division of the NS into a sensorimotor system and a system of the interior (in charge of bioregulatory functions).

- Developmental cognitive neuroscience reveals the co-development of body and brain, and the progressive organization of the brain through self-sustained behavioural interactions.
 - There are no sufficient genomic resources to specify the connectivity of vertebrate brains and there is evidence for generalist genes participating in brain development. As a result, the development of neurodynamic organization is highly open to the environment.
 - What is characteristic and distinctive of brain development is the extent to which it relies on the available plasticity, its dependency on endogenous neural activity and the duration of the developmental process. Particularly significant is the fact that cognitive development is strongly behaviour driven: from prenatal stages, sensorimotor coordination generates the regularities and organization that leads to subsequent developmental stages, incorporating environmental-contextual features that are, in turn, the result of its own activity.
- Large scale models of brain dynamics reveal an integrated dissipative organization of neural activity and its self-regulation.
 - Rather than the individual-neuron-based serial processing of incoming stimuli, it is the temporal synchronization of reciprocally connected clusters of neurons with endogenous activity what constitutes the pillar of current neurodynamic approaches to large-scale brain functioning.
 - Within this framework the problem is not so much the establishment of the processing of incoming stimuli but the coordination of the activity of diverse and specialized local clusters to produce a global coherent behaviour.
 - The solution to this problem is the formation of neurodynamic structures understood as *a 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 situation, spontaneously forms a cross-correlated pattern through a coherent synchronous integration of local clusters.*
 - All three major cortical divisions (sensory, motor, and association) and their endless specialized ramifications can, at any given time, influence or directly participate on the formation of a dominant large-scale neurodynamic structure.
 - At the scale of full behavioural scenes, the neurodynamic organization of large-scale brain activity is seen as a sequence of dissipative dynam-

ic structures that are created and destroyed to make way for new ones successively.

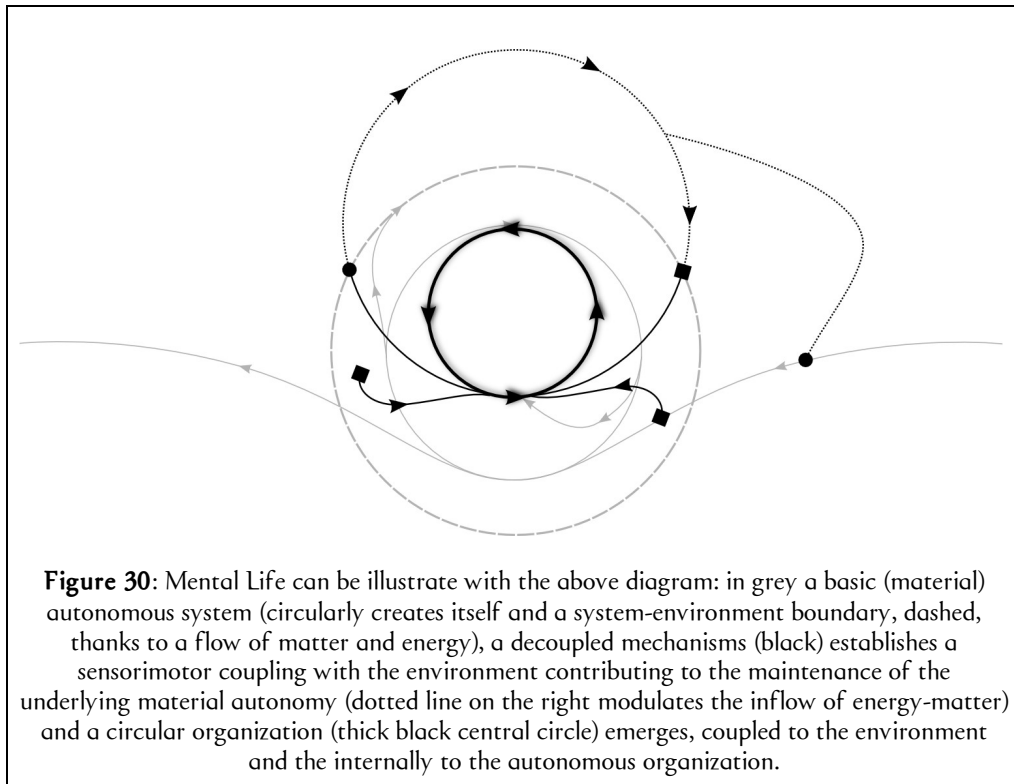
- The emerging simplified picture is that of a behavioural organization through two integrated axis:
 - At the horizontal axis sensorimotor coupling is sustained by the rapid formation of an assembly of distributed neural processes coordinated through phase-locked synchronization in the gamma-band. This axis is interpreted as constituting a self-organized web of neurodynamic structures (instantiated on the synaptic architecture of the cortex).
 - The intersection between the horizontal axis and a vertical axis of emotional integration (in the theta band) provides a form of self-regulation: a) the channelling or modulation of sensorimotor neurodynamic structures (attention and appraisal) and b) the transformation of their anatomical basis through synaptic potentiation and depression (learning) when some internal dissonance takes place (error assessment).
- A new level of normativity can be said to appear as a result. It is not any more the satisfaction of biologically adaptive norms that renders behavioural phenomena functional, but their ongoing contribution to the maintenance of a behavioural identity, that resulting from a history of interactions, instantiated on a web of neurodynamic structures and regulated by its emotional embodiment.
- Mental life brings-forth the emergence of a new mode of agency. This mode of agency overcomes the limitations and problems that adaptive behaviour suffered to satisfy the minimal structure of intentional agency. This is so because the new level of agency is constituted directly within the behavioural domain, so that norms are both cause and effect of sensorimotor interactions (not derivatives of an underlying metabolic infrastructure). As a result there is no dissociation between the source of the norms and the regulation according to these norms, vertical integration permits a continuous monitoring and reconfiguration of action according to the self-organized and holistic activity of (potentially) the entire brain.
- The phenomenon of mental death, that arises in prolonged exposure to regimes of sensorimotor deprivation, solitary confinement and torture reveals the dependency of Mental Life on its structured engagement with the world.

2. A CONCEPT FOR MINDS

At the beginning of this thesis I promised to deliver a naturalized, universalizable, and minimal conceptual model for minds by providing the specificity of the domain in which mind occurs and its organizational specificity:

- **Domain specificity:** Mental Life occurs in the domain of behaviour, a high dimensional dynamical system that is hierarchically 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.
- **Organizational specificity:** Mental Life is an adaptive web of sensorimotor dynamic structures created, sustained and regulated through an open history of interactions with body and world.

The resulting model can be illustrated with figure 30.



The proposed conceptual model (linguistically expressed) is generative in the sense that identity, normativity and causal asymmetry (or pro-activity) can be said to result from the proposed organization. The self-organized and cohesive web of sensorimotor structures constitutes an *individuality* as a form of

active differentiation and preservation of an organized totality. A level of distinctive *normativity* is generated by the differential participation of behavioural processes on the stability dependencies and constraints required for the continuation of sensorimotor organization. Finally the dynamic flow of behaviour appears knocked or nested on a highly integrated internal neurodynamic organization providing the means by which the generation of *actions* is asymmetrically laden towards it: Mental Life generates a form of agency

In relation to the epistemological constraints stated in chapter 2, the concept of Mental Life is *universal* (or at least of wide generality) on that no particular molecular or anatomical detail was in any way essentially involved on its definition. It is *naturalizable* on that the model: a) was constructed from the bottom-up (departing from physical and biochemical organization) up to the evolution of cognition, b) it is amenable to the operational framework of dynamic (simulation) modelling (as shown by some pioneering robotic experiments—Di Paolo) and c) it falls in congruence with some current approaches to cognitive development and large scale brain functioning. It is *minimalist* on that it satisfies all but no more than the necessary and sufficient conditions to generate a distinctive level of individuality, normativity and agency. It was argued that nothing below the proposed model may be able to account for mindfulness, in particular adaptive behaviour rooted on biological/metabolic normativity faces serious problems to account for intentional agency. In addition, the model of Mental Life remains open to incorporate different sensorimotor modalities, a social developmental and behavioural context, the inclusion of tools and technologies, different time-scales and differential contribution of specific processes.

The project is, admittedly, unfinished. The conceptual linguistic model presented would benefit from a robotic simulation model in which holistic properties emerge and be analysed. Such minimal simulation model should include at least two neurodynamic sensorimotor structures, created in an interactive developmental process, stability conditions for this coupled structures should be identified and the system should be shown to operate preserving this minimal organization and compensating perturbations interactively. This simulation work has been partly developed but still remains unfinished and could not be included as an integral part of this thesis (although its construction and preliminary experiments have served as a conceptual guide and intuitive pump to develop the present work).

Other possible expansions of the current work include:

- A detailed reconstruction of the notions of goal, intentionality and meaning within the context of Mental Life and their relation the computational notions of representation (Fodor 1975) and information (Dretske 1981³) and some current attempts to reconstruct these no-

3 Dretske, F. I. (1981) *Knowledge and the flow of information*. University of Chicago Press.

tions in embodied and situated approaches to cognitive science (Wheeler 2005, Bickhard 1993, 2004).

- The connection of the concept of Mental Life with other philosophical notions, in particular with the Aristotelian notion of *psyche* and the Heideggerian notion of *Dasein*.
- A detailed discussion with the philosophical tradition that roots intentionality in evolution (Millikan 1984, Dennett 1987, 1995).

I have argued that the study of complex mechanisms has opened the possibility for re-conceptualizing the mind as the dynamic flow of behaviour knotted in neurodynamic organization, a fleshy living machine that extends its ghost-like shadow to an always openworld of needful significance.

