THE UNIVERSITY *of* EDINBURGH

The Biological Roots of Cognition and the Social Origins of Mind

Autopoietic Theory, Strict Naturalism and Cybernetics

Mario Villalobos

PhD in Philosophy

University of Edinburgh

August 2014

To my family

Abstract

This thesis is about the ontology of living beings as natural systems, their behavior, and the way in which said behavior, under special conditions of social coupling, may give rise to mental phenomena. The guiding questions of the thesis are: 1) What kinds of systems are living beings such that they behave the way they do? 2) How, through what kinds of mechanisms and processes, do living beings generate their behavior? 3) How do mental phenomena appear in the life of certain living beings? 4) What are the natural conditions under which certain living beings exhibit mental phenomena? To answer these questions the thesis first assumes, then justifies and defends, a Strict Naturalistic (SN) stance with respect to living beings. SN is a metaphysical and epistemological framework that, recognizing the organizational, dynamic and structural complexity and peculiarity of living beings, views and treats them as metaphysically ordinary natural systems; that is, as systems that, from the metaphysical point of view, are not different in kind from rivers or stars. SN holds that if in natural sciences rivers and stars are not conceived as semantic, intentional, teleological, agential or normative systems, then living beings should not be so conceived either. Having assumed SN, and building mainly on the second-order cybernetic theories of Ross Ashby and Humberto Maturana, the thesis answers question 1) by saying that living beings are (i) adaptive dynamic systems, (ii) deterministic machines of closed transitions, (iii) multistable dissipative systems, and (iv) organizationally closed systems with respect to their sensorimotor and autopoietic dynamics. Based on this ontological characterization, the thesis answers question 2) by showing that living beings' behavior corresponds to the combined product of (i), (ii), (iii) and (iv). Points (i) and (ii) support the idea that living beings are strictly deterministic systems, and that, consequently, notions such as information, control, agency or teleology—usually invoked to explain living beings' behavior—do not have operational reality but are rather descriptive projections introduced by the observer. Point (iii) helps to understand why, despite their deterministic nature, living beings behave in ways that, to the

observer, appear to be teleological, agential or "intelligent". Point (iv) suggests that living beings' sensorimotor dynamics are closed circuits without inputs or outputs, where the distinction between external and internal medium is, again, an ascription of the observer rather than a functional property of the system itself. Having addressed the basic principles of living beings' behavior, the thesis explores the possible origin of (truly) mental phenomena in the particular domain of social behavior. Complementing Maturana's recursive theory of language with Vygotsky's dialectic approach the thesis advances, though in a still quite exploratory way, a sociolinguistic hypothesis of mind. This hypothesis answers questions 3) and 4) by claiming that the essential properties of mental phenomena (intentionality, representational content) appear with language, and that mind, as a private experiential domain, emerges as a dialectic transformation of language.

Contents

Declaration of Authorship

I, Mario Villalobos, declare that this thesis is my own work and has not been submitted for any other degree or professional qualification.

Mario Villalobos

31st August 2014

Note on Publications

Parts of this thesis have been published, or are forthcoming, in the following articles:

Villalobos, M. (2012). Machines, life and cognition: a second-order cybernetic approach. *Proceedings of 4th AISB Symposium on Computing and Philosophy, University of Birmingham* (pp. 41-47).

Villalobos, M. (2013a). Enactive cognitive science: revisionism or revolution? *Adaptive Behavior, 21* (3), 159-167. DOI: 10.1177/1059712313482953

Villalobos, M. (2013b). Autopoiesis, life, mind and cognition: bases for a proper naturalistic continuity. *Biosemiotics, 6* (3), 379-391. DOI: 10.1007/s12304-013-9174-8

Villalobos, M. & Ward, D. (2014). Living systems: autopoiesis, autonomy and enaction. *Philosophy and Technology*. DOI 10.1007/s13347-014-0154-y

Abramova, E. & Villalobos, M. (accepted). Stability and functional closure: the apparent (Ur)-intentionality of living beings and the game of content. *Philosophia*.

Acknowledgments

I would like to thank all the people who, in one way or another, to a greater or lesser extent, have contributed to make this work possible. Many thanks to Susan Oyama, Ezequiel Di Paolo, Tom Froese and Liz Swan for helpful and detailed comments on some parts of this thesis. Many thanks to Dan Hutto, Mark Bishop, John Protevi and Robert Rupert for occasional but illuminating philosophical discussions.

Many thanks to my supervisors Mark Sprevak, Dave Ward, Andy Clark and, at the beginning, Julian Kiverstein. With different styles and degrees of participation, but equal kindness and intellectual rigor, they have supported me in countless ways both academically and extra-academically. Mark and Dave have become, at least in my heart, something more than supervisors, which I take as an honor and a privilege.

Many thanks also to the Mind and Cognition Discussion Group, with Robert, Richard, Francesca and Andrea. Their support has been crucial during my years in Edinburgh. From outside philosophy, Richard Shillcock, Alex Papadopoulos-Korfiatis and Ekaterina Abramova have been a great intellectual company too.

I am very grateful to Becas Chile Conicyt, which has fully funded my doctorate studies.

Many thanks to Simone Magni and his comrades, my political family.

Many thanks to Tito, Pacha, Ivan and Hector Anselmo, my original family.

Above all, many thanks to my wife Loreto Maturana for being here, there and everywhere, and for the two little treasures she brought to my life, Simon and Martin.

Introduction

This thesis is about the ontology of living beings as natural systems, their behavior, and the way in which said behavior, under special conditions of social coupling, may give rise to mental phenomena.

The guiding questions of the thesis are:

1) What kinds of systems are living beings such that they behave the way they do?
2) How do mental phenomena appear in the life of certain living beings?

The thesis answers the first question by claiming that livings are:

1) Adaptive dynamic systems
2) Deterministic machines of closed transitions
3) Multistable dissipative systems, and
4) Organizationally closed systems with respect to their autopoietic dynamic and sensorimotor activity

The thesis answers the second question by claiming that mental phenomena, essentially understood as intentional and representational phenomena, appear with language, and that language, in turn, emerges as a recursive phenomenon in the domain of communicative behaviors of third-order biological systems.

To elaborate these answers the thesis first assumes, then justifies and defends, a Strict Naturalistic stance with respect to living beings. Strict Naturalism is a metaphysical and epistemological framework that, recognizing the organizational, dynamic and structural complexity and peculiarity of living beings, views and treats them as metaphysically ordinary natural systems; that is, as systems that, from the metaphysical

point of view, are not different in kind from rivers or stars. Strict Naturalism holds, basically, that if in natural sciences the behavior of rivers and stars is not explained by appealing to semantic, intentional, teleological, agential or normative elements, then living beings' behavior should not be so explained either.

Cybernetic allies

I am not alone in this enterprise. My allies are two cyberneticists whose works, in my opinion, have set bases for a Strictly Naturalistic understanding of living beings, their behavior and adaptation. They are Ross Ashby and Humberto Maturana.

Ross Ashby and Humberto Maturana are present, in one way or another, in almost everything that is said in this thesis. They have become to me, for better or worse, not only allies but models of how to approach those phenomena that interest me, elaborate the concepts and terminology, formulate the questions, and build the answers.

After reading and studying them with effort and dedication over these years, I wish I could have assimilated a bit of their analytic power, their conceptual depth and intellectual courage. Most likely this has not been the case, but more important to me is that I have found in them a way of viewing living beings, including human beings, that I think is worth exploring, developing and deepening.

This thesis might be viewed, in a sense, as an invitation to reconsider the cybernetic tradition not from a merely historical point of view, but rather as a living source of ideas and theoretical tools that may help us to illuminate, with a different light, certain areas, topics and problems in cognitive sciences.

Why (old fashioned) Cybernetics and not (more contemporary) Dynamical Systems Theory?

Dynamical Systems Theory (DST) has usually been identified as the contemporary, updated and improved, offspring of cybernetics (van Gelder,

1998; Grush, 1997). If this is the case, why in this thesis do we stick with cybernetics instead of using 'new brand' DST?

The cybernetic research program, considered as a whole, contains two main streams: first-order cybernetics and second-order cybernetics (Dupuy, 2009; Müller and Müller, 2007). First-order cybernetics basically corresponds to the study of living beings and artificial systems in terms of dynamical systems, most of the time, through the use of mathematical formalisms (the canonical example here is Wiener, 1948). Second-order cybernetics, instead, corresponds to an epistemological reflection about the observer (the scientist, the cyberneticist) and her descriptive/explanatory practices (von Foerster, 2003). A good part of second-order cybernetics, especially in Maturana's work, has to do with identifying and preventing certain descriptive or explanatory fallacies that are more or less recurrent in the study of living beings. As von Foerster points out, whereas first-order cybernetics is the study and theory of the observ*ed* systems, second-order cybernetics is the study and theory of the observ*ing* systems (von Foerster, 2003).

DST constitutes, in many aspects, a development and sophistication of first-order cybernetics, and in that sense, effectively, one can see it as an updated version of cybernetics. Nonetheless, DST, at least until now, has almost entirely disregarded the epistemological work developed in second-order cybernetics. There is nothing in DST, or at least I have not seen anything, like an explicit and systematic epistemological theorization about the observer and her descriptive/explanatory practices; i.e., anything like a Second-Order Dynamical Systems Theory.

The epistemological reflection concerning the observer and her descriptive/explanatory practices is, as we shall see, a central component of this thesis. A good part of our research problem has to do with examining the way in which the observer (i.e., the cognitive scientist) approaches living beings (observed systems), frames her descriptions and elaborates her explanations. A basic strategy in our research, as we shall see, is to try to reveal the observer-relative character of more or less popular notions and explanatory constructs in cognitive sciences (e.g., intentionality, internal representation, teleology).

DST, thus, takes just a half of the cybernetic tradition, whereas in this thesis we need the entire armory.

Another relatively absent aspect in DST that plays a key role in our research is the *metaphysical* analysis of dynamic systems. DST has a certain tendency to focus on the effectiveness and parsimony of its explanatory models, without taking too much care about the metaphysical aspects. DST develops useful and very sophisticated mathematical models for describing, explaining and predicting some behaviors in biological and artificial systems, and, as long as these models work, does not seem terribly worried about the metaphysical specifications of the modeled systems.

For example, for DST, the structural dynamic of a concrete system may be either deterministic or stochastic (van Gelder, 1995b); there seems to be no particular (explicit, justified) metaphysical position on this point. For us it is, on the contrary, very important to specify the way, whether deterministic or not, in which dynamic systems undergo their changes of state.

Ashby and Maturana take, make explicit, and defend a *deterministic* metaphysical position with respect to dynamic systems in general and living beings in particular. They, and I with them, as we shall see in Chapter 2, think that a deterministic conception of living beings is crucial to properly explaining and understanding their behavior.

DST has also displayed, I think for the same reason, a rather liberal or ambiguous attitude toward the cognitive notion of internal (neural) representation (van Gelder and Port, 1995; Kelso, 1995). The partial hostility DST shows towards this notion comes, mainly, from methodological considerations rather than from an explicit and systematic metaphysical analysis of dynamic systems.

In this thesis we will reject the notion of internal (neural) representation as a way of explaining living beings' behavior, but we will do it on the basis of ontological and metaphysical arguments. We will not say that internal representations do not have explanatory value because they are overly simplistic descriptive tools (van Gelder, 1995b), or do not add anything to our understanding (Chemero, 2000). We will not say that (minimal, weak) internal representations are dispensable because, from an epistemological point of view, they do not meet what Ramsey calls the "job-description

challenge" (Ramsey, 2007). We will not recommend avoiding the overuse of internal representations for the sake of some healthy scientific pragmatism (Haselager, 2004).

What we will hold, building on Maturana's metaphysical principle of structural determinism, is that internal representations are explanatory fictions that do not have biological reality.

These, I guess, are the main reasons why, for the specific purposes of this thesis, we have preferred good and old cybernetics rather than 'brand new' DST.

Style and relation to the literature

Ashby and Maturana are not philosophers (though many of their conceptual developments and theoretical reflections might be interpreted as philosophical works in their own right). Ashby is a cyberneticist who focuses, mainly, on the neurophysiology of adaptive behavior. Maturana is a neurophysiologist who develops, basically, a cybernetic theory of cognitive phenomena in general. Both of them are well trained in empirical research, and develop their respective theoretical systems in dialogue with experimental works.

The way in which they present their ideas and results is, however, to a large extent, axiomatic and theoretically self-contained. In Ashby these features are reinforced, in part, by his tendency to mathematical formalisms; in Maturana, perhaps, by his tendency to the creation of neologisms. When presenting their ideas, they do not compare or contrast them with the ideas of alternative approaches, and very rarely do they engage in direct discussions with rival theories. They proceed rather by building their own theoretical systems trying to preserve, as much as they can, a strict internal coherence in the conceptual structure of their explanatory models.

This is a philosophical thesis in the field of cognitive sciences, or so I want to think. However, inspired by the styles of Ashby and Maturana, many times I behave more as an outsider than a philosopher of cognitive science. During the exposition, there are many topics and discussions that might have been addressed by framing them within the established philosophical

literature, identifying the different positions about the points at issue, and placing, highlighting the contrasts, this thesis's position within that context. I have, however, deliberately chosen to do something different.