Maybe philosophical problems are hard not because they are divine or irreducible or meaningless or workaday science, but because the mind of *Homo sapiens* lacks the cognitive equipment to solve them (...) Humanly thinkable thoughts are closed under the workings of our cognitive faculties, and may never embrace the solutions to the mysteries of philosophy. (...) [T]he mind owes its power to its syntactic, compositional, combinatorial capacities. Our complicated ideas are built out of simpler ones, and the meanings of the relations that connect them. (...) We grasp matter as molecules, atoms, and quarks; life as DNA, genes and a tree of organisms (...) All are assemblies of elements composed according to laws, in which the properties of the whole are predictable from the properties of the parts and the way they are combined. (...) But there is something peculiarly holistic and everywhere-at-once and nowhere-at-all and all-at-the-same-time about the problems of philosophy. Sentience is not a combination of brain events or computational states. (...) the 'I' is not a combination of body parts or brain states or bits of information, but a unity of selfness over time, a single locus that is nowhere in particular. Free will is not a causal chain of events and states, by definition. Although the combinatorial aspect of meaning has been worked out (...) the *core* of meaning—the simple act of referring to something—remains a puzzle, because it stands strangely apart from any causal connection between the thing referred to and the person referring. (...) Our thoroughgoing perplexity about the enigmas of consciousness, self, will, and knowledge may come from a mismatch between the very nature of these problems and the computational apparatus that natural selection has fitted us with. (Pinker 1997:561—565)

References

- Abouhamad, W.N., Bray, D., Schuster, M., Boesch, K.C., Silversmith, R.E., & Bourret, R.B. (1998). Computer-aided resolution of an experimental paradox in bacterial chemotaxis. *Journal of Bacteriology* 180, 3757—3764.
- Ackley, D. and Littman, M. (1992) Interactions between Learning and Evolution. In Langton, et al. (Eds.) *Proc. Artificial Life II*, Redwood City CA: Addison-Wesley, pp. 487—509.
- Adler, J. (1969) Chemoreceptor in Bacteria. *Science* 166:1588—1597.
- Alberch, P (1982) Developmental constraints in evolutionary processes. In Bonner, J.T. (Ed.) *Evolution and Development. Dahlen Conference*. New York: Springer Verlag, pp. 313—332.
- Albert, J. (1999) Computational modeling of an early evolutionary stage of the nervous system. *Biosystems* 54(1-2): 77—90
- Alexandre, G. & Zhulin, I. B. (2001) More than one way to sense chemicals. *Journal of Bacteriology*, 183(16):4681—4686.
- Alexandre, G., Greer, S.E. & Zhulin, I.B. (2000) Energy taxis is the dominant behaviour in *Azospirillum brasilense*. *Journal of Bacteriology* 182:6042—6048.
- Allport, G. W. (1935) The functional autonomy of motives. *American Journal of Psychology* 50: 141—156.
- Anderson, P.A.V. (1980) Epithelial conduction: its properties and functions. *Progress in Neurobiology* 15:161—203.
- Anet, F. (2004) The place of metabolism in the origin of life. *Current Opinion in Chemical Biology* 8:654—659
- Anokhin, A.P., Birbaumer, N., Lutzenberger, W., Nikolaev A, and Vogel F. (1996) Age increases brain complexity. *Electroencephalography and Clinical Neurophysiology* 99 (1): 63—8 .
- Apel, K.O. (1998) *Teoría de la verdad y ética del discurso*. Paidós, Barcelona.
- Arbib, M. A. (2003) Schema theory. In Arbib, M.A. (Ed.) *The Handbook of Brain Theory and Neural Networks*. 2nd Edition. MIT Press, pp. 993—998.
- Arshavsky, Y., Deliagina, T. & Orlovsky, G. (1997) Pattern Generation. *Current Opinion in Neurobiology* 7:781—789.
- Ashby, W.R. (1952). *Design for a Brain. The origin of adaptive behaviour*. Chapman and Hall, 1978 edition.
- Ashby, W.R. (1956) *An Introduction to Cybernetics*. Chapman and Hall.
- Ashby, W.R. (1962) Principles of the self-organizing system. In Von Foerster, H. & Zopf, Jr.G.W. (Eds.) *Principles of Self-Organization: Transactions of the University of Illinois Symposium*. London: Pergamon Press, pp. 255—278.

- Avery, L. & Thomas, J. H. (1997) Feeding and Defecation. In Riddle, D. L., Blumenthal, T., Meyer, B. J. and Pries, J. R. (Eds) *C. elegans II*. Cold Spring Harbor Laboratory Press.
<http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=ce2.chapter.815>
- Bak, P. (1997) *How Nature Works: The Science of Self-Organized Criticality*. Oxford University Press, Oxford, 1997.
- Baldwin, J. M. (1896). A new factor in evolution. *The American Naturalist*, **30**: 441—451, 536—553.
- Baluska, F., Volkmann, D. & Menzel, D. (2005) Plant synapses: actin-based domains for cell-to-cell communication. *Trends in Plant Science* **10** (3): 106—111.
- Barandiaran, X. (2004) Behavioral Adaptive Autonomy. A milestone in the Alife route to AI? In Pollack, J., Bedau, M.A., Husbands, P., Ikegami, T., Watson, R.A. (Eds.) *Proceedings of the 9th International Conference on the Simulation and Synthesis of Living Systems (ALife9)*, MIT Press, pp. 514—521.
- Barandiaran, X. & Feltre, R. (2003) Conceptual and methodological blending in cognitive science. The role of simulated and robotic models in scientific explanation. In *Volume of abstracts of the 12th International Congress of Logic, Methodology and Philosophy of Science*, Oviedo (Spain), August 7-13, 2003. p. 171.
- Barandiaran, X. & Moreno, A. (2006a) Alife Models as epistemic artefacts. En Rocha, L.M., Yeager, 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.
- Barandiaran, X. & Ruiz-Mirazo, K. (Eds.) (2008) Special issue on “Modelling Autonomy”. *BioSystems* **91**(2).
- Barlow, H.B. (1969) Pattern recognition and the responses of sensory neurons. *Annals of the New York Academy of Sciences* **156**:872—881.
- Barr, M.M. and Garcia, L.R. Male mating behavior (June 19, 2006), WormBook, ed. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.78.1, <http://www.wormbook.org>.
- Bechtel, W. & Abrahamsen, A. (2005) Explanation: a mechanist alternative. *Stud. Hist. Phil. Biol. & Biomed. Sci.* **36**: 421—441.
- Bechtel, W., & Richardson, R. C. (1993) *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press.
- Bedau, M.A. (1997) Weak Emergence. *Philosophical Perspectives* **11**: 375—399.
- 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.
- Beer, R. (1998) Representations and Cognitive Explanations: Assessing the Dynamicist's Challenge in Cognitive Science. *Cognitive Science* **22**: 295—318.
- Beer, R. D. (1995) A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence* **72**:173—215.

- Beer, R.D. (1990). *Intelligence as Adaptive Behavior: An Experiment in Computational Neuroethology*. Academic Press.
- Beer, R.D. (1995) A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence* **72**:73—215.
- Beer, R.D. (2000) Dynamical approaches to cognitive science. *Trends in Cognitive Sciences* **4**: 91—99.
- Békés, F (1975) Simulation of kinetics of proliferating chemical systems. *Biosystems* **7**: 189—195.
- Berntson, G.G. & Cacioppo, J.T. (2000) From Homeostatic to Alldynamic Regulation. In Cacioppo, J.T., Tassinary, L.G. & Berntson, G.G. (eds.) *Handbook of Pshychophysiology*, Cambridge University Press, pages 459—481.
- Bertschinger, N., Olbrich, E., Ay, N. & Jost, J. (2008) Autonomy: An information theoretic perspective. *BioSystems* **91**(2):331—345.
- Bichat, X. (1800) *Recherches Physiologiques sur la Vie et la Mort*.
- Bickhard, M.H. (1993) Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence* **5**:285—333.
- Bickhard, M.H. (2004) Process and Emergence: Normative Function and Representation. *Axiomathes* **14**:135—169.
- Bickhard, M.H. & Terveen, L. (1995). *Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution*. Elsevier Scientific.
- Bickle, J. (2003) *Philosophy and Neuroscience. A Ruthlessly Reductive Approach*. Kluwer Academic.
- Block, N. (1980) What is Functionalism? In Block, N. (Ed.) *Readings in the Philosophy of Psychology*. Cambridge, MA: Harvard University Press.
- Block, N. (1996) What is Functionalism. On-line revised entry on functionalism. In D.M Borchert (Ed.) *The Encyclopaedia of Philosophy Supplement*. MacMillan.
<http://www.nyu.edu/gsas/dept/philo/faculty/block>.
- Boden, M. (Ed.) (1996) *The Philosophy of Artificial Life*, Oxford University Press, 1996.
- Boden, M. A. (1999) Is metabolism necessary? *British Journal for the Philosophy of Science*, **50**(2): 231—248.
- Boerlijst, M. & Hogeweg, P (1995) Attractors and Spatial Patterns in Hypercycles with Negative Interactions. *J. theor. Biol.* **176**: 199—210.
- Bonabeau, E. (1999) (Ed.) *Stigmergy. Special issue of Artificial Life Journal* **5**(2). MIT Press.
- Bonner, J.T. (1988) *The Evolution of Complexity by Means of Natural Selection*. Princeton University Press.
- Bonner, J.T. (2003) On the origin of differentiation. *Journal of Biosciences* **28**(4): 523—528.
- Bourgine, P & Stewart, J. (2004) Autopoiesis and Cognition, *Artificial Life* **10**(3):327—345.
- Braitenberg, V. (1984) *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press.
- Bray, D., Levin, M.D. & Lipkow K. (2007) The Chemotactic Behavior of Computer-Based Surrogate Bacteria, *Current Biology* **17**: 12—19.
- Bray, D., Matthew, L. & Lipkow, K. (2007) The Chemotactic Behavior of Computer-Based Surrogate Bacteria. *Current Biology* **17**:12—19.