I have preferred to concentrate on exposing, as clearly as I can, with a considerable degree of detail, and allowing plenty of time for numerous examples and illustrations, the basic cybernetic concepts and intuitions that ground the philosophical ideas defended in this thesis.

The payoff, I want to believe, is that the reader will find a quite accessible presentation of views and theoretical constructions which, especially in the case of Maturana, sometimes are difficult to follow.

On the negative side, the thesis, because of this, somehow lacks the kind of dialectics that is more or less usual in philosophical works. Occasionally, I make reference to certain cognitive theories or approaches that might have played the role of 'philosophical opponents', but I do not engage in direct and open discussions. (I have done some of this dialectic exercise in separate papers. See "Notes on publications"). I hope, in future works, to do the job of explicitly confronting the main ideas of this thesis with the more or less established approaches and theories that compose the contemporary landscape of cognitive sciences.

Minimal caveat

The way in which I use and present the works of Ashby and Maturana in this thesis may give, at times, the impression that they are two authors who, surprisingly, agree about almost everything. That, of course, is not the case. Ashby and Maturana have very important points in common, but also several significant differences. In this thesis I only concentrate on the shared aspects, and, mainly for the sake of the exposition, I make them appear entirely lined up.

Outline of the thesis

The plan of the thesis is as follows. In Chapter 1, roughly, we present the research problem and the metaphysical and epistemological assumptions that

will be assumed throughout the thesis. We start, building on Maturana's second-order cybernetics, by talking about the observer and her explanatory practices in general. Then we concentrate on a particular explanatory system, namely the naturalistic system. There we introduce the notion of Strict Naturalism, and make explicit the way living beings will be conceived of, treated and studied in the thesis. The chapter ends by clarifying the precise sense in which living beings can be viewed as adaptive dynamic systems (point 1 in our ontological characterization of living beings).

Chapter 2 introduces, examines and justifies a series of metaphysical principles that are taken to be valid for dynamic systems in general and living beings in particular. We start by reviewing the precise sense in which living beings may be understood as 'machines', and illustrate, through a nice example provided by Ashby, the potential significance of this idea for certain explanatory practices in cognitive science (a point that will be fully developed in Chapter 5). Then we move on to argue that living beings are not only machines but *deterministic* machines. We pay special attention to the Maturanian 'principle of structural determinism' (PSD), and draw some important consequences for the study of living beings' behavior.

In Chapter 3 we focus on the apparent teleological character of living beings. We argue that said appearance is, ultimately, just that, an appearance, and that the real phenomenon behind it is a complex form of (deterministic) stability. The appreciation of living beings as far-from-equilibrium thermodynamic systems, and the Ashbyan notion of 'ultrastability' play a central role in this line of argument. They lead us to the conclusion that living beings are 'multistable dissipative systems' (point 3 in our ontological characterization), and that at least a good part of what we take to be their "purposeful" behavior corresponds in reality to complex forms of stability.

Chapter 4 addresses what is, perhaps, one of the most counterintuitive aspects of the thesis; the sensorimotor closure of living beings. First, we start by considering the notion of organizational closure at the level of the autopoietic dynamic of living beings, but, for reasons that are explained, we do not assume with respect to this point any special philosophical commitment. Things change when we examine the notion of closure at the

level of living beings' sensorimotor dynamic. Here we draw important and metaphysically loaded conclusions about perception and action in living beings. We argue that living beings' sensorimotor systems, including the nervous system (when they have one), work as closed systems wherein neither inputs nor outputs exist. We discuss and defend some philosophical consequences that follow from this view, and prepare the scenario for testing the explanatory power of our Strict Naturalistic conception of living beings.

Chapter 5 is a brief but important chapter. There we test, taking some representative cases of animal behavior, the explanatory power of our Strict Naturalistic conception of living beings. We conclude that at least a good part of living beings' behavior may be adequately explained in Strictly Naturalistic terms. That is, through explanatory constructions that, as is the case in standard natural science, do not need to appeal to intentional, semantic (representational), teleological, agential or normative elements.

Having reached this point, in Chapter 6 we face the question of the origin of mental phenomena. To answer this question we elaborate, in a rather exploratory and speculative way, a sociolinguistic hypothesis of mind. Building on Maturana's recursive theory of language, we argue that some of the essential properties of mental phenomena, such as intentionality and representational content, emerge with language, and that language, in turn, emerges as a recursive phenomenon in the communicative domain of social systems. In this view mind appears as an originally social phenomenon, whose private and individual dimension emerges as a process of dialectical transformation of language.

ally, a brief conclusion underlines the main philosophical points of the thesis and points out possible lines of research in the future.

Chapter 1

Metaphysical and epistemological prolegomena: some basic notions

In this chapter we are going to fix some basic metaphysical and epistemological notions that will help us to address our research problem in subsequent chapters. Most of what we are going to say about the behavior of living beings in this thesis depends on the general framework that we are going to offer here, so I want to be explicit and transparent about that. The reader must not take this chapter as a mere metaphysical preliminary, distant and not connected at all with the 'concretely cognitive' stuff. On the contrary, this chapter should be read rather as an invitation to construe the problem of cognition from a particular philosophical angle. Thus, although the chapter speaks of 'metaphysically abstract' things, almost nothing of what is said in it is philosophically innocent. If the reader is invited to see the problem of intelligent behavior and cognition in a certain way, it is because that way will frame and condition in turn the manner in which we are going to formulate the questions and construe the answers.

The metaphysical and epistemological framework that I will present here corresponds, to a large extent, to the metaphysical and epistemological framework that Maturana has built around his autopoietic theory through the years (Maturana, 1975, 1978, 1987, 1988, 1992, 2002, 2003). This framework is expanded, refined and developed in several ways, but always remaining loyal to the original spirit of Maturana's work, or so I honestly think. Since I consider that my conceptual contributions in this field are continuous and entirely compatible with Maturana's work, I do not flag them

as ‘mine’ in the text.

On more than one occasion the reader, especially the philosopher, will find that there are some connections, similarities or resonances between Maturana’s metaphysical frame and one or more philosophical systems already established in the literature. I am not going to do the job of flagging those similarities, and the reader is, of course, free to make the associations that she finds pertinent and useful for her understanding. But at the same time I recommend not to go too far with that kind of exercise. Maturana is a biologist who, although not ignorant or innocent about the history of metaphysics, has constructed his own conceptual system without situating it within the standard philosophical literature. This has given, for better or worse, enough room for commentators to make connections between Maturana’s system and certain allegedly akin philosophical schools. While some people connect Maturana’s metaphysics with Kant and some versions of contemporary pragmatist antirealism (Dougall, 2000, 1999), others put him in company of phenomenologists such as Heidegger or Merleau-Ponty (Dicks, 2011; Mingers, 1995). None of these associations, in my opinion, are necessary or particularly useful to grasp what is at play in Maturana’s metaphysical system. Many times, actually, they may be rather misleading. In this sense, I hope the chapter can speak for itself.

1.1 Observers, explanations and metaphysical frames: steps towards a Strict Naturalistic conception of living beings

We humans act as observers every time we perform an act of observation. An act of observation, in the Maturanian conception that I will follow here, corresponds basically to an attentional and linguistic act that operates upon a determinate experiential background. In the act of observation we attend to what happens or appears in our experiential field, and we do it by applying, implicitly or explicitly, one or more linguistic operations. By ‘experiential field’ I mean not only our perceptual experience of the external world but

also our internal experience, our subjectivity (feelings, thoughts, etc.). And by 'linguistic operations' I mean, roughly, psychological operations in which words or linguistic concepts[1] are required; e.g., categorization, description, judgment, inference, explanation, prediction, etc.

According to this definition, observation is an essentially linguistic act; therefore, reserved only for creatures with language. Non-linguistic creatures, even having a rich sensory life and a sophisticated behavioral repertory, do not count as observers. To the extent that the only creatures capable of language known hitherto—at least the only ones whose linguistic capacities we can verify directly and without ambiguity—are human beings, in this thesis, for all practical purposes, the expression 'observer' will always refer to a normally developed human being, to a person.

Our acts of observation—or observations, for short—vary with respect to a series of aspects and dimensions. Observations may be, in relative terms, more or less simple or complex, direct or indirect, subjective or objective, rigorous or careless, acute or superficial, spontaneous or controlled, etc. For example, in comparison with the standard scientific observation, which is rigorous, controlled and quite elaborated, our ordinary (commonsense) observations look rather simple, spontaneous and careless.

We have said that observations are essentially linguistic acts; therefore, operations mediated by concepts or words.[2] It is worth noting, though, that the central point in this idea is the conceptual *mediation* in itself, not the application of one or another concept in particular. For example, think of a man of an Amazonian tribe who has never had contact, direct or indirect, with any element of modern culture. When this man sees a laptop for the first time, surely he does not say "Ah, a laptop!" (not even in his own dialect). He lacks the concept LAPTOP, but that does not mean that his

[1] The qualification of 'linguistic' concepts is necessary because, for some philosophers, not all concepts are linguistic in nature (see for example Bermúdez, 2003). Here, nonetheless, every time we speak of concepts without further specifications, we will be solely referring to linguistic concepts; i.e., roughly, to the mental equivalents of words.

[2] This notion of observation sharply contrasts with van Fraassen's one, for whom 'observations' are basically unaided and non-conceptually mediated perceptual acts (see van Fraaseen, 1980).

perception is emptied of all conceptualization. The man takes the laptop in his hands, focuses the attention on it, inspects, examines it, etc., and after a couple of minutes he concludes that he has no idea of what that object can be. To conclude that, the man has gone through a line of reasoning in which, as a minimum, he has (i) compared the object with some known objects, (ii) wondered about its nature or possible functions, (iii) outlined some tentative answers, (iv) tested these hypotheses through the manipulation of the object, (v) evaluated the results, and (vi) finally decided that he has no idea of what the object can be. Yet this negative judgment, this recognition of ignorance, is far from being a conceptual vacuum. In his reasoning the Amazonian man has applied, presumably, concepts such as OBJECT (or THING), RECTANGULAR, HARD, SMOOTH SURFACE, UTILITY, etc., and in his conclusion, actually, has put the laptop under a complex concept that is highly abstract: UNKNOWN OBJECT.

Being an attentional act, the act of observation is also essentially an act of consciousness, of awareness. We cannot observe (conceptualize, examine, etc.) that which we are not aware of, that which is out of our attentional focus. What is left out of our attentional focus may, many times, interfere or modulate our observations (e.g., what psychologists study under the name of 'implicit memory', 'priming effect' and similar), but that is something different. This means that, although the observer is always a human being, the human being is not at every moment and in every respect an observer. In our daily life, actually, a fundamental part of our functioning occurs in non-observational terms. Those familiarized with the classical Heideggerian distinction between 'readiness-to-hand' and 'presence-at-hand' may, perhaps, appreciate this point in a relatively easy way. In our quotidian experience, we find ourselves 'immersed' in a series of actions whose realization does not require our conscious attention, or in which certain objects (tools) become 'phenomenologically transparent' to us. Such experiential condition is what Heidegger called 'readiness-to-hand'. But also we continuously face situations in which our attention and conscious thematization is required. This second condition, in which the world (some portion or aspect of it) appears as something separated from our experiential flow (as an object), is what Heidegger called 'presence-at-hand' (or in some

cases ‘un-readiness-to-hand’; a sort of attenuated version of ‘presence-at-hand’. See Wheeler, 2014). Without necessarily establishing a strict parallel with the Heideggerian scheme, one might say that, in our daily life, there is always a part of us that is simply ‘absorbed’ in the action, operating in non-observational terms (i.e., without conceptualization or conscious thematization), and another part that, in a ‘detached’ attitude, operates in observational terms.

Usually, when we speak of observation we tend to think of visual perception, of someone watching something (e.g., when we want to represent graphically an observer or an act of observation, generally we draw an eye). Nonetheless, the experiential base of an observation need not be visual. Any perceptual modality, any sensorial channel can provide enough experiential material for an act of observation. The act of observation, recall, is defined here essentially as a linguistic (conceptually mediated) attentional act. For example, the blind person that recognizes a laptop by touching it acts as an observer. The physician that, using a stethoscope, auscultates the heartbeats of a patient acts as an observer too. All professional tasters and smellers hired by the companies to test the quality of their products (chocolates, perfumes, etc.), act not only as observers but as expert observers. Even more, every time we examine introspectively our thoughts or feelings—whose psychological content, strictly speaking, we cannot see, hear, touch, taste or smell—we act as observers too. The only difference is that, in this case, our attention is directed at our mental life, our subjective experience.

1.1.1 Explanations, naturalism and naturalization

One of the most common practices in us observers is asking for and giving explanations. Explanations—viewed not from a logical viewpoint but rather from a *pragmatic linguistic* perspective—are basically answers to certain kinds of questions (van Fraassen, 1980; Achinstein, 1983). The questions

that ask for explanations usually start with interrogative adverbs such as "how" or "why", and may be formulated both explicitly and implicitly. When we ask for an explanation, what we want to know is essentially the "how" (the mechanism, the process, the functioning) or the "why" (the causes, the reasons) of something (a phenomenon, a state of affairs, a behavior, etc.). Asking for and giving explanations are linguistic acts that have to do, in the case of the interrogative act, with the feeling or recognition of a certain degree of ignorance (and the desire to dissipate it), and in the case of the explaining act, with the proposal of an answer associated to a certain knowledge (or presumption of knowledge)[3]. The linguistic unities (words, concepts, sentences, etc.) included in an explaining act may take different forms. Nonetheless, most of the time explanations take the form of a set of propositions. That is, most of the explanations consist in *explanatory propositions*.

In every explaining act we can distinguish at least three basic elements: a) the explanation itself (the propositional content of the explanatory act), b) that which is explained (the explanandum), and c) that which explains the explanandum (the explanans). For example, someone asks you "Why are you getting to the meeting so late?" And you reply "Because there was a traffic jam on the way." The explanandum is the fact that you are late to the meeting. The explanation is the propositional content of the utterance "Because there was a traffic jam on the way." And the explanans is the fact in the world that there was a traffic jam on the way. There are different kinds of explaining acts (as we shall see), but most of them share this basic structure.

I would like to remark, before going further, some trivial—and, perhaps thus, usually overlooked—points about explanations. Explanations are things that we human beings say to other human beings or to ourselves. All the explanations are formulated, thought, said or written by someone at a determinate moment. That is, there exist only *human* explanations.

[3] According to the standard pragmatic taxonomy, an interrogative act is a form of 'directive' or 'exercitive' act, and an explanatory act is a form of 'representative' or 'expositive' act (Searle, 1975; Austin, 1962). Here I use the notion of 'linguistic act', and not the more traditional pragmatic notion of 'speech act', basically to include both overt (spoken, written) and covert (purely thought) explaining acts.

Explanations take place in an experiential context that has to do with human curiosity and understanding. We ask for explanations because we feel curious about something, because we want to understand something. Explanations, when successful, are discursive acts through which we obtain a particular kind of experience; the experience of understanding[4]. But explanations, as discursive acts, involve in turn a particular kind of experience, i.e., the experience of explaining. Explanations, so to speak, are born, live and die without ever leaving the realm of human experience. Nonetheless, explanations operate in an experiential metadomain with respect to what that they aim to explain, i.e., with respect to their explanandum. Explanations do not constitute, nor can they replace, what they aim to explain (Maturana, 1988). This distinction is trivial but important.

An explanation of a phenomenon X is a linguistic construction that we use to understand, know or make sense of X, but it is not X. Even more, X may have such and such ontological structure, such and such metaphysical constitution, but its explanation, to be successful, does not necessarily have to mirror such structure or constitution. The conceptual reconstruction of X, with strict fidelity to its metaphysical determinations, is just *one* of the many ways in which we can explain X. For example, a child asks "Why is it raining?" and his grandpa replies "Because trees need water to survive." Watching the rain falling over the trees, the child comments in agreement "Trees are happy with the rain", and goes back to play. The explanation, from a pragmatic point of view, is successful to the extent that the child feels satisfied with it (at least until the next day). Yet although it is true that trees need water to survive, that condition is not the physical condition that produces the rain. The atmospheric event that we call rain has certain physical determinations that have to do with the mass, volume, density, altitude and electric charge of the clouds, the ambient temperature, the atmospheric pressure, etc. A strict reconstruction of the physical mechanisms that produce the rain will speak of these kinds of variables, and the fact that

[4] This is not to say that explanations are the *only* way of obtaining understanding. See for example Lipton (2009) and Khalifa (2013).

trees need water to survive will not appear in any place. That is the plane of the metaphysical constitution of the phenomenon of rain. But in the pragmatic plane of the explanatory act, in the experiential domain of understanding, saying that trees need water to survive works perfectly well, at least for the child's requirements.

In general, the pragmatic success of our explanations can be relatively independent of the degree in which they reflect the ontological determinations of the explanandum. In other words, explanatory goodness or success does not always mean explanatory correctness (Khalifa, 2013).

The value of these trivialities will become, I hope, clearer as long as we begin to address the central topic of this thesis. But the basic point to bear in mind is that explanations that seem effective from a pragmatic point of view, may well be incorrect from the point of view of their ontological adequacy. Many explanations, intuitive or even able to evoke a certain experience of understanding in the observer, prove to be linguistic constructions that have little or nothing to do with the metaphysical determinations, the structure and functioning of the system to be explained (see, for example, in Chapter 2, Ashby's analysis of the concept of 'memory' as an explanatory construct).

We can ask for explanations about almost everything and in a practically endless form (children know this very well and enjoy exasperating their parents with interminable chains of "Why?"). Depending on the observer or community of observers that receives and evaluates them, explanations may be legitimate or illegitimate, satisfactory or unsatisfactory, strong or weak, better or worse, etc. An explanation is legitimate when it meets the criterion of validation from the observer or community of observers that has asked for the explanation. Criteria of validation are sets of assumptions, principles, rules or conventions that define the type of linguistic act (its form and content) that can be accepted as an explanation. Illegitimate explanations are explanations that, for one or another reason, do not meet such criteria. Satisfactory explanations are those that satisfy, at least for a moment, the curiosity of the observer or community of observers that has asked for explanations (unsatisfactory explanations are those that fail to do so).

To an observer, a legitimate explanation may be more or less satisfactory or unsatisfactory, whereas an illegitimate explanation is almost always

unsatisfactory. I say ‘almost’, because in certain circumstances the observer can negotiate her criteria of validation for the sake of other psychological benefits. For example, a naturalistic layperson, in normal conditions, does not accept explanations that resort to spiritual elements, Tarot cards or things like that. They are at odds with her basic ontology and are viewed as illegitimate explanations. For that very reason, these explanations are not satisfactory; they do not dissipate her curiosity. Nonetheless, under exceptional conditions, typically highly stressful situations (a natural catastrophe, the diagnosis of a terminal disease, etc.), this very naturalistic layperson may start to accept spiritual or occultist explanations. Not necessarily because she has changed her basic ontology, say, because she now thinks that there exist real spiritual forces in the world, but rather because these kinds of explanations may offer an opportune and much needed emotional compensation, a sort of “deep sense of life and death” that helps her to cope with bewilderment, fear or uncertainty. Our questions, and the curiosity that grounds them, have components both intellectual and emotional, and our attitude towards different kinds of explanations changes or moves according to them (Gopnik, 1998).

Two or more explanations that are legitimate and equally satisfactory to a community of observers may be estimated according to additional criteria, such as simplicity, internal coherence, congruence with some already established knowledge, or others. In that way, if the observers want, they may come to decide which explanation is the “best one”.

Since criteria of validation are conventions, they may take different forms. Some criteria may focus on the propositional aspects of the explanation (e.g., its semantic content), whilst others may focus on extra-propositional or pragmatic aspects (e.g., authority of the speaker, rhetoric or stylistic aspects, etc.). Some criteria require the rational demonstration of the explanatory proposition, whilst others require, in addition, some kind of empirical support.

One of the basic criteria, not the only one, with which the observer evaluates the legitimacy of an explanation is the degree of congruence or compatibility between the metaphysical frame that she subscribes to and the metaphysical frame presupposed by the explanation. By ‘metaphysical

frame' I understand, roughly, a certain ontology governed by a set of metaphysical principles. An observer or community of observers X considers that an explanation Y is legitimate if Y refers to entities and phenomena that form part of—or that at least are compatible with—the ontology subscribed to by X, and if Y does not violate any metaphysical principle considered as valid for said ontology. For example, to a pagan culture, an explanation that resorts to magic phenomena, spells, demons and malign forces is a perfectly legitimate explanation, basically because it resorts to entities (demons, malign forces) and phenomena (magic, spells) that form part of the pagan metaphysics. To some religious communities (e.g., the catholic community), an explanation that resorts to magic phenomena and spells is not a legitimate explanation, but one that resorts to miracles is (miracles have a divine origin, while magic and spells are considered pagan heresies). To an animistic community, the most sensible explanation for the occurrence of a tsunami is that the sea has got angry for some reason (typically the misbehavior of a member of the community) and has decided to punish them in an exemplary way.

Each explanatory system—magical, religious, animist—is valid within a determinate metaphysical frame, and each metaphysical frame corresponds to a set of beliefs shared by a concrete community. These beliefs, most of the time, operate in implicit form; i.e., they are rarely defined in an explicit and systematic way. Through its metaphysical frame each community defines, with more or less precision, that which counts as real and that which counts as unreal or fictitious, that which seems ontologically possible and that which seems ontologically impossible. Every metaphysical frame is, ultimately, a particular human construction.

Naturalistic explanations are no different in this sense. They are valid only within a determinate metaphysical frame, namely the naturalistic metaphysical frame. This point may seem quite obvious; naturalistic explanations are, of course, those that follow a naturalistic metaphysical frame. What may not be so obvious is the meaning of the expression 'naturalistic metaphysical frame'. What are the main characteristics of a naturalistic metaphysics? What is naturalism? When we hold that something is natural, what do we mean?

Naturalism

There are several kinds of naturalism, and various senses in which one can understand terms such as ‘natural’, ‘nature’ or ‘naturalism’ (Flanagan, 2006; Rosenberg, 1996). At the most basic metaphysical level, when we say that something is natural we simply mean that it is not supernatural; where supernatural includes, in a non-exhaustive list, entities or phenomena such as deities, demons, ghosts, goblins, witches, magic, spells, miracles, etc. (the metaphysical reasons for considering these entities and phenomena supernatural will be examined in Chapter 2). This is a basic distinction that separates the realm of the natural from the realm of the supernatural (Stroud, 1996).

Now, within the realm of the natural (i.e., non supernatural), we use the word natural to make a further distinction. We say that something is natural to mean that it is not artificial, i.e., that it has not been created or altered by human beings. For example, we distinguish between natural and artificial diamonds, natural and artificial lakes, etc. This is also the sense in which we distinguish between natural and social phenomena, and, correlatively, between natural and human (or social) sciences. We say that the human social world, and everything that human beings produce and create there (institutions, cities, laws, etc.), is not natural, not because it is supernatural but because it is cultural. That is, because it has been created according to the will of human beings or by means of social conventions.

We have therefore two senses of ‘natural’ here: natural as opposed to supernatural, and natural as opposed to unnatural (artificial, cultural). To avoid confusions, I propose to use the word Natural, capitalized, to name the realm of everything that is not supernatural, and the word natural, in lower case, to name that portion of Nature that is not artificial, manmade or cultural. So, while planets, volcanoes, political constitutions and philosophical theories are all Natural entities, only planets and volcanoes are natural in the aforementioned narrower sense. It follows from this characterization, for example, that the so called formal sciences (e.g., mathematics, logic, cryptography, etc.) are Natural sciences too, in spite of not being empirical sciences. They deal with abstract entities such as

numbers, propositions and codes; all of them human creations and therefore non supernatural.

Under the same criterion, it also follows that a good portion of philosophy works in fact as a Naturalistic discipline, at least in its contemporary version. The overwhelming majority of the contemporary philosophical research, in its distinct branches, develops through argumentations and critical reflections that do not assume any serious commitment to any supernatural element (or at least this author has rarely found them in his readings). Philosophers may use Gedankenexperimente and play with many kinds of possible worlds, sometimes positing entities or conditions with supernatural characteristics, but they rarely take that exercise at face value. What they do is to use the conceivability of those thought elements as an argumentative tool to examine either the validity or the soundness of some reasoning, not to defend the existence of supernatural entities.

Notice that this Naturalistic interpretation of philosophy is not necessarily equivalent to the metaphilosophical stance that conceives of philosophy as a discipline that is (or that must be) continuous with natural sciences. I am not saying that philosophy is or must be an extension of natural sciences. My characterization, as I see it, is more liberal or modest, and leaves enough room for establishing diverse forms of relationship between philosophy and science.

Now, the distinction between the natural world and human (cultural) world must be understood properly. Not everything that originates with humans necessarily counts as manmade. For example, the gaseous composition of the atmosphere is constantly modified (in a minimal proportion, but modified nonetheless) by our respiratory metabolism as species, but it is also modified by the way in which we use a series of chemical products, many of them synthesized by us. In the first case we alter the atmosphere as a result of our existence as living beings, just like microalgae and trees do. And we do it in the way in which our physiology, not our will, determines (taking in oxygen and delivering CO_2). In the second case, instead, we alter the atmosphere as a result of our intentional behavior, in a way that is determined by our decisions as consumers. In the

first case our intervention is as biological organisms, i.e., we cause a natural alteration of the atmosphere. In the second case we intervene as human beings; i.e., we cause an artificial alteration of the atmosphere.

Within the realm of the Natural, the distinction between the genuinely human and the natural is not entirely neat, as it depends, to a large extent, on the kind of philosophical anthropology that one subscribes to (i.e., the features or properties that one assumes as distinctively human). Still, we can make some general distinctions, or better said, prevent at least a couple of basic and recurrent confusions. There are two main ways of mistaking the relationship between the human world and the natural world. The first one is to attribute natural character to entities or phenomena whose character is artificial (social, conventional). The second one is to attribute human character to entities or phenomena whose character is natural. The first mistake consists in a sort of over-naturalization, and the second one is typically known as anthropomorphism.