- 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.
- Brenner, S. (1974) The genetics of *Caenorhabditis elegans*. *Genetics* **77**:71—94.
- Bressler, S.L. & Kelso, S.J.A. (2001) Cortical coordination dynamics and cognition. *Trends in Cognitive Sciences* **5**(1): 26—36.
- Broad, C. D. (1925) *The Mind and Its Place in Nature*. On-line version <http://www.ditext.com/broad/mpn/mpn.html> [downloaded 11-23-2007]
- Brooks, R.A. (1991) Intelligence without representation. *Artificial Intelligence* **47**:139—159.
- Brown, T. S. (1911) The intrinsic factors in the act of progression in the mammal. *Proc. Royal Soc. London, Series B* **84**:308—319.
- Burnet, F.M. (1959) *The Clonal Selection Theory of Acquired Immunity*. Cambridge: Cambridge University Press.
- Burzsáki, G. & Draguhn, A. (2004) Neuronal oscillations in cortical networks. *Science* **304**: 1926—1929.
- Campbell, D. T. (1974) Downward Causation. In Ayala, F. J. & Dobzhansky, T. (Eds.) *Hierarchically Organised Biological Systems*. Studies in the Philosophy of Biology, Berkeley, Los Angeles: University of California Press, pp. 179—186.
- Campbell, D.T. (1974) Evolutionary Epistemology. In P.A. Schilpp (ed.) *The Philosophy of Karl Popper*, pp. 413—463. Ashby, W.R. (1952) *Design for a Brain. The origin of adaptive behaviour*. Chapman and Hall, 1978 edition, pp. 413—463.
- Campbell, R.D., Josephson, R.K., Schwab, W.E. & Rushforth, N.B., (1976) Excitability of nerve-free hydra. *Nature* **262**:388—390.
- Canolty, R.T. et al. (2006) High gamma power is phase-locked to theta oscillations in human neocortex. *Science* **313**: 1626—1628.
- Carew, T.J. (2002) Understanding the consequences. *Nature* **417**: 803—806.
- Cartwright, N. (1983) *How the Laws of Physics Lie*. Oxford: Oxford University Press.
- Casti, J. (1997) *Would-be Worlds*. New York: John Wiley.
- Cechetti, D.F. & Topolovec, J.C. (2002) Cerebral Cortex. In Ramachandran, V.S. (Ed.) *Encyclopedia of the Human Brain Vol. 1.*, Academic Press, pp. 663—679.
- Chang, D.-I., Li, X.-C., Lou, X.-M., Lee, Y.-S., Kim, H.-K., Weiss, K., Kandel, E.R. & Kaang, B.-K. (1998) Transfection into *Aplysia* sensory neurons of an exogenous octopamine receptor coupled only to adenylyl cyclase allows genetic dissection of the relative contributions of PKA and PKC to short-term presynaptic facilitation. *Society of Neuroscience Abstracts* **24**:1190.
- Chemero, A. (2003) An outline of a theory of affordances. *Ecological Psychology* **15**: 181—195.
- Chemero, A. (2008) *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.
- Chemero, A. & Turvey, M. T. (2008) Autonomy and Hypersets. *BioSystems* **91**(2):320—330.
- Chialvo, D.R. (2004) Critical brain networks. *Physica A* **340**:756—765.
- Chiel, H. J., & Beer, R. D. (1997) The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in Neurosciences* **20**:553—557.
- Chomsky, N. (1965) *Aspects of a theory of syntax*. Cambridge, MA: MIT Press.

- Christensen, W. (2007) Evolutionary origins of volition, In Spurrett, D., Kincaid, H., Ross, D. & Stephens, L. (eds.) *Distributed Cognition and the Will: Individual Volition and Social Context*, Cambridge, Mass.: MIT Press, pp. 255—288. <http://cogprints.org/4831/>
- Christensen, W. (1999) *Interactivist-Constructivist Approach to Adaptive Intelligence and Agency*. PhD thesis; University of Newcastle, Australia. Available at: <http://www.newcastle.edu.au/departement/pl/staff/WayneChristensen/dissertation.htm>
- Christensen, W. & Hooker, C. (2000) Autonomy and the emergence of intelligence: Organised interactive construction. *Communication and Cognition - Artificial Intelligence* 17(3-4):133—157.
- Christensen, W.D. & Bickhard, M.H. (2002) The Process Dynamics of Normative Function. *Monist* 85 (1): 3—28.
- Church, A. (1936) An Unsolvable Problem of Elementary Number Theory. *American Journal of Mathematics* 58: 345—363.
- Churchland, P. S. (2002) *Brain-Wise: Studies in Neurophilosophy*. Cambridge, MA: MIT Press
- CIA (1983) *Human Resource Exploitation Training Manual*. Last retrieved 25 of March 2008: <http://www.gwu.edu/~nsarchiv/NSAEBB/NSAEBB27/02-01.htm>.
- Clancey, W.J. (1989) The Frame of Reference Problem in Cognitive Modeling. *Proceedings of 11th Annual Conference of the Cognitive Science Society*, Ann Arbor. Lawrence Erlbaum Associates, pp.107—114.
- Clark, A. (1996). Happy couplings: Emergence and explanatory interlock. In Boden, M. (Ed.) *The Philosophy of Artificial Life*. Oxford: Oxford University Press, pp. 262—281.
- Clark, A. (1997) *Being there. Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.
- Clark, A. & Chalmers, D. (1998) The extended mind. *Analysis* 58:10—23.
- Clark, A., & Thornton, C. (1997) Trading spaces: Computation, representation, and the limits of uninformed learning. *Behavioral and Brain Sciences* 20: 57—90.
- Cliff, D. T. (1991) Computational Neuroethology: A Provisional Manifesto. In Meyer, J. A. & Wilson, S.W., (Eds.) *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior (SAB'91)*, MIT Press, pp.29—39.
- Collier, J. (1988) Supervenience and Reduction in Biological Hierarchies. In Matthen, M. & Lin-sky, B. (Eds.) *Philosophy and Biology: Canadian Journal of Philosophy Supplementary Volume* 14:209—234.
- Collier, J. (1999). Autonomy and Process Closure as the Basis for Functionality. In Chandler, J.L.R. & van de Vijver, G., (Eds.), *Closure: Emergent Organizations and their Dynamics*. Volume 901 of the New York Academy of Sciences.
- Collier, J. (2004) Self-organization, Individuation and Identity, *Revue Internationale de Philosophie* 59: 151—172.
- Collier, J. D. & Hooker, C. A. (1999) Complexly organised dynamical systems. *Open Systems and Information Dynamics* 6(3), 241-302.
- Copeland, B.J. & Proudfoot, D. (1999) Alan Turing's forgotten ideas in computer science. *Scientific American*, April 1999: 98—103.
- Cornish-Bowden, A., Cárdenas, M. L., Letelier, J.C., Soto-Andrade, J., & Guíñez Abarzúa, F. (2004) Understanding the parts in terms of the whole. *Biology of the Cell* 96: 713—717.

- Crick, F. & Koch, C. (1990) Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2: 263—275.
- Csendes, T. (1984) A simulation study on the chemoton. *Kybernetes* 13 (2): 79.
- Cummins, R. (1975) Functional Analysis. *The Journal of Philosophy* 72(20):741—765.
- Daley, A.J., Girvin, A., Kauffman, S.A., Wills, P.R., & Yamins, D. (2002) Simulation of chemical autonomous agents. *Z. Phys. Chem.* 216:41—49.
- Damasio, A.R. (1994) *Descartes' Error. Emotion, Reason and the Human Brain*. New York: G.P. Putnam's Sons.
- Dawkins (1976/2006) *The selfish gene*. Oxford University Press, 30th Anniversary edition.
- Dennett, D. (1987) *The intentional stance*. Cambridge, MA: MIT Press.
- Dennett, D. (1994) Artificial Life as Philosophy. *Artificial Life* 1(3): 291—292.
- Dennett, D. (1995) *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon & Schuster.
- Descartes, R. (1637/1996) *Discourse on the Method of Rightly Conducting the Reason and Seeking for Truth in the Sciences*. Weissman, D (Ed.) Descartes, R. *Discourse on the Method and Meditations on First Philosophy*, Yale University, NY, 1996.
- Descartes, R. (1972) *Treatise of man*. Cambridge, MA: Harvard University Press.
- Dewey, J. (1893) The Reflex Arc Concept in Psychology. *Psychological Review* 3:357—370.
- Dewey, J. (1914) Psychological doctrine in philosophical teaching. *Journal of Philosophy* 11:505—512.
- Di Paola, V., Marijuán, P.C. & Lahoz-Beltra, R. (2004) Learning and evolution in bacterial taxis: an operational amplifier circuit modeling the computational dynamics of the prokaryotic 'two component system' protein network. *BioSystems* 74: 29—49.
- 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. (2003) Organismically-inspired robotics: homeostatic adaptation and natural teleology beyond the closed sensorimotor loop. In Murase, K., Asakura, T. (Eds.), *Dynamical Systems Approach to Embodiment and Sociality*. Advanced Knowledge International, Adelaide, pp. 19—42.
- Di Paolo, E.A. (2005) Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences* 4:97—125.
- Di Paolo, E.A., Noble, J. & Bullock, S. (2000) Simulation models as opaque thought experiments. In Bedau, M.A., MacCaskill, J.S., Packard, N.H. & Rasmussen, S. (Eds.) *Artificial Life VII: The Seventh International Conference on the Simulation and Synthesis of Living Systems*, MIT Press, pp. 497—506.
- Di Paolo, E.A. & Iizuka, H. (2008) How (not) to model autonomous behaviour. *Biosystems* 91(2): 409—423.
- Dreyfus, H. L. (1972) *What Computers Can't Do: A Critique of Artificial Reason*. Harper and Row.
- 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.
- Dunn, N. A., Pierce-Shimomura, J. T., Conery, J. S. & Lockery, S. R. (2004) A Neural Network Model of Chemotaxis Predicts Functions of Synaptic Connections in the Nematode *C. elegans*. *Journal of Computational Neuroscience* **17**(2):137—147.
- Dunn, N.A. Pierce-Shimomura, J.T. Conery, J.S. Lockery, S.R. (2006) Clustered Neural Dynamics Identify Motifs for Chemotaxis in *Caenorhabditis elegans*. In *International Joint Conference on Neural Networks*, pp. 547—554.
- Dunn, N.A., Conery, J.S. & Lockery, S.R. (2007) Circuit Motif for Spatial Orientation Behaviors Identified by Neural Network Optimization. *Journal of Neurophysiology* **98**: 888—897 DOI:10.1152/jn.00074.2007
- Durbin, R.M. (1987) *Studies on the Development and Organisation of the Nervous System of Caenorhabditis elegans*. Ph.D. Cambridge University.
- Dyson, F. (1982) A model for the origin of life. *J. Mol. Evol.* **18**: 344—350.
- Dyson, F. (1999) *Origins of Life*. Cambridge, MA: Cambridge University Press, Second edition.
- Eagleman, M. (2004) The where and when of intention. *Science* **303**:1144—1146.
- Ebert, A. & Dyck, M.J. (2003) The experience of mental death: The core feature of complex posttraumatic stress disorder. *Clinical Psychology Review* **24**: 617—635.
- Edelman, G. (1987) *Neural Darwinism. The Theory of Neuronal Group Selection*. New York: Basic Books.
- Edelman, G.M. (1999) Building a Picture of the Brain. *Ann. N.Y. Acad. Sci.* **882**: 68—89.
- Edelman, G.M. (1987) *Neural Darwinism: the Theory of Neuronal Group Selection*. Basic Books.
- Edelman, G.M. (1989). *The remembered present*. Basic Books.
- Edelman, G.M. & Tononi, G. (2000) *A universe of consciousness. How matter becomes imagination*.
- Eigen, M., Gardiner, W., Schuster, P. & Wincker-Oswatitch, R. (1981) The origin of genetic information. *Scientific American* **244**(4): 88—118.
- Elman, J.L., Bates, E., Johnson, M.H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: a connectionist perspective on development*. Cambridge, MA: MIT Press.
- Emmeche, C. (1994) *The Garden in the Machine*. Princeton University Press.
- Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, **2**:704—716.
- Engel, A.K. & Singer, W. (2001) Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences* **5**(1):16—25.
- Etxeberria, A., Moreno, A., Umerez, J. (Eds.) (2000) Special issue on the contribution of artificial life and the sciences of complexity to the understanding of autonomous systems. *CCAI: Communication a& Cognition - Artificial Intelligence* **17**(3/4): 111—254.
- Faumont, S. and Lockery, S. (2006) The awake behaving worm: simultaneous imaging of neuronal activity and behavior in intact animals at millimeter scale. *Journal of Neurophysiology* **95**(3):1976—81.
- Fernando, C.T., & Rowe, J. (2007) Natural selection in chemical evolution. *Journal of Theoretical Biology* **247**: 152—167.