Examples of over-naturalization can be found in what Marx understood as ideological conceptions of the social reality. Social orders which, being social, are essentially historical and in principle modifiable, are presented as natural orders that, because they are natural, cannot or must not be modified. For example, classical liberalists of the XVIII and XIX centuries, in their defense of the then flourishing capitalist economic system, usually engaged in this kind of over-naturalization. They presented the capitalist system as a 'natural' order (a sort of providential harmony) in which any human attempt to intervene could only cause imbalance and malfunctioning. These views, said Marx, aimed to legitimize (consciously or unconsciously) hierarchies of domination and relations of exploitation sanctioning them as 'natural', namely independent of human will and unchangeable. In general terms, there is over-naturalization every time we assign natural character to entities or conditions whose origin is human (in the relevant sense). The idea of over-naturalization will be taken up again later on in this thesis. It will be argued that a good part of the so called philosophical projects of 'naturalization'—conspicuously, the naturalization of intentionality and representational content, norms and purposes,—are ultimately projects of over-naturalization.

The opposite mistake is anthropomorphic projection, the humanization of

natural entities or phenomena. When one of us attributes human features to entities such as planets, rivers or volcanoes, the anthropomorphic projection is easy to detect (as in the example of the animist community mentioned before). Yet when one of us attributes human features to non-human entities that have certain closeness to us, the task is more difficult. For example, all non-human entities that belong to our proximate genus, i.e., animals or living beings in general, are susceptible of unnoticed anthropomorphic projections. Ethologists, and specially primatologists, are well aware of this anthropomorphic tendency and maintain a permanent methodological vigilance toward it (Wynne, 2007, 2004; Tyler, 2003; Mitchell and Hamm, 1996; Kennedy, 1992).

Similarly, most of the non-human entities that are products of human design, i.e., artifacts, tools and artificial systems in general, are susceptible to unnoticed anthropomorphic projections too (von Foerster, 1970). Engineering systems are usually described and interpreted with notions such as 'governor', 'controller', 'command', 'instructions', and similar. These notions, strictly speaking, are social-communicational metaphors that belong to the world of human relations. In strictly engineering contexts, I would say, these metaphors are innocent. The problem, for us, arises when certain biological systems are studied from the viewpoint of design, by analogy to engineering systems. Then, we start to hear that such and such biological structure 'controls' or 'monitors' such and such function, that such and such subsystem 'commands' and 'regulates' such and such task, 'giving instructions' to such and such subsystems, and so on and so forth.

Especially interesting is the case of entities that, whilst not being human, form part of the biological structure of humans. The most relevant case for our discussion here is that of the brain and the nervous system. These systems—that are not human but that have a relevant participation in the generation of all behaviors that we consider distinctively human—have been one of the favorite targets for all kinds of (sometimes very sophisticated) anthropomorphic projections. And philosophers of mind and cognitive scientists, truth be told, have not been particularly skillful at detecting them. Far from that, I would say that many of them have been, wittingly or unwittingly, their active promoters.

For example, every time the brain and the nervous system are analyzed by analogy to engineering systems, they typically appear as control systems, commanding and monitoring tasks, receiving and sending instructions. Other times they are examined by analogy to human organizations, with plans, goals, tasks to fulfill, control hierarchies, communication channels and flows of messages, codes, etc. Or sometimes, directly, they are endowed with the very intellectual human abilities whose neural basis one is trying to understand. The brain appears 'inferring' such and such things about the world, 'predicting' such and such events, 'formulating and contrasting hypotheses' about such and such state of affairs, etc.

One of the main strategies of this thesis is to insist in the idea that the brain and the nervous system are entities that, though they exist *in* us, do not exist *like* us, and that a proper naturalistic characterization of their functioning must exclude every kind of anthropomorphic projection.

Some anthropomorphic projections are very difficult to detect, not because of the closeness of the target, but because of the high degree of abstraction of the projected human issue. If I say that my laptop is not working because it is sad and tired, you will detect the anthropomorphic projection easily. If, instead, a physicist says that quantum subatomic systems are probabilistic systems, a cardiologist that your heart is malfunctioning, or a botanist that the number of rings in a trunk carries information about the age of the tree, you will not think, at least at first instance, that they are committing any anthropomorphic projection. Nonetheless, as it will be argued later on, epistemic states, normative values, modal spaces, among other categories, seem to be nothing more than mental constructions that we use as observes to make sense of what we observe. They are ways of framing our observation, not properties of the observed systems (not at least when the observed systems are natural systems). In cases like these, where the anthropomorphic projection has to do specifically with our condition as observers, I shall speak of 'observational projections'. Due to their highly abstract character, their 'transparency', so to speak, these forms of anthropomorphism are the most dangerous. They take subtle forms and are presented many times under apparently innocent labels such as 'realism', 'objectivism' or 'emergentism'. We will review, at different

moments along this thesis, several of these observational projections. Their detection and identification are the most important for our philosophical enterprise.

Over-naturalization and anthropomorphism are unsound forms of Naturalism; they confuse the realm of the natural with the realm of the cultural-artificial, or, at a more abstract level, the constitutive conditions of the observation with the constitutive conditions of the observed systems. A sound or strict Naturalism, hereafter Strict Naturalism, is that which is free of over-naturalizations, anthropomorphisms and observational projections. Strict Naturalism is the only kind of Naturalism that we will consider acceptable throughout this thesis.

Naturalization

We have distinguished between Naturalism and naturalism. Correlatively, we can also distinguish between Naturalization and naturalization. Let us start with the notion of Naturalization. What does it mean to Naturalize something? In a minimal sense, to Naturalize something, say a phenomenon X, is simply to conceive of X in terms that are compatible with a Naturalistic metaphysical frame (i.e., with a non-supernatural metaphysics). For example, to Naturalize the origin of the universe is, in the first place, to assume that whatever happened there happened in strict accordance with a Naturalistic metaphysics. Two people, a believer and an atheist, may ignore almost everything about the origin of the universe, but they still can hold different conceptions of it. The believer says that whatever happened in the origin of the universe, it happened by the action of God. The atheist, instead, says that whatever happened there it happened as a Natural phenomenon, without the intervention of any divine action. The atheist does not have at hand any explanation about the origin of the universe, her ignorance is almost absolute, but she does have a Naturalistic conception of it. To that extent, and in the most basic sense, she has Naturalized the origin of the universe.

This is a minimal sense of Naturalization, but one that is fundamental, or so I want to argue. A more ambitious version would require not only a

conception but also an explanation in Naturalist terms. The believer might reply, "You can say that you have Naturalized the origin of the universe only once you have provided a Naturalistic explanation about it, not a mere conception. No Naturalistic explanation, no Naturalization." I agree that, for a defense of Naturalism, it is very good to have a Naturalistic explanation at hand, but I think that is not the decisive point. Naturalism is essentially a matter of metaphysical conception, of ontology, and this aspect should have priority over epistemological considerations. We will analyze this idea in detail soon (see the example of the exorcism below). Let us now review the notion of naturalization.

What does it mean to naturalize something? In a minimal sense, to naturalize something, say a phenomenon X, is to conceive of X in terms that respect a Strict Naturalistic frame. That is, if X is a natural entity or phenomenon, to conceive of X as being natural, and if X is a human-cultural entity or phenomenon, to conceive of X as being human-cultural. In other words, to naturalize something is to liberate it from either anthropomorphic projections (including observational projections) or over-naturalizations, depending on the particular case. Aristotelian (or ancient) physics was essentially a Naturalistic system (no miracles, no magic), yet within that category it was an anthropomorphic system (an unsound naturalism). It attributed psychological states such as purposes and interests to physical phenomena (e.g., "bodies *seek* their natural resting state"). Modern (or Galilean) physics did not need to Naturalize Aristotelian physics but only naturalize it; i.e., to remove any trace of psychologism and teleology in the physical phenomena.

As in the case of Naturalization, I think that naturalization is primarily a matter of metaphysical conception, and only secondarily a matter of explanation. Why this priority of metaphysical considerations over epistemological considerations? Why should a determinate metaphysical conception have priority over a determinate explanatory construction? Let me illustrate the idea through a somewhat dramatic example. In many communities and for a long time, epileptic seizures were viewed as manifestations or proofs of demoniac possession. Their "treatment" consisted usually in horrible tortures (exorcisms) or even sacrifices.

Suppose we are back in the Middle Ages. A common citizen and a priest are having a long discussion about how to treat a child that suffers from epileptic fits. The citizen claims that the seizures have nothing to do with alleged demoniac forces—which he considers fictions—but with certain biological disorders, and therefore that exorcisms are useless and unnecessary tortures. The priest claims the contrary—for him demoniac forces are not fictions but real entities—and suggests that, instead of wasting time discussing the issue, they should start the exorcism as soon as possible. To close the debate about the causes of the seizures (and their consequent treatment), the audience suggests that the final decision must be made on the basis of the "best explanation". If the best explanation is purely biological, then there is no need to assume the existence of demoniac forces, and exorcism is discarded as a treatment. But if the best explanation is in terms of demoniac forces, then they are entitled to believe that demoniac forces really exist, and hence that exorcism is justified as a treatment.

The medieval citizen, ignorant about the physiological anatomy of the brain and its complex balance of neurotransmitters, manages to elaborate a rather poor, vague and unconvincing explanation for the seizures. The priest, with triumphal air, provides a full and detailed explanation of the way in which evil takes over the soul of weak persons, and of how exorcisms, when properly executed, manage to drive away the demoniac forces.

In a scenario like this, surely most of us would like to favor the citizen's opinion, even in absence of a good explanation. We would prefer a Naturalistic conception of the seizures rather than a detailed supernatural explanation of their causes. In a case like this, I guess, we would not be happy with a rule of the type "inference to the best explanation" (informally speaking). We would like to make the audience understand, for the sake of the child, that these two explanations cannot be compared and measured under the same terms; that they belong to entirely different metaphysical frames, and that what should be assessed in the first place are precisely these metaphysical frames, not the explanations as isolated elements.

The moral of this example should be relatively clear, but I need to make it explicit:

1. Before starting to compare and evaluate two or more competing

explanations, it is a good practice to check the metaphysical frames upon which they are grounded. If they share the same metaphysical frame, we can proceed to compare them in a direct way on the base of a set of accorded criteria. If, instead, the explanations are grounded on different metaphysical frames, then we have to step back and start to evaluate the corresponding metaphysical frames.

2. Sometimes it is better to have at hand a robust (well established, safe) metaphysical frame accompanied by a bad (incomplete, poor, unclear) explanation than a very good (complete, detailed, clear) explanation based on a weak (uncertain, unstable) metaphysical frame. A modest hut built on firm ground may be, in the long term, better than a luxurious palace built on swampy ground.

What is the point of all these considerations? The point is this: What I want to hold in this thesis is an unrestricted commitment to a Strict Naturalistic *conception* of the world and its phenomena, irrespective of whether or not we are able to enrich said conception with a Strict Naturalistic *explanation*. If in this thesis, and despite our best efforts, we cannot provide a Strict Naturalistic explanation of those phenomena that interest us, we will conclude that, by this time, we have failed to provide such an explanation. Yet we will never negotiate, as a way of obtaining some sort of substitute explanation, our Strict Naturalistic conception of them. That is, we will never recur to anthropomorphisms, observational projections or over-naturalizations as explanatory strategies.

1.2 What we want to explain and understand in this thesis, and the way we want to do it

As it was mentioned in the Introduction, in this thesis we want to explain and understand basically two things. First, we want to explain and understand living beings' behavior. How—i.e. through what kinds of processes and mechanisms—do living beings generate their behavior? What is it about the

nature of living systems that causes them to behave the way they do? Second, we want to understand how mental phenomena appear in the life of some living beings; the natural conditions under which certain living beings exhibit mental phenomena. Those are our questions.

To answer these questions we will adopt, as it has been explained, a Strict Naturalistic stance. According to this stance, living beings are natural systems and must be studied as such; that is, by appealing to the same ontological assumptions and explanatory principles that we use to study any natural system in general. Which are these ontological assumptions and explanatory principles? We have said that Strict Naturalism is a specific form of Naturalism, namely that which is free of over-naturalizations, anthropomorphisms and observational projections. But, in concrete, how do we translate this requirement to the case of living beings? How strict should our Strict Naturalism be with respect to living beings? Let us review this point through some examples.

When we study a river, we do not think of it as moving according to goals, purposes or ends. We do not say—when we are engaged in serious scientific discourse—that it *tries* to reach the sea, or that, in spite of the obstacles (rocks, cliffs), it *succeeds* in reaching the sea. We do not say these kinds of things because the river, strictly speaking, is not trying to do anything; it just obeys natural laws. We assume that natural systems are non-teleological (purposeless) systems.

In the same way, and for the same reason, we do not think of natural systems as acting in terms of means and ends. The Sun generates light through thermonuclear fusion, but the Sun does not use thermonuclear fusion as a means to generate light. Natural entities, strictly speaking, do not use things; they are not users. A user is an entity for which certain things appear as means for certain ends. But the Sun, in producing thermonuclear fusion, does not have in view any end in particular; it just happens that the emission of light is a consequence of thermonuclear processes. Natural systems do not interact with the world in terms of means and ends; they only interact in causal terms. Likewise, natural entities neither face nor solve problems. Nature, taken in itself, has no problems. A state of affairs may turn into a problem only for an evaluative and purposeful entity, an entity that appraises

or judges a certain state of affairs as positive or negative with respect to some ends. Again, the Sun generates light through thermonuclear fusion, but the Sun, by producing thermonuclear fusion, is not 'solving the problem' of light generation.

When we observe a solar eclipse, we do not say that the Moon, blocking the Sun and shadowing the Earth, has 'made a mistake'. The Moon's movements are never right or wrong, good or bad, successful or unsuccessful in themselves; they just are. The Moon does not have any duty to fulfill; its behavior is not a matter of *must* or *ought*. We conceive of natural systems as non-normative systems.

When we study a lightning flash, we do not think of it as strictly speaking evaluating possibilities of action, as choosing, out of a set of alternatives, its own path to the ground. Nor do we say that the lightning controls or regulates, as a pilot, its trajectory to the ground. We do not think of natural systems as agents that choose what to do and control their behavior. We assume that natural systems behave according to laws and physical conditions that are given.

When we study the correlative movements of the Moon and tides, we observe a strict covariation. Yet we do not think of them as coordinating their displacements by means of messages, signs, codes, instructions, or some form of semantic interaction in which they act as interpreters. We do not think of tides as responding to the Moon's movements on the basis of informational contents, representations or intentional states. We assume that natural systems relate to each other in strictly physical-causal terms.

Ultimately, and perhaps for all the aforementioned reasons, i.e., because they are not the kinds of systems that can be evaluated in terms of success or failure, because they do not try to fulfill any task or goal, because their behavior is never a matter of choice, we do not think of natural systems as being more or less intelligent, or more or less stupid. Nor do we explain their behavior in epistemic terms. If a group of comets pass very close to Jupiter, and all except one of them collide and disintegrate, we do not say that this comet has a better knowledge of gravitational fields. Natural systems do not have epistemic states; they are neither knowledgeable nor ignorant about

anything. Natural systems' behavior is neither intelligent nor stupid; it is simply the result of laws and physical conditions that are given.

In this thesis, *Strict Naturalism* is the claim that living beings are natural systems, and that they must be studied by appealing to the same ontological assumptions and explanatory principles that we use to study any natural system in general. In so doing, notice, we are not denying that living beings have unique and distinctive features as natural systems (we will review some of them); we only assert that those features do not set living beings apart from the rest of natural systems. Following this requirement, in this thesis we will characterize living beings' behavior in terms of what Maturana, in line with Ashby's cybernetic approach (Ashby, 1960), calls 'structural drift' (Maturana, 2003).

The notion of structural drift has two complementary connotations. First, the concept of *drift* means that the system exists and behaves in a deterministic way, without purposes, possibilities of choice, or ability to control its behavior (as opposed to a boat controlled by a helmsman who chooses a route according to some goal). Second, that the drift is *structural* means that the system's behavior is determined solely by material or physical factors (forces, energy, etc.), not by epistemic, normative, intentional or semantic factors. Assuming that living beings are systems in structural drift is to assume, therefore, that they (i) do not follow any purpose or goal, (ii) do not have possibility of choice, and (iii) do not have control over their behavior. If living beings are drifting systems, then they are strictly deterministic systems. Complementarily, if the drift of living beings is structural (not epistemic, normative, semantic or intentional), then (i) living beings do not host epistemic states or processes, (ii) their behavior is never intelligent or stupid, good or bad, successful or unsuccessful, and (iii) their relationship with the environment is strictly structural (not epistemic, normative, semantic or intentional).

Thus, to say it with Ashby's words, in this thesis "[i]t will be assumed throughout that [...] an animal [or a living system in general] behaved in a certain way at a certain moment because its *physical and chemical nature at that moment allowed it no other action*" (Ashby, 1960, p. 9. Emphasis added).

This is a way of thinking about living beings that we will assume as our starting point, but also, a way of thinking that the thesis, through its different chapters, will try to vindicate. That is, unless we find legitimate, compelling reasons to think otherwise, we will understand living beings according to the Strict Naturalistic framework above described, i.e., as systems in structural drift. (The thesis will try to show that there are not such compelling reasons).

1.2.1 The cognitive construction of living beings

When viewed from a Strict Naturalistic framework, living beings seem to lose the epistemic and intentional flavor to which we, cognitive scientists and philosophers, are used to. When we observe a living being in its environment, we typically tend to assume, implicitly or explicitly, at least four things.

First, we tend to see the living being as trying to solve some problem or fulfill a task (e.g., how to get food, how to escape from predators, how to reproduce, etc.). Generally, we assume that living beings try to stay alive as much as they can, or that they want to ensure the continuity of the species. When we do that, we frame living beings within a normative and teleological context, which allows us to evaluate whether or not their behavior is successful or adequate. On this basis, generally, we decide whether or not the organism's behavior is intelligent. By extension, and also by decomposition, we tend to do the same with the organism's subsystems. For example, in animals, we tend to see their nervous system, and specially the brain, as solving sub-problems or fulfilling sub-tasks (e.g., detecting features of the environment, commanding responses), and through them, as contributing to meeting the general goals of the organism.

Second, we tend to see living beings as agents that, before a problem or determinate situation, have the ability to choose what to do, or that have some degree of control or regulation over their behavior. In animals, for example, the brain is usually viewed as a control system that commands,

regulates and monitors the organism's behavior. Basically, living beings appear to us as systems that, somehow, escape brute and blind determinism, and to that extent, as systems that manifest some degree of agency or freedom of action.

Third, we tend to assume that living beings are systems open and oriented to the environment, with sensory 'windows' that allow them to pick up stimuli from the external world (sensation and perception), and effector organs that allow them to act upon said external world (action). Intuitively, we see living beings as having an external world before them; a world toward which they (need to) direct their sensory windows and their action, and with respect to which they need to adapt.

Fourth, as long as living beings are viewed as agents that try to solve problems, their interaction with the environment (and by extension their internal states), are usually interpreted in epistemic or cognitive terms. Living beings not only exist in the world, they 'know', or 'need to know', the world in which they exist. The epistemic interpretation of living beings, or their subsystems (typically the nervous system), may come in different versions. One version, perhaps the more traditional, assumes that living beings know the world in which they live in a more or less indirect way, i.e., through the mediation of internal representations or some form of internal modeling. Other, less traditional versions, see living beings as having a non-representational epistemic contact with the environment, either in terms of direct ecological information about opportunities for action (e.g., Gibson, 1966, 1979; Chemero, 2009), or in terms of sensorimotor (contentless) intentionality, sense-making, meaning creation and similar phenomenological characterizations (e.g., enactive approaches. See Weber and Varela, 2002; Thompson, 2007; Hutto and Myin, 2013). What remains as a common denominator is the view of living beings as epistemic or cognitive agents.

When combined, these four assumptions constitute what we might call the 'cognitive construction' of living beings; that is, living beings as objects of study for the particular interests of the cognitive scientist and philosopher. Our Strict Naturalism claims that the cognitive construction of living beings is precisely that, a construction, not a characterization of their ontology as

natural systems. In certain contexts, this 'construction' of living beings as cognitive agents may be a legitimate exercise. I wish, however, to distinguish sharply between an ontological characterization of a natural system and its observer-relative construction.

For example, when a chemist sees a piece of gold, what she sees is a determinate molecular composition and a series of natural properties (atomic weight, melting point, malleability, conductivity, etc.). When the economist sees a piece of gold what he sees is richness, a certain amount of value and an associated price in the market. The ontology of gold as a natural element corresponds to the characterization provided by the chemist (gold as a natural kind). The economist deals with gold as a precious metal, as an object of value within the context of human interests and institutions (gold as a conventional kind). Atomic weight and electron configuration are intrinsic properties of gold; market value and price are relative and context dependent properties. The status of gold as a precious metal is relative to the needs and interests of a certain kind of economic agent that operates in a certain historical context. In principle, depending on the economic agent and its historical context, any metal might be elevated, at some moment, to the category of precious metal.

The chemist, if asked, may explain the value attributed to gold as rooted, in part, in its chemical properties: "We assign to gold the value that we assign, in part, because of its non-corrosiveness, its non-reactiveness, its malleability, and its easy smelting." The economist, only under a severe confusion, would try to explain the chemical properties of gold as rooted in its market value.

The object of study of the chemist is, thus, ontologically primary (gold as a natural kind) with respect to that of the economist (gold as a conventional kind). This ontological hierarchy does not render the economist's knowledge superfluous or less important than that of the chemist, or his discipline, economics, as less serious and rigorous than chemistry. Yet the hierarchy is there, and it is worth acknowledging it to prevent potential misinterpretations.

For example, when the chemist says that the economic value of gold is rooted, in part, in its natural properties as a mineral, she is not *reducing* the

economic value of gold to its natural properties. She is not saying that the market value of gold simply equates to its chemical properties. What she means is that the natural properties of gold may explain, in part, why, *once put in the context of human institutions and interests*, gold has for us the economic value that it has. If the economist misinterpreted the chemist as 'naturalizing' the market value of gold, he would be, in reality, over-naturalizing his object of study; i.e., taking as natural something that is cultural and conventional. The market value of gold is something Natural, not something natural. A Strict Naturalistic economist would recognize the constructed and conventional character of his object of study, and would not ask for over-naturalizations.

This is the sense in which, as the title says, this thesis wants to speak of the "biological *roots* of cognition." When viewed from a Strict Naturalistic frame, living beings appear as non-epistemic, non-cognitive, non-intentional, non-semantic, and non-intelligent systems; that is, devoid of those aspects that cognitive science cares about. This characterization corresponds to the primary ontology of living beings as natural systems; i.e., as natural kinds. A Strictly Naturalistic cognitive science, the kind of cognitive science that I favor, would acknowledge the constructed character of living beings as cognitive systems, and see, as Ashby (1962) and Maturana (2003) will show us, that, in principle, any physical system in interaction with its surrounding might be considered, under certain criteria, as an 'intelligent' or 'cognitive' system. Complementarily, such a cognitive science would reject any attempt to over-naturalize the aspects that constitute its object of study (i.e., intelligence, teleology, normativity, intentionality, etc.).

In the same way that the chemist may explain, without reduction, the economic value assigned to gold as partially rooted in its natural properties (its primary ontology), so in this thesis we will try to explain the cognitive and intelligent status *attributed* to living beings as rooted, in part, in their primary ontology as natural systems. This explanatory construction, metaphysically austere and devoid of those 'cognitive' aspects to which we are used, perhaps will not be the best one, in comparison to more standard 'cognitive' explanations, in terms of detail, clarity or coverage. Yet, as in the case of the medieval citizen in our previous example, it will have the merit

of placing living beings in strict continuity with the all other natural systems, which is one of our goals in this thesis.

1.2.2 Ontology of Living beings

What kinds of natural systems are living beings such that they behave in ways that we qualify as intelligent, cognitive, purposeful or intentional? Which are, from a Strict Naturalistic viewpoint, the main ontological features of living beings? Here is our answer.

From a Strict Naturalistic viewpoint, living beings may be characterized as:

1. Adaptive dynamic systems
2. Deterministic machines of closed transitions
3. Multistable dissipative systems
4. Organizationally closed systems with respect to (i) their sensorimotor activity, and (ii) their molecular productive processes (i.e., autopoietic systems)

These four features can explain, we think, living beings' behavior, and also help us to understand why we observers tend to appraise said behavior as intelligent, cognitive and purposeful.

Point 1 says that living beings are adaptive dynamic systems. What does this mean? What is a 'system'? When is a system a 'dynamic' system? In which sense are living beings 'adaptive' systems? The rest of this chapter will be dedicated to examine and answer these questions. (Points 2, 3 and 4 will be examined in Chapters 2, 3 and 4 respectively).

1.3 Distinctions, unities, domains, identities

Every time we observers point out or individuate a unity (thing, object, event, phenomenon, etc.) in the concrete field of our sensorial experience or in the abstract space of our thought, we perform an act of distinction by means of which we separate that which is attended to from a background (concrete or abstract) according to some criterion of distinction that we use in an explicit or implicit manner (Maturana, 1975; Spencer-Brown, 1979). The space, ambit or domain in which a unity (and its correlative background) is distinguished may be any; natural, social, physical, conceptual, etc. Thus, planets, chairs, subatomic particles, mathematical equations, political constitutions, good deeds, wars, unicorns, are all unities of a certain kind distinguished in a certain space or domain (concrete, abstract, real, fictional) defined by one or more observers. A unity, to put it more simply, may for all practical purposes be understood as that which we can mention in language.