- Fernando, C.T., & Rowe, J. (2008) The origin of autonomous agents by natural selection. *Biosystems* **91**(2):355—373.
- Ferrée, T.C., Marcotte, B.A. & Lockery, S.R. (1997) Neural network models of chemotaxis in the nematode *Caenorhabditis elegans*. *Advances in Neural Information Processing Systems* **9**:55—61.
- Field, R.J. & Noyes, R.M. (1974) Oscillations in Chemical Systems IV. Limit cycle behavior in a model of a real chemical reaction, *J. Chem. Phys.* **60**:1877-1884.
- Finlay, B.L. and Darlington, R.B. (1995) Linked regularities in the development and evolution of mammalian brains. *Science* **268**: 1578—1584.
- Fodor, J.A. (1968) *Psychological explanation*. New York: Random House.
- Fodor, J.A. (1975) *The language of thought*. New York: Thomas Y. Crowell.
- Fodor, J.A. and Pylyshyn, Z. (1988) Connectionism and Cognitive Architecture: a Critical Analysis. *Cognition* **28**:3—71.
- Ford, J.M., Krystal, J.H. & Mathalon, D.H. (2007) Neural Synchrony in Schizophrenia: From Networks to New Treatments. *Schizophrenia Bulletin* **33**(4):848—852.
doi:10.1093/schbul/sbm062
- Franklin, S. & Graesser A. (1996) Is it an Agent, or Just a Program? A Taxonomy for Autonomous Agents *Proceedings of the Workshop on Intelligent Agents III, Agent Theories, Architectures, and Languages* Lecture Notes In Computer Science; Vol. 1193, pp:21—35.
- Freeman, W. (1997a) Nonlinear neurodynamics of intentionality. *Journal of Mind and Behavior* **18**(2-3): 291—304.
- Freeman, W. (1997b) Three centuries of category errors in studies of the neural basis of consciousness and intentionality. *Neural Networks* **10**: 1175—1183.
- Freeman, W. (2000) *How Brains Make Up Their Minds*. Columbia University Press.
- Freeman, W. (2005) Emotion is from preparatory brain chaos; irrational action is from premature closure. *Behavioral and Brain Sciences* **28** (2): 204—205.
- Freeman, W.J., Kozma, R. & Werbos, P. J. (2001) Biocomplexity: adaptive behavior in complex stochastic dynamical systems. *BioSystems* **59**(2):109—123.
- Fries, P. (2005) A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences* **9**(10):474—480.
- Fries, P, Nicolic, D. & Singer, W. (2007) The gamma cycle. *Trends in Neurosciences* **30**(7):309—316. doi:10.1016/j.tins.2007.05005
- Friston, K.J. (2000) The labile brain (I, II & III). *Phil. Trans. R. Soc. Lond. B* **355**: 215—265.
- 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.
- Gabor Miklos, G. L., Campbell, K. S. W., & Kankel, D. R. (1994) The rapid emergence of bio-electronic novelty, neuronal architectures, and organismal performance. In Greenspan, R.J. (Ed.) *Flexibility and Constraint in Behavioural systems*, John Wiley and Sons, pp. 269—293.

- Gánti, T. (1971) *The Principle of Life*. 1st edition. (In Hungarian) Gondolat, Budapest.
- Gánti, T. (1984/1991) *The Chemoton Theory. Vol I: Theoretical Foundations of Fluid Machines*. OMIKK, Budapest, 1984.
- Gánti, T. (2002) On the early evolutionary origin of biological periodicity. *Cell Biology International* **26** (8): 729—735.
- García Azkonobieta, T. (2005) *Evolución, desarrollo y (auto)organización. Un estudio sobre los principios filosóficos de la evo-devo*. Tesis Doctoral, Universidad del País Vasco, - Euskal Herriko Unibertsitatea, San Sebastián.
- Ghysen, A. (2003) The origin and evolution of the nervous system. *International Journal of Developmental Biology* **47**:555—562.
- Gibson, J.J. (1979) *The ecological approach to visual perception*. Boston: Houghton-Mifflin.
- Gibson, K.R. (1991) Myelination and behavioral development: A comparative perspective on questions of neoteny, altriciality, and intelligence. In Gibson, K. R. & Petersen, A. (Eds.) *Brain maturation and cognitive development: Comparative and cross-cultural perspectives*. Hawthorne, NY: Aldine de Gruyter, pp. 29—64
- Giere, R. (1979/1997) *Understanding Scientific Reasoning*. Harcourt Brace College Publishers, 4th edition 1997.
- Giere, R. (1999) *Science without laws*. University of Chicago Press.
- Giere, R. (2004) How Models Are Used to Represent Reality. *Philosophy of Science* **71**: S742—S752.
- Gilbert, S.F. (2001) Ecological Developmental Biology: Developmental Biology Meets the Real World. *Developmental Biology* **233**:1—12.
- Gilbert, S.F. (2003) *Developmental Biology*. Sinauer Associates.
- Glennan, S. (1996) Mechanisms and the nature of causation. *Erkenntnis* **44**: 49–71.
- Glennan, S. (2002) Rethinking mechanistic explanation. *Philosophy of Science* **69**: S342—S353.
- Glennan, S. (2005) Modeling Mechanisms *Stud. Hist. Phil. Biol. & Biomed. Sci.* **36**: 443—464.
- Godfrey-Smith, P. (1996) *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.
- Godfrey-Smith, P. (2002) Environmental Complexity and the Evolution of Cognition. In R. Sternberg and J. Kaufman (eds.) *The Evolution of Intelligence*. Lawrence Erlbaum Associates, pp.233—249.
- Godfrey-Smith, P. (2005) Folk Psychology as a Model. *Philosophers' Imprint* **5**(6). <www.philosophersimprint.org/005006/>
- Goethe, J.W. (2002) *Goethe y la Ciencia. Una antología de los escritos científicos de Goethe*. Siruela.
- Goodman, M.B. (2006) Mechanosensation (January 06, 2006), WormBook, ed. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.62.1, <http://www.wormbook.org>.
- Goodwin, B., Holder, N. & Wyllie, C. C., (1983) *Development and evolution*. Cambridge: Cambridge University Press.
- Grassé P.P. (1959). La reconstruction du nid et les coordinations inter-individuelles chez *Bellicositermes natalensis* et *Cubitermes* sp. La theorie de la stigmergie: Essai d'interpretation des termites constructeurs. *Ins. Soc.*, **6**: 41—83.

- Grassian, S. (1993) *Psychiatric effects of solitary confinement*. Public report based on declaration submitted in September 1993 in Madrid v. Gomez, 889FSupp.1146. http://www-prisoncommission.org/statements/grassian_stuart_long.pdf
- Grassian, S., & Friedman, N. (1986) Effects of Sensory Deprivation in Psychiatric Seclusion and Solitary Confinement. *International Journal of Law and Psychiatry* **8**:49—65.
- Grey-Walter, W. (1950) An imitation of life. *Scientific American*, **182**(5): 42—45.
- Griffiths, P.E. & Stotz, K. (2002) How the mind grows: a developmental perspective on the biology of cognition. *Synthese* **122**: 29—51.
- Habermas, J. (1981/1987) *The Theory of Communicative Action*. Cambridge, MA: Polity.
- Haken, H. (1977) *Synergetics - An introduction: Nonequilibrium phase transitions and self-organization in physics, chemistry and biology*. Springer Verlag.
- Haney, C. (2003) Mental Health Issues in Long-Term Solitary and “Supermax” Confinement. *Crime & Delinquency* **49**(1):124—156. DOI: 10.1177/0011128702239239
- Harnad, S. (1990) The symbol grounding problem. *Physica D* **42**:335—346.
- Harold, F. M. (2001) *The way of the cell. Molecules, Organisms and the Order of Life*. Oxford University Press.
- Haselager, W. & Gonzalez, M.E. (2005) The meaningful body: On the differences between artificial and organic creatures. In Loula, A, Gudwin, R. & Queiroz, J. (Eds) *Artificial Cognition Systems*. Hershey, PA: IDEA Group Inc., pp. 238—250.
- Hebb, D. O. (1958). The Motivating Effects of Exteroceptive Stimulation *American Psychologist* **13**(3): 109—113.
- Held, R. and Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology* **56**(5):872—876.
- Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science* **15**:137—175.
- Hendriks-Jansen (1996) *Catching ourselves in the act: situated activity, interactive emergence, evolution, and human thought*. Cambridge, MA: MIT Press.
- Herrmann, C. S., Munk, M.H.J. & Engel, A.K. (2004) Cognitive functions of gamma-band activity: memory match and utilization. *Trends in Cognitive Sciences* **8**(8): 347—355.
- Hilborn, R.C. (1994) *Chaos and Nonlinear Dynamics*. Oxford University Press.
- Hinton, G. E. & Nowlan, S. J. (1987) How learning can guide evolution. *Complex Systems* **1**: 495—502.
- Hobert, O. (2006) Specification of the nervous system (August 8, 2005), WormBook, ed. The *C. elegans* Research Community WormBook, doi/10.1895/wormbook.1.12.1, <http://www.wormbook.org>.
- Hoffmeyer, Jesper. (1996): *Signs of Meaning in the Universe*. Bloomington: Indiana University Press.
- Holland, J. H. (1992) *Adaptation in Natural and Artificial Systems*. Cambridge, MA: MIT Press.
- Holland, O. (2003) Exploration and high adventure: the legacy of Grey Walter. *Phil. Trans. R. Soc. Lond. A* **361**: 2085—2121.
- Holroyd, C. B. & Coles, M. G. H. (2002) The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review* **109**:679—709.)