The domain in which the observer distinguishes a unity defines the domain of existence of the unity, and the unity, in its specificity, defines in turn a particular domain of phenomena and ontological determinations. The domain of phenomena of a unity, roughly speaking, is all that may happen with the unity; i.e., the total space of its vicissitudes. The domain of its ontological determinations is all that the unity is or may be; i.e., the total space of its possible determinations or properties. Different unities have different ontological and phenomenic domains. For example, If X is a crystal glass, it may break; if X is a volume of liquid water, it cannot break. If X is a civil law, it can be derogated; if X is a galaxy, it cannot be derogated. If X is a human action, it can be altruist or egoist; if X is a mountain, it can be high or low, but not altruist or egoist, etc. In other words, each unity admits certain predicates (descriptions) and excludes others, which has to do with its particular space of existence and its specificity as a unity.

When the observer perceives or conceives of a unity as interacting with a certain surrounding or medium, she deals with an interacting unity. When, on the contrary, the observer perceives or conceives of a unity as having no interaction with any surrounding or medium, she deals with an isolated unity

(e.g., the Universe, as a whole, is usually considered as an isolated unity).

When the observer cannot (or does not want to) distinguish further unities within the already distinguished unity, she deals with a simple unity. When, on the contrary, the observer distinguishes subunities within the already distinguished unity (because she can or prefers to do so), the observer deals with a composite unity. A simple unity appears as an entity without components, i.e., as an unanalyzable or atomic entity, and its properties as simple or fundamental properties (Maturana, 1975, 1980, 1981). For example, if a unity X, say a cube, is solid, the observer accepts the solidity of X as its constitutive property without asking whether there are some subunities in X (components) whose properties and relations might account for the solidity of X (X simply *is* solid, period). In the case of a composite unity, the observer sees the unity as constituted by a set of components arranged in a certain way, and its properties as based on the properties of its components plus the particular way in which they are arranged (Maturana, 1975, 1980, 1981). For example, the observer may explain the solidity of X (the cube) in terms of the cohesive force among the molecules that compose it.

A composite unity corresponds to what we usually call 'system'; an assembly of components interconnected in a certain way. Depending on the particular space or domain in which we make our distinctions, we may speak of physical systems, legal systems, financial systems, etc. Living beings are a particular version of physical systems.

A simple unity exists as a totality. A composite unity, or system, exists not only as a totality but also as a set of components. These two domains of existence (as a totality and as a set of components) constitute to the observer different domains of phenomena that require, in many occasions, different descriptions. This means that the observer, in her descriptions, will find that there are some concepts or predicates that apply at the level of the totality, but not at the level of its components, and vice-versa. In our previous example the observer explained the solidity of X (the cube) in terms of the cohesive force of the constituent molecules. Suppose now the observer looks for further specifications, asking whether the constituent molecules are, in turn, solid or not. We applaud her curiosity, but recognize at the same time

that she has asked the wrong kind of question. The descriptive categories ‘solid’, ‘liquid’ and ‘gaseous’ do not apply to molecules. Rather—we explain to her—, it is the particular organization that molecules adopt among them (close together or scattered) that determines the supra-molecular properties ‘solid’, ‘liquid’ and ‘gaseous’.

When two domains are orthogonal, that means, in Maturana’s terminology, that each domain specifies and has its own space of associated phenomena, and that these cannot be transferred, without committing a fallacy or category mistake, from one domain to the other. For example, we can say that a unicellular organism has died, but we cannot say the same about its lipidic components, as lipids, taken in themselves, are not living unities. An excessive accumulation of lipids may bring as a result the death of the cell, yet what dies is always the cell, not its lipidic components.

When a unity is perceived (in the case of concrete unities) or conceived of (in the case of abstract unities) as an unchanging entity, we deal with a static unity. When the unity is perceived or conceived of as undergoing one or more changes, we deal with a dynamic unity (we will provide more terminological details in Chapter 2). For example, if an observer looks at a pebble for a couple of hours (in a room with constant and homogeneous environmental conditions), and ignores (for whatever reason) the dynamic of the subatomic particles that compose it, the pebble will appear to her as a static unity. The same observer, equipped with a special device that detects the movement of the subatomic particles, will perceive the pebble as a dynamic system. My ideas about international politics change from time to time; they are changing entities in a propositional space. Platonic *Ideas*, as conceived of in Plato’s metaphysics, are immutable and eternal entities; they are static unities in an ‘intelligible world’.

The changes undergone by a dynamic unity may be internally generated (endogenous changes) or triggered by external factors (exogenous changes). In the first case we speak of ‘active’ dynamic unities (they are dynamic in virtue of their endogenous changes), and in the second case we speak of ‘passive’ dynamic unities (they undergo changes only thanks to the action of external factors). This last case, of course, is a real possibility only for

interacting dynamic unities. Isolated unities, if dynamic, are so only in virtue of their own dynamic.

Strictly speaking, as long as physical objects are conceived of as composed by atoms (which are in themselves dynamic unities), every physical object can be considered, in principle, as an active dynamic system. Conventionally, nonetheless, macromolecular physical objects are not considered active dynamic systems *merely* because of their atomic composition. Both a chair in a terrace and an erupting volcano are physical unities composed by atoms, but we do not see them as being dynamic in the same way. A chair in the terrace is usually viewed as a passive dynamic system, i.e., as a system that exhibits certain changes through time, but only because external factors (environmental humidity, solar radiation, the usage of people) act upon it. An erupting volcano, on the contrary, is viewed as an active dynamic unity, i.e., endowed with an endogenous magmatic activity. In this thesis we will follow this convention.

When a dynamic system remains always passive or always active, depending on the case, we say that the system has a fixed dynamic regime. When the system is able to alternate, without loss of organization, between passive and active dynamics, we say that the system has a variable dynamic regime. A chair, for example, is a dynamic system of fixed regime (it is always passive). A volcano, on the contrary, is a dynamic system of variable regime; when dormant (inactive), it behaves as a passive dynamic system, and when erupting, as an active dynamic system. Artificial machines with energy supply, such as engines or electronic devices (TVs, radios, laptops, etc.), are examples of dynamic systems of variable regime too. Whether the regime of a dynamic system is fixed or variable depends, to a large extent, on its thermodynamic regime. Living beings, as it will be shown in Chapter 3, are active dynamic systems of fixed regime (because of their condition of far-from-equilibrium thermodynamic systems).

When the dynamic of a unity is exclusively endogenous or exclusively exogenous, we speak of 'partial' dynamic unities. Thus, passive interacting dynamic unities (e.g., a chair on the terrace) and active isolated dynamic unities (e.g., the Universe considered as a whole) are both partial dynamic unities. When the dynamic of a unity is the result of both exogenous *and*

endogenous factors, we speak of 'total' dynamic unities. By definition, only active interacting unities behave as total dynamic unities. Stars, active volcanoes, tornadoes, but also working artificial machines, are examples of total dynamic systems; they exhibit an endogenous dynamic, and at the same time, interact with a physical environment that affects them to a greater or lesser degree. Living beings are a particular version of total dynamic systems.

Importantly, in some active dynamic systems one can distinguish not only internal changes but processes, operations or mechanisms that exhibit a certain pattern or configuration. Generically, these unities can be called 'systems of processes'. In a system of processes the relevant components are the processes performed by the system (combustion processes, oxidation processes, selection processes, etc.), not the elements through which said processes are carried out (combustible materials, chemical compounds, people, etc.). Depending on the particular way in which the processes are configured, we may distinguish between serial and parallel systems, linear (open) and circular (closed) systems, centralized and distributed systems, etc. We will come back to these systems later on when, following Maturana's autopoietic theory, living beings are characterized as closed systems of processes of chemical production (Chapter 4).

1.3.1 Class identity

Usually, when we distinguish a specific unity, we do so by recognizing it (implicitly or explicitly) as a member of a certain class; i.e., we define its identity by its belonging to a specific class. For instance, I distinguish that which is in front of me as a chair insofar as I recognize it as belonging to the class of chairs. The class of chairs is not a concrete entity out there among the chairs, but rather the mental construct or formation that I use to identify certain objects as chairs[5].

[5] This notion of class, which I will use in this section, is basically a conceptualist

Every class is formed according to some classification criterion whose nature (well defined or ambiguous, subjective or objective, constitutive or relational, institutive or non-institutive, etc.) determines the nature of the class so formed. Thus, well-defined criteria generate well-defined classes (e.g., the class of prime numbers under 100), and badly defined criteria generate classes whose extension is imprecise (e.g., the class of slightly pretentious intellectuals). Subjective criteria generate subjective classes (e.g., the class of all things that I like), and objective criteria—in the sense that their validity is not defined by the personal preferences or desires of anyone in particular—generate objective classes (e.g., the class of electrically nonconductive materials).

Constitutive criteria are those that focus on the unity itself without considering its relations with other unities, and relational criteria are those that attend to said relations. In one case we specify constitutive properties, and in the other relational properties. Constitutive criteria may focus either on the superficial features of the unities (their external appearance), or on more essential properties such as their internal composition or organization (the *organization* of the unities will be a key notion for our subsequent analysis). In both cases, superficial or 'deep', the object or unity is considered as an isolated entity. Relational criteria, on the contrary, place the unity in a matrix of relations (actual or potential) and specify its properties in virtue of said relations (e.g., the way in which the unity affects other unities, its functional role within a certain process or mechanism, etc.). Thus, for example, someone may define the class of chairs as the class of all those material objects that exhibit a particular spatial organization among their pieces (constitutive criterion), while someone else may define it simply as the class of all those material objects on which one can sit (relational criterion).

Institutive criteria are those that stipulate or institute the properties that will define the class members as such. For example, an academic community

---

notion. From a conceptualist viewpoint classes are primarily logical entities, i.e., products of a certain intellectual activity. What may be concrete and extra-logical, in many cases, is their extension (the particulars that they cover).

decides to institute an annual award for best doctoral thesis. In awarding the prize to a different student every year, the community is defining and actually creating the members of the class ‘winners of the award for best doctoral thesis’. Most of our social institutions work in this way, i.e., by creating classes whose members exist uniquely by convention (e.g., the class of married people, the class of prime ministers, the class of illegal immigrants, etc.). Non-instituted classes are classes in which the identity of their members (the distinctive property that defines them as members of a particular class) is not instituted or created by the observer. Examples of non-instituted classes are the so called ‘natural kinds’. For example, gold, as a mineral, is a natural kind (whereas as a precious metal or monetary reserve, it is a conventional class).

Recall we are understanding classes as mental constructs. What is peculiar in the case of instituted or conventional classes, as opposed to non-instituted classes, is *the way* in which one or more observers generate the classes in question, not the fact that the classes are generated by one or more observers. Let us try to clarify the difference. Someone becomes the new prime minister of UK in virtue of a procedure (a convention officially recognized) that we ourselves have created and whose application confers to a determinate person the property ‘prime minister’. A community of observers (biologists) examines a vast sample of cells and generates the classes (taxa) ‘prokaryotic cells’ and ‘eukaryotic cells’. Although these taxa, *qua* classes, are the product of the collective work of the biologists, it is clear that the biologists are not the creators of the procedure (i.e., the evolutionary process) by which, millions of years ago, some cells without a nucleus became nucleated cells. It is one thing to form a class by means of an abstraction that picks up or draws some generic property in the unities under consideration (e.g., presence or absence of nucleus in the cells), and another is to form a class by means of a stipulation that creates the very generic property (e.g., being a prime minister) that will define which unities do or do not belong to the class so constituted.

By distinguishing a unity as a member of a certain class, we identify its ‘class identity’. This operation is simple routine for us as observers, but we are not always aware of its possible complications. Peter is a human being;

i.e., a particular that belongs to the class of 'human beings'. Now, if Peter is a human being, then he is also a living being (a particular that belongs to the class of 'living beings'), and if he is a living being, then he is also a physical entity (a particular of the class 'physical entities'). But Peter is not only a human being; he is also a man, a blue eyed person, a British citizen, a university student, a son, a friend, a cellist in the local orchestra, a sympathizer of the Conservative Party, etc. Peter belongs to many different classes and therefore has many different identities. Some identities last for a couple of years (university student), some of them a lifetime (living being). How can so many different identities coexist in Peter? The answer is simple: they can coexist because they take place in different domains, and because their values respect the ontological dependencies that exist among these domains (no class identity is incompatible with any other). This equals to saying that Peter exists in many different domains, each one with its own phenomenic and ontological domain. And here is where things may get complicated. Sometimes it is not easy to distinguish between different domains, especially when they are ontologically contiguous. Let us see this through some examples.

The day that Peter gets married his identity as a single man disintegrates and a new one appears; married Peter. Although Peter's biological structure is 'involved' in said identity change (Peter's body goes wherever Peter goes), it does not constitute per se the domain in which such identity change takes place. The change of marital status takes place in a domain of social norms or civil laws, and when we say that Peter has got married we are distinguishing him as a civil subject in the domain of such social laws, not as a biological system in the domain of physical bodies (e.g., it would not make sense to say that the totality of Peter's cells have got married). All this looks (I hope) relatively clear and reasonable, but now consider a trickier case. Medical exams have confirmed that Peter suffers viral hepatitis, i.e., that Peter is sick. When we say that Peter is sick, what do we mean? It seems, prima facie, that this identity is constituted in the biological domain of Peter; something that goes on in his liver. Under closer examination, however, this does not seem so. What goes on in Peter's liver is that a certain viral population is interacting with the hepatic cells, resulting in a certain pattern

of morphological and physiological changes. Those are the facts. Now, whether this biological condition in Peter's liver counts as a disease or not (or makes Peter a sick person) is something that, strictly speaking, is defined at a different level.