- Hooker, C. A. (1995) *Reason, Regulation and Realism: Toward a Naturalistic, Regulatory Systems Theory of Reason*. Albany, NY: State University of New York Press.
- Hopfield J.J. (1982) Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci.*, **79**: 2554—2558.
- Horchler, A., Reeve, R., Webb, B. & Quinn, R. (2004) Robot phonotaxis in the wild: a biologically inspired approach to outdoor sound localization. *Advanced Robotics* **18**(8):801—816.
- Horgan, J. (1995) From Complexity to Perplexity. *Scientific American* **272**:104—109.
- Horridge, G.A. (1968) The origin of the nervous system. In Burne G.H. (Ed.) *The structure and function of the nervous tissue*, vol. 1: 1—31.
- Huang, B. & Pitelka, D. R. (1973) The contractile process in ciliate. *Stentor coeruleus*. *The Journal of Cell Biology*, **57**:704—728.
- Humphreys, P. (2002) Computational models. *Philosophy of Science* **69**: S1—S11.
- Hunter, L. (1993) Molecular biology for computer scientists. In L. Hunter (Ed.) *Artificial Intelligence and Molecular Biology*. AAAI Press, Ch. 1, pp. 1—46.
- Hutchins, E. (1995) *Cognition in the Wild*. Cambridge, MA: MIT Press.
- Iizuka, H. & Di Paolo, E. (2007) Toward Spinozist Robotics: Exploring the Minimal Dynamics of Behavioral Preference. *Adaptive Behavior* **15**(4): 359—76.
- Ikegami, T. & Suzuki, K. (2008) From a homeostatic to a homeodynamic self. *BioSystems* **91**(2):388—400.
- Inoue, M. & Kaneko, K. (2006) Condition for intracellular adaptive dynamics for chemotaxis. *Physical Review E* **74** (1): 011903.
- Izhikevich, E.M. (2007) *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. MIT Press.
- Izquierdo-Torres, E. & Di Paolo, E. A. (2005) Is an Embodied System Ever Purely Reactive? In Capcarrere, M.S., Freitas, A.A., Bentley, B.J., Johnson, C.G. & Timmis, J. (Eds.) *Advances in Artificial Life. Proceedings of the Eighth European Conference (ECAL 2005)*, Lecture Notes in Computer Science series, Springer Verlag, pp. 252—261.
- Izquierdo-Torres, E. and Harvey, I. (2007) Hebbian Learning using Fixed Weight Evolved Dynamical 'Neural' Networks. In H. A. Abbass et al. (eds.), *Proceedings of the First IEEE Symposium on Artificial Life. (IEEE-ALife'07)*, IEEE Press, pp. 394—401.
- Jacob, P. (2003) Intentionality. *Stanford Encyclopedia*. <http://plato.stanford.edu/entries/intentionality/>
- Jen, E. (2002) Stable or Robust? What is the difference? On-line paper: <http://www.santafe.edu/~erica/stable.pdf>
- Jensen, O. & Colgin, L. L. (2007) Cross-frequency coupling between neuronal oscillations. *Trends in Cognitive Sciences* **11**(7):267—269.
- Jensen, O., Kaiser, J. & Lachaux, J-P. (2007) Human gamma-frequency oscillations associated with attention and memory. *Trends in Neurosciences* **30**(7):317—24.
doi:10.1016/j.tins.2007.05.001
- Jerison, H.J. (2002) Evolution of the Brain. In Ramachandran, V. S. (Ed.) *Encyclopedia of the Human Brain, Vol. II*, Academic Press, pp. 251—267.
- Johnson, M. H. (1997) *Developmental cognitive neuroscience: An introduction*. Oxford: Blackwell

- Johnson, M. H. (2001) Functional brain development in humans. *Nature Reviews Neuroscience* **2**:475—483.
- Jonas, H. (1966/2001) *The phenomenon of life: Toward a philosophical biology*. Evanston, IL: Northwestern University Press.
- Jonas, H. (1968) Biological Foundations of Individuality. *International Philosophical Quarterly* **8**: 231—251.
- Kampis, G. (1991) *Self-modifying systems in biology and cognitive science*. Pergamon Press.
- Kandel, E.R. (2001) The Molecular Biology of Memory Storage: A Dialogue Between Genes and Synapses. *Science* **294**:1030—1038.
- Karmiloff-Smith, A. (2005) Ontogeny Genetics, and Evolution: A Perspective from Developmental Cognitive Neuroscience. *Biological Theory* **1**(1): 44—51.
- Karmiloff-Smith, A. & Thomas, M. (2005) Can Developmental Disorders Be Used to Bolster Claims From Evolutionary Psychology? A Neuroconstructivist Approach. In Taylor-Parker, S., Langer, J. and Milbrath, C. (Eds.) *Biology and Knowledge Revisited. From neurogenesis to psychogenesis*. Lawrence Elbawm. pp. 307—321.
- Kass-Simon, G. and Scappaticci, A.A. (2002) The behavioral and developmental physiology of nematocysts. *Canadian Journal of Zoology* **80**:1772—1794.
- Kauffman, S. (1986) Autocatalytic Sets of Proteins. *Journal of Theoretical Biology* **119**: 1—24.
- Kauffman, S. (1993): *The origins of order: Self-organization and selection in evolution*. Oxford: Oxford University Press.
- Kauffman, S. (2000) *Investigations*. Oxford University Press.
- Kauffman, S.A. (1971) Cellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems. *Journal of Cybernetics* **1**: 71—96.
- Kaufman A, Dror G, Meilijson I, Ruppin E (2006) Gene expression of *Caenorhabditis elegans* neurons carries information on their synaptic connectivity. *PLoS Computational Biology* **2**(12):e167. DOI:10.1371/journal.pcbi.0020167
- Keijzer, F. (2005) Theoretical behaviorism meets embodied cognition: two theoretical analyses of behavior. *Philosophical Psychology* **18**(1):123—143.
- Keijzer, F. (2006) Differentiating animality from agency: Towards a foundation for cognition. In R. Sun & N. Miyake (Eds.) *Proceedings of CogSci/ICCS 2006*, Alpha NJ: Sheridan Printing, pp.1593—1598.
- Keller, E. (2007) The disappearance of function from ‘self-organizing systems’. In Boogerd, F, Bruggeman, F, Hofmeyr, J.-H., & Westerhoff, H.V. (Eds.) *Systems Biology. Philosophical Foundations*, Amsterdam: Elsevier, Amsterdam, pp. 303—317.
- Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.
- Kelso, J.A. (1984) Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology - Regulation Integration and Comparative Physiology* **246**: R1000—R1004.
- Kiontke, K. and Sudhaus, W. Ecology of *Caenorhabditis* species (January 09, 2006), The *C. elegans* Research Community (Ed.), *WormBook*, doi/10.1895/wormbook.1.37.1, <http://www.wormbook.org>.
- Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P. (1983) Optimization by simulated annealing. *Science* **220**:671—680.

- Kitano, H. (2001) Systems Biology: Toward System-level Understanding of Biological Systems. In Kitano, H. (Ed.) *Foundations of Systems Biology*. Cambridge MA: MIT Press, pp. 1—38.
- Kitano, H. (2004) Biological Robustness. *Nature Review Genetics* **5**:826—837
- Klipp, E., Herwig, R., Kowald, A., Wierling, C. and Lehrach, H. (2005) *Systems biology in practice*. Wiley-VCH.
- Knuuttila, T. (2005) *Models as Epistemic Artefacts: Toward a Non-Representationalist Account of Scientific Representation*. Prima Oy, University of Helsinki.
- Kohler, I. (1962) Experiments with goggles. *Scientific American* **206** (5): 62—72.
- Kohler, I. (1964) The formation and transformation of the perceptual world. *Psychological Issues* **3**: 1—173.
- Kolb, B. & Whishaw, I.Q. (2005) *An Introduction to Brain and Behavior*. Second edition. Worth Publishers.
- Kolmogorov, A.N. (1965) Three Approaches to the Quantitative Definition of Information. *Problems of Information Transmission* **1**(1): 1—7.
- Koopowitz, H., & Keenan, L. (1982) The primitive brains of platyhelminthes. *Trends in Neurosciences* **5**(3):77—79.
- Kovas, Y. & Plomin, R. (2006) Generalist genes: implications for the cognitive sciences. *Trends in Cognitive Sciences* **10** (5): 198—203.
- La Cerra, P. & Bingham, R. (1998) The adaptive nature of the human neurocognitive architecture: An alternative model, *Proc. Natl. Acad. Sci. USA* **95**: 11290—11294.
- Laland K. N., Odling-Smee J. & Feldman M. W. (1999) Evolutionary consequences of niche construction and their implications for ecology. *Proceeding of the Nationall Academy of Sciences USA* **96**:10242—10247.
- Langton, C. (1989) Artificial Life. In Langton, C. (Ed.), *Artificial Life*. Redwood City CA: Addison-Wesley, pp.1—47.
- Langton, C. (1990) Computation at the edge of chaos. *Physica D* **42**(1-3):12—37.
- Langton, C.G. (1989) Artificial life. In Langton, C. G. (Ed.) (1989) *Artificial life*. Reading, MA: Addison-Wesley, pp.1—44.
- Lasagna G. and Germoglio M. (2002) Isolation and sensory deprivation - clinical observations following eight years of collaboration. *Annals of Burns and Fire Disasters* **15** (4):195.
- Levi-Montalcini, R. (2001) *La galaxia mente*. Crítica.
- Lewis, E. B. (1992) Clusters of master control genes regulate the development of higher organisms. *Journal of the American Medical Association*, **267**: 1524—1531.
- Lewis, M.D. (2005a) Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and Brain Sciences* **28**: 169—194.
- Lewis, M.D. (2005b) Self-organizing individual differences in brain development. *Developmental Review* **25**(3-4):252—277.
- Lewontin, R. (2000) *The triple helix*. Harvard University Press.
- Llinás, R. (1988) The intrinsic electrophysiological properties of mammalian neurons: insights into central nervous system function. *Science*, **242**(4886): 1654—1664.
- Llinás, R.R. (2001) *The I of the vortex. From neuron to self*. Cambridge, MA: MIT Press.

- Llinás, R., Ribary, U., Contreras, D. & Pedroarena, C. (1998) The neuronal basis for consciousness. *Philosophical Transactions of the Royal Society of London, Series B*, **353**:1841–1849.
- Loh, M., Rolls, E.T. & Deco, G. (2007) A dynamical systems hypothesis of schizophrenia. *PLoS Comput Biol* **3**(11): e228. doi:10.1371/journal.pcbi.0030228
- Lorenz, E. (1963) Deterministic nonperiodic flow. *Journal of Atmospheric Sciences* **20**:130–141.
- Lorenz, K. (1977) *Behind the Mirror*. London: Methuen.
- Luisi, P. L. (2003) Autopoiesis: a review and a reappraisal. *Naturwissenschaften* **90**(3): 49–59.
- Lungarella, M. and Sporns, O. (2006) Mapping information flow in sensorimotor networks. *PLoS Computational Biology* **2**(10): e144. DOI: 10.1371/journal.pcbi.0020144
- Luu, P., Tucker D.M. & Makeig, S. (2004) Frontal Midline Theta and the Error-Related Negativity: Neurophysiological Mechanisms of Action Regulation. *Clinical Neurophysiology* **115**(8):1821–1835.
- Lyon, P. (2006) The biogenic approach to cognition. *Cognitive Process* **7**:11–29.
- Mackie, G.O. (1990) The elementary nervous system revisited. *American Zoologist* **30**(4):907–920.
- Mackie, G.O. (2004b) Central Neural Circuitry in Jellyfish *Aglantha*. *Neurosignals* **13**:5–19.
- Mackie, G.O. & Meech, R.W. (1985) Separate sodium and calcium spikes in the same axon. *Nature* **313**:791–793.
- Maes, P. (1994) Modelling Adaptive Autonomous Systems. *Artificial Life* **1**:135–162.
- Maes, P. (Ed.) (1990) *Designing Autonomous Agents*. Cambridge, MA: MIT press.
- Margulis, L. (1992) *Symbiosis in Cell Evolution: Microbial Communities in the Archean and Proterozoic Eons*, W.H. Freeman.
- Margulis, L. & Sagan, D. (1995). *What is Life?* New York: Simon and Schuster.
- Markram, H. (2006) The Blue Brain Project. *Nature Reviews Neuroscience* **7**: 153–160.
- Mattick, J.S. (2004). The hidden genetic program of complex organisms. *Scientific American*, October 2004: 61–67.
- Maturana H.R. (1970) Biology of Cognition. In Maturana & Varela (1973/1980) *Autopoiesis and Cognition*. Reidel.
- Maturana, H.R. (2002) Autopoiesis, structural coupling and cognition: A history of these and other notions in the biology of cognition. *Cybernetics and Human Knowing* **9**: 5–34.
- 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.
- Mavelli, F. & Ruiz-Mirazo, K. (2007) Stochastic simulations of minimal self-reproducing cellular systems. *Philosophical Transactions of the Royal Society of London B* **362**(1486):1789–1802.
- Maynard-Smith, J. & Szathmari, E. (1995) *The Major Transitions in Evolution*. Freeman.
- Mayr, E. (1963) *Animal species and evolution*. Cambridge, MA: Harvard University Press.
- McCoy, A. (2006) Invisible in Plain Sight: CIA Torture Techniques Go Mainstream. *Amnesty International Magazine*.
http://www.amnestyusa.org/magazine/invisible_in_plain_sight.html

- McCulloch, W. S. and Pitts, W. H. (1943). A logical calculus of the ideas immanent in the nervous system. *Bulletin of Mathematical Biophysics* **5**:115—133.
- McFarland, D. (1985) *Animal behaviour: psychobiology, ethology and evolution*. Essex: Longman Scientific & Technical.
- McGregor, S. & Fernando, C. (2005) Levels of Description: A Novel Approach to Dynamical Hierarchies. *Artificial Life* **11**: 459—472.
- Merleau-Ponty, M. (1942) *La structure du comportement*. Presses Universitaires de France.
- Meyer-Lindenberg A. (1996) The evolution of complexity in human brain development: an EEG study. *Electroencephalography and Clinical Neurophysiology* **99** (5): 405—411.
- Meyer-Lindenberg A., Ziemann, U., Hajak, G., Cohen, L. & Berman, K.F. (2002) Transitions between dynamical states of differing stability in the human brain. *PNAS* **99**(17): 10948—10953.
- Millikan, R.G. (1984) *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT press.
- Mills, R. & Watson, R. A. (2005) Genetic assimilation and canalization in the Baldwin effect. In Capcarrere, M. et al. (Eds.) *Proceedings of the Eighth European Conference on Artificial Life*, Springer Verlag, pp. 353—362.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002) Network Motifs: Simple Building Blocks of Complex Networks. *Science* **298**: 824—827.
- Mitchell, M. (1996) *An Introduction to Genetic Algorithms*. Cambridge MA: MIT Press.
- Mitchell, M. (2006) Complex systems: Network thinking. *Artificial Intelligence* **170**(18):1194—1212.
- Moreno, A. & Barandiaran, X. (2004) A Naturalized Account of the Inside-Outside Dichotomy. *Philosophica* **73**: 11—26.
- Moreno, A. & Lasa, A. (2003) From Basic Adaptivity to Early Mind. *Evolution and Cognition*, **9** (1):12—30.
- Moreno, A. & Umerez, J. (2000) 'Downward Causation' at the Core of Living Organization. Andersen, P.B., Emmeche, C., Finnemann, N.O. & Christiansen, P.V. (Eds.) *Downward Causation*. Aarhus: Aarhus University Press, pp. 99—117.
- Moreno, A., Etxeberria, A. & Umerez, J. (2008) The autonomy of biological individuals and artificial models. *BioSystems* **91**(2):309—319.
- Moreno, A., Ruiz-Mirazo, K. & Barandiaran, X. (2008) The impact of the paradigm of complexity on the foundational frameworks of biology and cognitive science. In Hooker, C. (Ed.) *Philosophy of Complex Systems* (Vol. 10 of *Handbook of the Philosophy of Science*), New York: Elsevier, to appear.
- Mori, I. (1999) Genetics of chemotaxis and thermotaxis in the nematode *Caenorhabditis elegans*. *Annual Review Genetics* **33**:399—422.
- Morowitz H.J. (1999) A theory of biochemical organization, metabolic pathways, and evolution. *Complexity* **4**:39-53.
- Morowitz, H. J. (1992) *Beginnings of Cellular Life*. New Haven: Yale University Press.
- Morrison, M. (2000) *Unifying Scientific Theories*. Cambridge: Cambridge University Press.
- Moss, L. (2006) Redundancy, Plasticity, and Detachment: The Implications of Comparative Genomics for Evolutionary Thinking, *Philosophy of Science* **73**:930—946.

- Müller, G. B. (2003) Embryonic motility: environmental influences and evolutionary innovation, *Evolution and Development*, **5**(1): 56—60.
- Munakata, Y., Casey, B.J. & Diamond, A. (2004) Developmental Cognitive Neuroscience: progress and potential. *Trends in Cognitive Sciences*, **8**(3): 122—128.
- Neibur, E. & Erdős, P. (1991) Theory of locomotion of nematodes: Dynamics of undulatory progression on a surface. *Biophysics Journal* **60**:1132—1146.
- Newell, A. (1980) Physical Symbol Systems. *Cognitive Science* **4**: 135—183.
- Nicolis, G. & Prigogine, I. (1977) *Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order through Fluctuations*. Wiley-Interscience.
- Niebur, E. & Erdős, P. (1993) Theory of the locomotion of nematodes: control of the somatic motor neurons by interneurons. *Mathematical Biosciences* **118**:51—82.
- Nithianantharajah, J. & Hannan, A.J. (2006) Enriched environments, experience-dependent plasticity and disorders of the nervous system. *Nature Review Neuroscience* **7**: 697—710.
- Orgel, L. (2000) Self-organizing biochemical cycles. *Proc. Natl. Acad. Sci. USA* **97**(23): 12503—12507.
- Parker, G. H. (1919) *The elementary nervous system*. Lippincott, Philadelphia.
- Pattee, H.H. (1972) Laws and constraints, symbols and languages. In Waddington, C.H. (Ed.), *Towards a Theoretical Biology 4, Essays*. Edinburgh University Press, Edinburgh, pp. 248—258.
- Pattee, H.H. (1973) The physical basis and origin of hierarchical control. In Pattee, H.H. (Ed.), *Hierarchy Theory: The Challenge from Complex Systems*. New York: Braziller, pp.71—108.
- Pessoa, L. (2008) On the relationship between emotion and cognition. *Nature Reviews Neuroscience* **9**: 148—158, doi: 10.1038/nrn2317
- Petitot, J., Varela, E.J., Pachoud, B., and Roy, J.-M. (Eds.) (1999) *Naturalizing Phenomenology*. Stanford, CA: Stanford University Press.
- Pfeifer, R. & Scheier, C. (1999) *Understanding Intelligence*. Cambridge, MA: MIT Press.
- Piaget, J. (1967 /1969) *Biología y Conocimiento. Ensayo sobre las relaciones entre las regulaciones orgánicas y los procesos cognoscitivos*. Siglo veintiuno editores.
- Pinker, S. (1997) *How the Mind Works*. Penguin Books.
- Plaetzer, K., Thomas, S. R., Falkner, R. & Falkner, G. (2005) The microbial experience of environmental phosphate fluctuations. An essay on the possibility of putting intentions into cell biochemistry *Journal of Theoretical Biology* **235**(4):540—554.
- Poincaré, H. (1890) Sur le problème des trois corps et les équations de la dynamique. *Acta Mathematica* **13**: 1-270.
- Poincaré, H. (1985) L'espace et la géométrie. *Revue the métaphysique et de morale* **3**:631—646.
- Polanyi, M. (1968) Life's irreducible structure. *Science* **160**:1308—1312.
- Port, R. & van Gelder, T. (1995) *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.
- Prescott, T.J. (2007) Forced moves or good tricks in design space? Landmark on the evolution of action selection. *Adaptive Behavior* **15**(1): 9—31.
- Prigogine, I. & Stengers, I. (1984) *Order out of Chaos*. Bantam Books, New York.