'Healthy' and 'sick', 'normal' and 'abnormal' are essentially normative categories; they distinguish between the functional and the dysfunctional, between good and bad, adequate and inadequate, right and wrong, etc. We say that the viral infection alters the proper or normal functioning of the liver, yet what we call 'the proper functioning of the liver' is nothing but the way the liver *ought to* work according to our expectations as observers. And the whole point is that, loosely paraphrasing William James (1909), biological phenomena are neither normal nor abnormal, they just *are*. Biological processes or mechanisms do not *have to* work in this or that way; they do not *have to* meet such and such expectation. Biological processes are what they are and occur as they occur as a matter of fact, not as a matter of norms.

A biologist and a doctor are observing the exams of Peter's liver. We ask "Is there anything wrong with Peter's liver?" The doctor says "Well, I am sorry to tell you there is something very wrong with Peter's liver. It is infected with a very dangerous virus that is producing a severe chemical imbalance." The (Strict Naturalistic) biologist pauses and says "Well, strictly speaking there is nothing wrong with Peter's liver. It is reacting exactly as an infected liver would do. Peter's liver is behaving in perfect conformity with its actual biochemical conditions; it is not violating any natural law."

There is a fine line between the pattern of biological functioning that we find in Peter's liver and the normative category that we use to qualify it. The biologist speaks from the point of view of the liver as a natural entity; the doctor from the normative point of view of the human expectations and preferences. Both points of view are legitimate and useful in their own terms; the crux is not to confuse them.

These considerations reinforce, from a different angle, the distinction that we drew in previous sections between the primary ontology of living beings as natural systems and their construction as objects of study for the interests of the cognitive scientist or philosopher. One thing is the class identity of

living beings as natural systems, and another their class identity as cognitive systems. Both of them are valid in their own terrain; the crux is not to confuse them.

1.3.2 Systems: structure, organization, adaptation

We will focus now on systems (i.e., composite unities) and on a particular way of defining their class identity, namely by attending to their organization. In every system we can distinguish, at least in principle, a determinate organization and structure, or better said, a determinate organization instantiated in a determinate structure. The organization of a system denotes "[t]he relations between components that define a composite unity (system) as a composite unity of a particular class" (Maturana, 1981, p. 24), whereas the structure denotes "[t]he actual components (all their properties included), together with the actual relations that […] realize a system as a particular member of the class of composite unities to which it belongs by its organization" (Maturana, 1981, p. 24). For example, that material object in front of me is a chair as long as it exhibits a certain spatial organization such that I can identify it as belonging to the class of chairs, while its structure corresponds, among others things, to its concrete material realization (e.g., wood or metal), size (big or small) and design (baroque style or Bauhaus style). In an abstract domain, a logical reasoning is a deductive inference as long as its premises and conclusion exhibit the kind of logical organization that is characteristic of deductive inferences; its structure, on the other hand, includes aspects such as the particular propositional content of its premises, their respective truth values, and the validity of the logical connection between the premises and the conclusion. In the case of the chair what matters is a certain *spatial* organization, whilst in the case of the deductive inference what is critical is a certain *logical* organization.

The relation that constitutes a determinate organization may belong to any type (spatial, logical, temporal, functional, etc.), as it depends on the

descriptive domain chosen by the observer. For example, we can identify the class identity of a vehicle engine in terms of the spatial organization (layout) of its pieces, in which case we speak of 'Straight/In-line' engines, 'V type' engines, and 'Boxer/Flat' engines, among others. But we can also identify an engine in terms of its operational organization or 'cycle' (the steps through which it converts combustible into kinetic energy), in which case we distinguish, among others, 'Two-stroke' engines, 'Four-stroke' engines, and 'Six-stroke' engines. The engine can be viewed as a set of physical pieces but also as a set of combustion operations (i.e., as a system of processes). Notice, this means that the 'components' or subunities under consideration are different in each case: material pieces in one case, combustion operations in the other.

Maturana, following a constitutive criterion, contends that what defines the class identity of a system is its organization, not its structure. A change in the structure of a unity *may or may not* lead to a change in its class identity, whilst a change in its organization leads, by definition, to a change in its class identity. For example, a metallic chair remains a chair as long as it maintains a particular organization among its pieces. While this organization is conserved, the chair can admit several structural changes without losing its class identity as a chair (e.g., we can change the length of its legs, or replace them by wooden legs, etc.). Now, if we set the chair on fire, we will observe a sequence of structural changes (i.e., the melting of the metal) that lead finally to the disorganization of the chair as a chair. Similarly, if we disassemble the chair and reorganize its pieces to make a table, the chair disappears as a system and a new system appears in its place. The chair exists as a chair as long as it maintains a certain organization that is recognizable for us, and ceases to exist when said organization changes or is lost. Its structure, on the contrary, can change within certain ranges without altering its class identity as a chair. In a more abstract domain, we can change the structure of a deductive inference within certain ranges without altering its identity as a deductive inference. A valid and sound deductive inference about the mortality of Socrates remains a deductive inference even if we change its content for the mortality of my pet, the truth value of one of its premises, or replace its conclusion by one that it does not

follow from the premises. The result will be simply an invalid and unsound deductive inference about the mortality of my pet. As long as the inference conserves the kind of logical organization that is characteristic of the deductive inferences, none of these structural changes will alter its class identity. Now, if we keep on changing its structure in different ways, there will be a moment at which the deductive inference will disappear as such and become, perhaps, an abductive one, an inductive one, or simply a set of sentences that do not constitute any reasoning at all.

Of course, what changes and what remains the same in each case depends on the space or domain in which the observer is making her distinctions. When we take a baroque chair and transform it into a Bauhaus chair, the specific unity that we identify as 'baroque chair' in the domain of 'styles of chairs' loses its class identity and disappears as such, while the more generic unity that we identify simply as 'chair' (whatever its style), remains the same. When we take the chair and transform it into a table, the unity chair loses its identity and disappears as such. Nonetheless, if our unity of observation corresponds to the broader category of 'furniture', it is clear that any change from chair to table or vice-versa will appear as irrelevant. In the physical world—let us leave aside for now the abstract domain of the purely thought unities—the disintegration of a unity always presupposes the conservation of a more basic unity with respect to which the disintegrated unity was a particular version. In the limit case, if our unities of observation are as generic and basic as 'physical phenomenon', 'physical magnitude' or 'physical unity' (including massless particles, energy fields, etc.), then we will observe that everything is conservation through the constant constitution and disintegration of physical entities (we will come back to this point in the next chapter when talking about 'systems of closed transitions').

When a system admits of one or more structural changes without loss of organization (i.e., without losing its class identity), we speak of a structurally plastic system. In our previous examples, we saw that both the chair and the deductive inference admitted a number of structural changes without losing their class identities. Chairs and deductive inferences are therefore instances of structurally plastic systems. When the system under consideration is a concrete physical unity in a concrete medium, say, a chair in a room,

Maturana calls this process—conservation of organization through structural changes in interaction with a medium—'adaptation' (Maturana, 1981). Understood in the context of Maturana's metaphysics, 'adaptation' is not an exclusively biological category.

> If the organization of a composite unity remains invariant while it undergoes structural changes [...] through its recurrent interactions in its medium [, we say that] its adaptation is conserved [...] Defined in this manner, [...] conservation of adaptation is not peculiar to living systems. It is a phenomenon that takes place whenever a plastic composite unity undergoes recurrent interactions with structural change but without loss of organization. (Maturana, 1980, pp. xx-xxi)

An adaptive system, from this point of view, is simply an interacting dynamic system endowed with structural plasticity. Passive dynamic systems, such as chairs, shoes or dormant volcanoes, exhibit passive adaptation, whereas active dynamic systems, such as stars or erupting volcanoes, exhibit active adaptation. Living beings' adaptation, as we shall see in chapters 3 and 4, is a particular case of active adaptation.

Notice that, defined in this way, adaptation is a purely descriptive concept. It is a way of naming the conservation of integrity that takes place in interacting dynamic systems, not an explanation of how this conservation takes place. In Maturana's conceptual system, adaptation is a condition that needs to be explained; it is an explanadum, not an explanans. The adaptation of each system, passive or active, living or not, must be explained attending to its particular organizational and structural features, and always respecting, as we shall see in the next chapter, certain fundamental metaphysical principles.

There are many aspects that one can consider in the analysis of an adaptive dynamic system, passive or active. For our purposes, nonetheless, there is one in particular that is crucial: the *way* in which the system undergoes its changes of state. The way in which a system undergoes its

changes of state depends on, or is ruled by, certain metaphysical principles that have to do, basically, with the ‘actuality’, ‘conservation’, ‘continuity’, and ‘determinacy’ of the system (among others). One discipline that has studied in an extensive way the metaphysics of the dynamic systems is cybernetics. We will dedicate the next chapter precisely to this topic; the metaphysical analysis of dynamic systems using the conceptual tools of cybernetic theory.

Chapter 2

Dynamic systems and deterministic machines: lessons from cybernetics

In the previous chapter we declared our unrestricted commitment to a Strict Naturalistic conception of the world and its phenomena, emphasizing the idea that said Naturalism, if sound, has to be free of anthropomorphisms, observational projections and over-naturalizations. After that we offered a basic metaphysical terminology in which we spoke of unities, systems, class identities, organization, structure, and other general categories. Toward the end of the chapter we focused our attention on dynamic systems, especially those endowed with structural plasticity and adaptation.

In this chapter we will analyze dynamic systems more in depth. Our central concern will be, once again, essentially metaphysical. We will concentrate on *the way* in which dynamic systems undergo their changes of state; the 'deep logic', so to speak, of their transitions, trajectory and behavior. For this metaphysical analysis we will use the conceptual tools of cybernetic theory, mainly in the versions of Ross Ashby and Humberto Maturana.

The metaphysical analysis of dynamic systems is important because it will help us to understand a bit more about the nature of living beings. Living beings are dynamic systems, and all that applies to dynamic systems applies to them too. Recall that one of our purposes in this first part of the thesis (Chapters 1 to 5), is to construe a Strictly Naturalistic characterization of living beings; understand their functioning, behavior and relationship to

the environment. To this purpose, of course, it is important to understand what is peculiar about living beings, their exclusive mark within the natural entities. But it is equally (or even more) important to understand what is *not* peculiar about them. We have said in the introductory chapter that living beings, from a Strict Naturalistic viewpoint, may be characterized as:

1. Adaptive dynamic systems
2. Deterministic machines of closed transitions
3. Multistable dissipative systems
4. Organizationally closed systems with respect to (i) their sensorimotor activity, and (ii) their molecular productive processes (i.e., autopoietic systems)

As we shall see along this thesis, out of these features, only 4 (ii) (i.e., autopoietic organization) seems to be an exclusive mark of living beings. All the rest are features that living beings share with all or some of the non-living physical systems. The important point for us is that living beings, in spite of their peculiar autopoietic organization (if they have such organization), must be treated as ordinary physical entities subjected to the same laws, principles and constraints that rule any physical system in general.

In Chapter 1 we already saw that living beings are adaptive dynamic systems (point 1). And we immediately remarked that they are not the only adaptive dynamic systems in the world, that many non-living systems are dynamically adaptive too. Now in Chapter 2 we will review the idea that living beings, like all the rest of the dynamic systems, are deterministic machines of closed transitions (point 2 in our list). We will see that, for those who subscribe to a Strict Naturalistic frame, it should be relatively trivial to point out that living beings are machines (or 'state-dependent systems', according to the cybernetic jargon that we will use later on) that instantiate closed transitions and whose configuration, trajectory and structure are deterministic.

Out of these metaphysical features, probably the most important for us, is the deterministic character of living systems. Ashby, Maturana, and I with

them think that to assume and understand the deterministic nature of the living beings (with respect to their configuration, trajectory and structure) is crucial for understanding their behavior, especially when this is appraised by us as "intelligent".

Most of the conventional explanations about the (intelligent) behavior of living beings tend to appeal to notions or concepts (e.g., agency, control, teleology) that, in one way or another, violate or overlook the deterministic conception of living beings. That is, many of them manage to explain living beings' behavior at the cost (unjustified to our eyes) of conceiving of them as non-deterministic dynamic systems. Why such a conception constitutes a cost, and moreover a high one, should become clear in this chapter.