- Pross, A. (2004) Causation and the origins of life. Metabolism or replication first? *Origins of Life and Evolution of the Biosphere* **34**: 307—321.
- Purves, D., Augustine, G.J., Fitzpatrick, D., Hall, W. C., Lamantia, A.-S., McNamara, J.O., & Williams, S.M. (Eds.) (2004). *Neuroscience*. Third Edition. Cambridge MA: Sinauer Associates, Inc.
- Putnam, H. (1960) Minds and Machines. In Hook, S. (Ed.) *Dimensions of Mind*, pp.148—179. New York: New York University Press.
- Quartz, S.R. (1999) The constructivist brain. *Trends in Cognitive Sciences* **3** (2):48—57.
- Quartz, S.R. & Sejnowski, T.J. (1997) The neural basis of cognitive development: a constructivist manifesto, *Behavioural and Brain Sciences* **20**: 537—596.
- Rankin, C.H. (2004) Invertebrate Learning: What Can't a Worm Learn? *Current Biology* **14**:R617—R618, DOI 10.1016/j.cub.2004.07.044
- Rankin, C.H., Beck, C.D.O., and Chiba, C.M. (1990) *Caenorhabditis elegans*: A new model system for the study of learning and memory. *Behavioral Brain Research* **37**:89—92.
- Ray, T. (1992) An approach to the synthesis of life. In Langton, C.G., Taylor, C., Farmer, J.D. & Rasmussen, S. (Eds.) *Artificial life II. Proceedings of the Workshop on Artificial Life*, pp. 325—371.
- Ray, T. S. (1994) An evolutionary approach to synthetic biology: Zen and the art of creating life. *Artificial Life* **1**(1-2): 195—226.
- Reigl, M., Alon, U. & Chklovskii, D.B. (2004) Search for computational modules in the *C. elegans* brain. *BMC Biology* 2004, 2:25. <http://www.biomedcentral.com/1741-7007/2/25>
- Robinson, S.R. (2005) Behavioral embryology. Hopkins, B. (Ed.) *The Cambridge Encyclopaedia of Child Development*. Cambridge University Press. pp.: 469—473.
- Rodriguez, E., George, N., Lachaux, J.P., Martinerie, J., Renault, B. & Varela, F.J. (1999) Perception's shadow: long- distance synchronization of human brain activity. *Nature* **397**:430—433.
- Rohde, M. & Di Paolo, E. (2005) “Value Signals” and Adaptation: An Exploration in Evolutionary Robotics. *CCNR Report* 584. <ftp://ftp.informatics.sussex.ac.uk/pub/reports/csrp/csrp584.pdf>
- Rose, J. K. & Rankin, C. H. (2001) Analyses of Habituation in *Caenorhabditis elegans*. *Learning & Memory* **8**:63—69.
- Rosen, R. (1958) A relational theory of biological systems. *Bulletin of Mathematical Biophysics*. **20**: 245—341.
- Rosen, R. (1991) *Life itself. A comprehensive enquiry into the nature, origin and fabrication of life*. New York: Columbia University Press.
- Rosenblatt, F (1958) The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review* **65** (6): 386—408.
- Rosenblueth, A., Wiener, N., & Bigelow, J. (1943) Behavior, purpose and teleology. *Philosophy of Science* **10**(1):18—24.
- Rosslénbroich, B. (2005) The evolution of multicellularity in animals as a shift in biological autonomy. *Theory in Biosciences* **123**:243—262.
- Rosslénbroich, B. (2006) The notion of progress in evolutionary biology – the unresolved problem and an empirical suggestion. *Biology and Philosophy* **21**:41—70

- Roth, G. and Dicke, U. (2005) Evolution of brain and intelligence. *Trends in Cognitive Science* **9** (5): 250—257.
- Ruiz-Mirazo, K and Moreno, A. (2000) Searching for the roots of autonomy: The natural and artificial paradigms revisited. *Communication and Cognition-Artificial Intelligence*, **17**(3–4): 209—228.
- Ruiz-Mirazo, K. (2001) *Condiciones físicas para la aparición de sistemas con capacidades evolutivas abiertas*. [Ph. D. Thesis: *Physical conditions for the appearance of autonomous systems with open-ended evolutionary capacities*] San Sebastián: University of the Basque Country.
- Ruiz-Mirazo, K. & Mavelli, F. (2008) On the way towards ‘basic autonomous agents’: Stochastic simulations of minimal lipid–peptide cells. *BioSystems* **91**(2):374—387.
- Ruiz-Mirazo, K. & Moreno, A. (2004) Basic Autonomy as a Fundamental Step in the Synthesis of Life. *Artificial Life*, **10**:235—259.
- Ruiz-Mirazo, K. Umerez, J. & Moreno, A. (2008) Enabling conditions for ‘open-ended evolution’. *Biology & Philosophy* **23**:67–85. DOI 10.1007/s10539-007-9076-8.
- Rumelhart, D.E., McClelland, J.L., & the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol.1: Foundations*. Cambridge, MA: MIT Press.
- Russel, E.S. (1916) *Form and Function*. A contribution to the history of animal morphology. London, John Murray. Available at <http://www.gutenberg.org>.
- Russell, S.J. & Norvig, P. (1995) *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.
- Ryle, G. (1949) *The concept of mind*. University of Chicago Press.
- Sanyal, S., Wintle, R.F., Kindt, K.S., Nuttley, W.M., Arvan, R., Fitzmaurice, P., Bigras, E., Merz, D.C., Hebert, T.E., van der Kooy, D., et al. (2004). Dopamine modulates the plasticity of mechanosensory responses in *Caenorhabditis elegans*. *The EMBO Journal* **23**:473–482.
- Schaffner, K. F. (2000) Behavior at the organismal and molecular level: The case of *C. elegans*. *Philosophy of Science* **67**: S273—S288.
- Searle, J.R. (2004) *Mind. A brief introduction*. Oxford University Press.
- Searle, J.R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* **3**(3):417—457.
- Segre, D., Ben-Eli, D. & Lancet, D. (2000) Compositional genomes: prebiotic information transfer in mutually catalytic noncovalent assemblies. *Proc. Natl. Acad. Sci. USA* **97**(8): 4112—4117.
- Seth, A. (2002) Agent-based modelling and the environmental complexity thesis. In Hallam, B., Floreano, D., Hallam, J., Hayes, G. & Meyer, J-A. (Eds.) *Proceedings of the seventh international conference on simulation of adaptive behavior on From animals to animats*, MIT Press, pp. 13—23.
- Seth, A. K. (2007) Measuring Autonomy by Multivariate Autoregressive Modeling. In Almeida e Costa, F et al. (Eds.) *Proceedings of the 9th European Conference on Artificial Life (ECAL 2007)*, Springer Verlag, pp. 475—484.
- Seth, A.K. & Edelman, G.M. (2004) Environment and behavior influence the complexity of evolved neural networks. *Adaptive Behavior* **12**(1):5—21.
- Shapiro, R. (2000) A replicator was not involved in the origin of life. *IUBMB Life* **49**(3):173—176.

- Sherrington (1940) *Man on his Nature*. Cambridge University UK: Cambridge University Press. 2nd edition, 1951.
- Simon, H.A. (1969) *The Sciences of the Artificial*. Cambridge MA: MIT Press, 3rd edition 1996.
- Simons, E.J. (1981) The role of electricity in plant movements. *The New Phytologist* **87**:11—37.
- Sismondo, S. (1999) Models, Simulations, and Their Objects. *Science in Context* **12**(2):247—260.
- Skarda, C.A. & Freeman, W.J. (1987) How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences* **10**:161–195.
- Skinner, B. F. (1938) *The Behavior of Organisms: An Experimental Analysis*. New Jersey: Prentice Hall.
- Smith, L. B. & Thelen, E. (2003) Development as a dynamical system. *Trends in Cognitive Sciences*, **7**(8): 343—348.
- Smithers, T. (1995) Are autonomous agents information processing systems? In Steels, L. and Brooks, R.A. (Eds.) *The artificial life route to artificial intelligence: Building situated embodied agents*. New Haven: Erlbaum.
- Smithers, T. (1997) Autonomy in Robots and Other Agents. *Brain and Cognition* **34**, 88—106.
- Solé, R.V., Munteanu, A., Rodriguez-Caso, C. & Macía, J. (2007) Synthetic protocell biology. *Philosophical Transactions of the Royal Society of London B* **362**:1727—1739.
- Spiegelman, S. (1967) An in vitro analysis of a replicating molecule. *American Scientist* **55**:221—264.
- Spiegelman, S. (1971) An approach to the experimental analysis of precellular evolution. *Quarterly Review of Biophysics* **4**(2): 213—253.
- Sporns, O. (2004) Complex neural dynamics. In Jirsa, V.K. and Kelso, J.A.S., (Eds.) *Coordination Dynamics: Issues and Trends* Springer-Verlag pp. 197—215.
- Sporns, O., Tononi, G. & Edelman, G.M. (2000) Theoretical Neuroanatomy: Relating Anatomical and Functional Connectivity in Graphs and Cortical Connection Matrices. *Cerebral Cortex* **10**:127—141.
- Squire, L. R., F. E. Bloom, S. K. McConnell, J. L. Roberts, N. C. Spitzer & M. J. Zigmond (Eds.) (2003) *Fundamental neuroscience*. San Diego, CA: Elsevier Academic Press.
- Staddon, J. E. R. (2003) *Adaptive Behavior and Learning*. Internet Edition: <http://psychweb-psych.duke.edu/departments/jers/abl/TableC.htm>
- Steels, L. (1991). Towards a Theory of Emergent Functionality. In Meyer, J. and Wilson, R., (Eds.), *Simulation of Adaptive Behaviour*. MIT Press, pp. 451—461.
- Steels, L. (1995) The Artificial Life Roots of Artificial Intelligence. In Langton, C. (Ed.) *Artificial Life: An Overview*, Cambridge, MA: MIT press, pp. 75—110.
- Sterelny, K. (2001) *The Evolution of Agency and Other Essays*, Cambridge University Press, Cambridge, UK.
- Stewart, J. (1996) Cognition=Life. Implication for higher-level cognition. *Behavioral Processes* **35**:311–326.
- Stock, J.B. & Levit, M. (2000) Signal transduction: Hair brains in bacterial chemotaxis. *Current Biology* **10**: R10—R14.

- Striedter, G.F. (2003) Epigenesis and evolution of brains: From Embryonic Divisions to Functional Systems. In Müller, G and Newman, S. *Origination of Organismal Form: Beyond the Gene in Developmental and Evolutionary Biology*, pp. 287—303.
- Striedter, G.F. (2006) Précis of *Principles of Brain Evolution*. Behavioral and Brain Sciences, *BBS* **29**: 1—36.
- Strogatz, S.H. (2001) Exploring complex networks. *Nature* **410**: 268—277.
- Sugiyama, T. & Fujisawa, T. (1978) Genetic analysis of developmental mechanisms in Hydra. II. Isolation and characterization of an interstitial cell-deficient strain. *J. Cell Sci.* **29**: 35—52.
- Suzuki, R. and Arita, T. (2004). Drastic Changes in Roles of Learning in the Course of Evolution. In Pollack, J. B., Bedau, M., Husbands, P., Ikegami, T. & Watson, R. A. (Eds.) *Proceedings of Artificial Life IX*, MIT press, pp. 369—374.
- Taft, R.J., Pheasant, M. and Mattick, J. (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays* **29**: 288—299.
- Taylor, R. (1950a) Comments on a Mechanistic Conception of Purposefulness. *Philosophy of Science* **17**(4):310—317.
- Taylor, R. (1950b) Purposeful and Non-Purposeful Behavior: A Rejoinder. *Philosophy of Science* **17** (4): 327—332.
- Thelen, E. & Smith, L. (1994) *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.
- Thelen, E. & Smith, L. B. (1994) *A Dynamic Systems Approach to the Development of Cognition and Action*, MIT Press, Cambridge, MA.
- Thompson, E. (2007) *Mind in Life. Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.
- Thomson, E. (2004) Life and mind: From autopoiesis to neurophenomenology. A tribute to Francisco Varela. *Phenomenology and the Cognitive Sciences* **3**: 381—398.
- Tononi, G. & Edelman, G.M. (1998) Consciousness and Complexity. *Science* **262**: 1846—1851.
- Tononi, G., Edelman, G., & Sporns, O. (1998) Complexity and coherency: integrating information in the brain. *Trends in Cognitive Science*, **2**(12):474—484.
- Tononi, G., Sporns, O. & Edelman, G.M. (1996) A complexity measure for selective matching of signals by the brain. *Proc. Nat. Ac. Sci.* **93**:3422—3427.
- Trewavas, A. (2003) Aspects of Plant Intelligence. *Annals of Botany* **92**: 1—20. doi:10.1093/aob/mcg101
- Tsuda, I. (2001) Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems, *Behavioral and Brain Sciences* **24**(5):793—847.
- Tuci, E., Harvey, I. & Quinn, M. (2002) Evolving integrated controllers for autonomous learning robots using dynamic neural networks. In Hallam, B., Floreano, D., Hallam, J., Hayes, G. & Meyer, J-A. (Eds.) *From Animals to Animats 7 - Proceedings of The Seventh International Conference on the Simulation of Adaptive Behavior (SAB'02)*, pp. 4—9.
- Tucker, D. M., Derryberry, D., & Luu, P. (2000). Anatomy and physiology of human emotion: Vertical integration of brainstem, limbic, and cortical systems. In J. C. Borod (Ed.), *The neuropsychology of emotion*. London: Oxford University Press.
- Turing, A.M. (1936) On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* **2**(42): 230—265.

- Turing A.M. (1952) The chemical basis of morphogenesis. *Philos. Trans. R. Soc. Lond. B.* **237**:37—72.
- Umerez, J. (2001) Howard Pattee's theoretical biology: a radical epistemological stance to approach life, evolution and complexity. *BioSystems* **60**:159—177.
- Umerez, J. & Moreno, A. (1995) Origin of Life as the First MST Control Hierarchies and Inter-level Relation. *World Futures* **45**(2):139—154.
- Umerez, J., Etxeberria, A. & Moreno, A. (1993) Emergence and Functionality. Ramaekers, J. (Ed.) *XIII International Congress on Cybernetics*, Namur: International Association of Cybernetics, pp. 519—523.
- Uttal, W.R. (2001) *The new phrenology*. MIT Press.
- Van Duijn, M., Keijzer, F.A. & Franken, D. (2006) Principles of minimal cognition: Casting cognition as sensorimotor coordination. *Adaptive Behavior*, **14**(2):157—170.
- van Gelder, T. (1998) The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences* **21**(5):615—628.
- Varela, F. (1979) *Principles of biological autonomy*. New York: Elsevier.
- Varela, F. (1992) Autopoiesis and a Biology of intentionality, In McMullin, B. & Murphy, N. (eds.) *Autopoiesis & Perception*. pp. 1—14. *Proceedings of a Workshop held in Dublin City University, August 25th and 26th 1992*. School of Electronic Engineering Technical Report, Dublin, 1994.
- Varela, F. (1992) Autopoiesis and a Biology of intentionality. In McMullin, B. & Murphy, N. (eds.) *Autopoiesis & Perception*. pp. 1—14. *Proceedings of a Workshop held in Dublin City University, August 25th and 26th 1992*. School of Electronic Engineering Technical Report, Dublin, 1994.
- Varela, F., Lachaux, J.P., Rodriguez, E. & Martinerie, J. (2001) The Brain Web: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience* **2**: 229—239.
- Varela, F., Maturana, H. & Uribe, R. (1974) Autopoiesis: The Organization of Living Systems, its characterization and a model. *BioSystems* **5**:187—196.
- Varela, F.J. (1997) Patterns of life: intertwining identity and cognition. *Brain and Cognition* **34**:72—87.
- Vidal Miranda, P. (2005) *La fin d'une hiérarchie fonctionnelle: assemblées neuronales et organisation de l'information en perception et mémoire visuelles*. Thèse de doctorat de l'Université de Paris 6.
- Volterra, A. & Meldolesi, J. (2005) Astrocytes, from brain glue to communicating elements: The revolution continues. *Nature Reviews Neuroscience*, advanced online publication:2—16. DOI: 10.1038/nrn1722.
- von Foerster, H. (1974) Cybernetics of Epistemology. In Keidel, W.D., Handler, W. & Spring, M. (eds.) *Kybernetik und Bionik* (vol 2). Munich, Oldenburg. Reprinted in: Foerster, H. (2003) *Understanding Understanding. Essays on Cybernetics and Cognition*. New York, Springer Verlag. (NB. Page numbers refer to the reprinted version.)
- von Uexküll, J. (1940/1982). *Bedeutungslehre*. (Bios 10. Johann Ambrosius Barth, Leipzig), [translated by Thure von Uexküll, 1982: The theory of meaning. *Semiotica* **42**(1): 25—82]
- Waddington, C. H. (1957) *The Strategy of the Genes*. London: George Allen & Unwin.
- Walter, W. G. (1950) An imitation of life. *Scientific American* **182**:42—45.

- Watson, J. B. (1914) *Behavior: An introduction to comparative psychology*. Henry Holt.
- Weaver, W. (1948) Science and Complexity. *American Scientist* **36**(4):536—544.
- Weber, A. and Varela, F. J. (2002) Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences* **1**: 97—125.
- Weber, B. H. & Depew, D. J. (2003) (Eds.) *Evolution and Learning. The Baldwin Effect reconsidered*. Cambridge, MA: MIT Press.
- Werner, G. (2006) Metastability, criticality and phase transitions in brain and its models. *BioSystems* **90**(2): 496—508.
- Wheeler, M. (2005) *Reconstructing the cognitive world. The next step*. Cambridge, MA: MIT Press.
- White, J.G., Southgate, E., Thomson, J.N. and Brenner S. (1986) The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Phil. Trans. R. Soc. Lond. B* **314**:1—40.
- Wicken, J.S. (1987) *Evolution, thermodynamics and information. Extending the Darwinian program*. Oxford: Oxford University Press.
- William, J. (1890) *The Principles of Psychology Vol. I*, Henry Holt.
- Wimsatt, W.C. (1974) Complexity and Organization. In R. Buck and R.S. Cohen (eds.), *PSA-1972, Boston Studies in Philosophy Science* **20**: 67—86.
- Wolkenhauer, O. (2006) *Systems Biology. Dynamic Pathway Modelling*. http://www.sbi.uni-rostock.de/dokumente/t_sb.pdf
- 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.
- 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.
- Wright, W.G. (2000) Neuronal and Behavioral Plasticity in Evolution: Experiments in a Model Lineage, *BioScience*, **50** (10): 883—894.
- Yaeger, L.S. & Sporns, O. (2006) Evolution of Neural Structure and Complexity in a Computational Ecology, In Rocha, L. et al. (Eds.) *Proceedings of Artificial Life X*, MIT Press, pp.330—336.
- You, Y., Kim, J., Cobb, M. and Avery, L. (2006) Starvation activates MAP kinase through the muscarinic acetylcholine pathway in *Caenorhabditis elegans* pharynx. *Cell Metabolism* **3**:237—245.
- Zaikin, A. N. and Zhabotinsky, A. M. (1970) Concentration wave propagation in a two-dimensional, liquid-phase self oscillating system. *Nature* **225**:535—537.
- Zhang, Y., Lu, H. and Bargmann, C. I. (2005) Pathogenic bacteria induce aversive olfactory learning in *Caenorhabditis elegans*. *Nature* **438**: 179—184. DOI:10.1038/nature04216
- Ziemke, T. (2001) The Construction of 'Reality' in the Robot: Constructivist Perspectives on Situated Artificial Intelligence and Adaptive Robotics. *Foundations of Science* **6**(1): 163—233.
- Ziemke, T. (2005) Cybernetics and Embodied Cognition: On the Construction of Realities in Organisms and Robots. Special issue on Heinz von Foerster (editor: Alex Riegler), *Kybernetes*, **34**(1-2):118—128.

Zubek, J.P. (Ed.) (1969). *Sensory Deprivation: Fifteen Years of Research*. New York: Appleton-Century-Crofts.

This book has been written and edited with
OpenOffice Write under X-evian, Debian and
Ubuntu GNU/Linux operating systems, using
fonts Bitstream Vera Charter, Nimbus Times
Roman & Gara.