Strictly speaking, determinism is a metaphysical assumption, a certain basic interpretative commitment about reality and its phenomena. As such, it might or not be true—speaking in a strong realistic metaphysical sense—about the world. (The same runs for indeterminism, its metaphysical opposite)[6]. Those who, like me, want to build a Strict Naturalistic cognitive science, do not need to hold that the physical world *is* really deterministic. We only need to hold that determinism, in comparison to indeterminism, is a more reliable and secure metaphysical assumption for the scientific study of natural phenomena in general, and, by extension, of living beings in particular. Even more, we do not need to claim that determinism is the best metaphysical assumption in every respect and under every circumstance. For example, to those whose priority is not a Strict Naturalistic study of living beings but rather, as in the case of Hans Jonas (1966) and his contemporary enactive followers (Varela, Thompson and Rosch, 1991; Thompson, 2007; Di Paolo, 2005; Froese and Ziemke, 2009), an existentialist and phenomenological interpretation of their constitution and behavior, the right kind of assumption is indeterminism and not determinism. If one wants to see living beings as free agents that face possibilities of action and act according to goals or purposes, what one needs is to conceive of them as non-deterministic systems, not as deterministic systems. Nonetheless, if what we want is a Strict Naturalistic explanation of living beings and their

[6] I will use the terms 'indeterministic' and 'non-deterministic' as synonyms.

behavior, then, so I will argue, it is determinism and not indeterminism that stands as our best metaphysical ally.

Now, if we are not going to be realistic about determinism, if we will offer it only as an adequate metaphysical assumption for the exclusive purposes of the scientific study of living beings, what is its philosophical force? Why determinism and not indeterminism? Although we cannot (and do not need to) show that the physical world is truly deterministic, it will be interesting to observe in this chapter that determinism, and not indeterminism, seems to form part of our primary metaphysical attitude as spontaneous observers (i.e., previous to the subscription to more elaborated metaphysical commitments, Naturalistic or not). We will see that determinism constitutes, so to speak, a sort of 'basal' metaphysical assumption that spontaneously rises from the regularities and coherences of our ordinary dealing with physical objects. In that sense, because of its closeness and smooth coordination with the operational coherences of our quotidian experience, it will be argued that determinism offers us a more intuitive and austere metaphysical frame than its indeterministic counterpart.

We will see that metaphysical indeterminism (not epistemological or moral indeterminism) always strikes us as something exceptional and extravagant, something that does not belong to the regularities of our ordinary experience. Actually, as we shall see, the violation of metaphysical determinism stands as one of the typical marks (though not the only one) of those entities and phenomena that we usually qualify as supernatural. People and cultures that believe in supernatural phenomena do not escape this basic attitude. They too interpret, tacitly, the violation of determinism as something entirely exceptional, something reserved only for extraordinary agents with special powers (deities, witches, etc.). The only difference between them and us Naturalistic people is that whereas they accept said exceptionality as real, we sanction it as fictitious. Yet both they and we recognize the extraordinary character of any physical phenomenon that escapes determinism.

Of course, being nothing more than an assumption, we might choose not the treat living systems as deterministic systems, and I am not here to claim that such an indeterministic strategy is illegitimate or impossible. This

project's only merit, perhaps, is that we make explicit the metaphysical assumption that guides our descriptions and explanatory attempts, and that we try to follow it with fidelity. Cognitive theories that, in one way or another, implicitly or explicitly, do not follow the restrictions derived from assuming a deterministic metaphysics for living systems owe us at least a justification of their choice. In the same way that it is not given, say, as an evident and proved metaphysical truth, that living beings are deterministic systems, so it is not given that they are—or must be seen as—indeterministic systems.

We hold that, when determinism is assumed, a series of ideas about the nature of living beings (conspicuously: agency, freedom of action, control, self-control, teleology, and others) are revealed as explanatory fictions; more or less useful depending on the interests of the observer, but fictions in the end. Cognitive theories that characterize living beings or their nervous systems, implicitly or explicitly, as possessing such properties (e.g., enactive theories that speak of agency and freedom of action (Di Paolo, 2005; Thompson, 2007), embodied theories that conceive of the nervous system as a control system teleologically oriented to certain tasks (Clark and Grush, 1999; Wheeler, 2005)) have a non minor pending task in their agenda. They have to 1) show how such properties are possible in deterministic systems, or, alternatively, 2) justify and elaborate an indeterministic metaphysics for dynamic systems in general, or a regional indeterministic metaphysics for living beings in particular. This, to my knowledge, has not been done so far, and until we receive some news about such a metaphysical justification, I think we are entitled, in the name of a prudent and conservative attitude, to assume determinism as our best version of Naturalism.

In a conservative version of Naturalism, also called classical Naturalism, physical systems are assumed to be deterministic systems. This doctrine, as we all know, has been questioned by some interpretations of quantum physics, where the behavior of the systems under measurement appears to be non-deterministic. The problem with this challenge is that there is nothing, even remotely, like a final word or general consensus among the physicists about the correct metaphysical interpretation of quantum theory (Sklar, 2009; Jaeger, 2003; Jenann, 2009; Rae, 2004). Quantum physicists and

philosophers of physics have been struggling, since the very inception of the theory (which is almost a century now), to find a way of making metaphysical sense of it (Goyal, 2011). (Just to give an idea, you can count in contemporary texts, easily, more than ten different metaphysical interpretations, some of them directly incompatible, of quantum theory. See for example Laloë, 2012). The reason has to do, in part, with the very point under discussion. Many interpretations of quantum theory suggest that fundamental physical reality is non-deterministic, and, as we will see, the violation of physical determinism cannot but strike us, against the background of the coherences and regularities of our ordinary experience, as something metaphysically extravagant and tough to swallow, something that requires exceptional metaphysical conditions.

Physicists and philosophers of quantum theory acknowledge that they have not been able, at least until now, to give us a unified and coherent metaphysical picture of the world and its phenomena (Goyal, 2011; Jenann, 2009; Jaynes, 2003, 1990, 1989). If this is the case, the challenge posed by quantum theory to the classic deterministic version of Naturalism should be taken rather with moderation. It may well be the case that, in the future, quantum theory manages to mature as a coherent and unified metaphysical paradigm (Goyal, 2010), replacing the classical one and giving us a whole new and comfortable ground for building a non-deterministic ontology of living beings, their biological processes and behavior. It may also be the case that, in line with new findings and theoretical developments, quantum theory is reabsorbed into a broader metaphysical frame that remains deterministic (Jaynes, 2003). Who knows? What is clear is that for now, and at least for the specific purposes of a cognitive science, it does not seem very prudent to embrace a non-deterministic conception of living beings on the base of such an uncertain scenario.

Ashby, Maturana, and I with them think that the deterministic conception of biological systems is consistent with our best established Naturalistic metaphysics, and that it is both prudent and reasonable to assume it as a base for the study of living beings.

We will now move on to addressing, point by point, the characterization of living beings as deterministic machines of closed transitions. Our guiding

questions will be: 1) In which sense can it be said that living beings are machines? 2) What does it mean to say that their transitions are closed? 3) In which sense or senses are they deterministic systems? The answers will be given in terms of some fundamental metaphysical principles. 1) Actuality principle (AP): living beings, like any other physical system, are machines in the sense that they exist as actuality. 2) Conservation principle (CP): that living beings instantiate closed transitions simply means that, like any other physical system, they are transitory organizations of a physical substratum that is conserved. 3) Determinacy principle (DP): living beings, like any other physical system, are deterministic systems with respect to their configuration, trajectory and structure.

As with Maturana's metaphysics in the previous chapter, here I also reconstruct, expand and modify to some extent Ashby's cybernetic terminology. Likewise, since I consider that my conceptual contributions in this area are continuous with, and completely loyal to, Ashby's theoretical system, I do not flag them as 'mine' in the exposition.

2.1 Machines

Cybernetics, in Ashby's classical version, is essentially a formal discipline dedicated to the study of machines in general (Ashby, 1956), whatever their constitution (material or formal, real or ideal) and their origin (natural or artificial). From a cybernetic point of view, machines are simply 'state-dependent' systems, or better said, systems whose trajectory is 'state-dependent'.

Every system exists—or may be conceived of as existing—in a certain time. Concrete systems exist in real time, whilst abstract or formal systems can be thought of as existing in a virtual or logical time. The passing of a system from one instant to another (in real, virtual or logical time) is called 'transition', and a sequence of transitions is called 'trajectory'. The transition of a system may be invariant or variant. It is invariant when the state of the system, after the transition, remains the same; it is variant when it changes.

When all the transitions of the system are invariant, we speak of a static or constant system. When one or more of its transitions are variant, we speak of a dynamic or transformer system. By convention, I will use the notions 'static'/'dynamic' when talking about physical concrete systems, and 'constant'/'transformer' when talking about abstract or formal systems. In the following example we have a dynamic system X with a trajectory of seven transitions, one of which is invariant (every transition is represented by an arrow →):

Temporal sequence	... 1	2	3	4	5	6	7	8 …
System's states	... A →	B →	C →	D →	E →	F →	F →	G …

In the following example we have a static system with a trajectory of seven transitions:

Temporal sequence	... 1	2	3	4	5	6	7	8 …
System's states	... A →	A →	A →	A→	A →	A→	A→	A …

A system, dynamic or static (transformer or constant), is a state-dependent system, i.e., a machine, if its current state, at every moment, is the result of the transition (variant or invariant) of *its previous* state, or, which is the same, for every current state of the system, the next state arises as a result of the transition (variant or invariant) of *its current* state.

According to this definition, the dynamic system X in the previous example is a state-dependent system. We can see that the state *B* at the time 2 is the result of the variant transition of the state *A* at the time 1, *C* at the time 3 the result of the variant transition of *B* at the time 2, *D* at the time 4 the result of the variant transition of *C* at the time 3, and so on and so forth. The key point in this pattern of transitions is that, at every moment, only the actual existing (current) state of the system is object of transition. Whether the transition is variant or invariant is not the essential point. Thus, the static

system in the previous example is also a state-dependent system, i.e., a static machine, because every one of its states arises as a result of the invariant transition of the previous state of the system. A state-dependent system, in other words, is essentially a system governed by AP. Let us examine this idea.

The past state of a system is a state that existed at some moment but that does not exist anymore (it has gone). The future state of a system is a state that still does not exist. Both past and future states are inexistent (non-actual) states at the moment in which the transition of the system takes place. In the previous example, at the time 2 the system is in state *B*, so *B* is the only state that can undergo a transition and give rise to the state which follows at the time 3 (i.e., *C*). Why cannot *C* be obtained from *D*, for example? Well, simply because at the time 2 *D* still does not exist. And why can *C* not be obtained from *A*? Well, simply because at the time in which the transition that results in *C* takes place (i.e., time 2), *A* does not exist anymore, it has gone.

A state-dependent system is a system governed by AP, in the sense that only that which exists as actuality (in the 'now') can be an object of transition (variant or invariant). That which does not exist (because it has gone or because it has not been instantiated yet) cannot be the object of any transition (variant or invariant). In other words, to be a machine is to be a system in which only the existent states can be objects of conservation (invariant transition) or modification (variant transition).

Now, if we think about the essential feature of machines, we will realize that although not every machine is a real physical system (e.g., formal machines such as the Turing machine), every real dynamic system (natural or artificial) is actually a machine. We said that a machine is a system whose transition is state-dependent. Well, real physical systems are systems whose succession of states follows precisely that metaphysical order. The reader can go back to the example of system X and see that, for a physical system, there is no way in which *C* could be obtained from *D* or *A*; i.e., from states that do not exist at the moment in which the transition that results in *C* takes place (the reader will find that this is the rule for every state in particular). Real physical systems are always state-dependent systems, namely

machines.

So, what does it mean to say that living beings are machines? It simply means that living beings are state-dependent systems; that they, like any other ordinary physical system, exist as actuality.

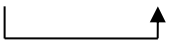
But the reader might say "Fair enough, we can say that we are clear now about the meaning of the concept 'machine', yet that does not mean that we have accepted the idea that living beings are state-dependent systems." What is the alternative, then? Why should not we think that living beings are 'state-*independent* systems'? Let us review, to appreciate the force of the notion of 'machine', cases in which AP is violated.

(i) Temporal order ... 1 2 3 4 5 …

System's states ... A ← B ← C ← D ← E …

(ii) Temporal order ... 1 2 3 4 5 …

System's states ... A → B C D→ E ...



In (i) we have that, at each moment, the present state of the system is the result of the transition of its future state, and that the past state of the system has been the result of the transition of its present state, and so on and so forth. The system is, somehow, 'running in reverse'. This is not a state-dependent system because the states of the system that are objects of transition are non-existent (non-actual) states. For example, a cat has had an accident and, after suffering for some minutes, has died. According to (i), the cat suffers as a consequence of the fact that in the next minute it will be dead; its current state is a modification (a variant transition) of its future state. Is this possible? Well, I do not know. What I do know is that this is not the standard way in which we interpret the dynamic systems that we find in our quotidian experience. Our ordinary metaphysics assumes that the cat

suffers now as a result of its previous condition, not as a result of his future condition.

In (ii) we have a situation in which, at the time 5, the state *E* of the system arises as a direct modification of a state that is not its immediate previous state (*D* at the time 4). According to this regime, for example, the way you look now might not be the result of your accumulated history during the last years but, surprisingly, the direct and immediate modification of your body when you were ten years old. Your body directly jumped, so to speak, from your childhood to your adultness, bypassing all the moments, hours and years in between. The moments that you lived as a teenager were real; it just happens that they took place in metaphysical discontinuity with the rest of your life. Is this kind of metaphysical tunneling possible? Well, I do not know. What I do know is that this is not the ordinary way in which we interpret the historical dynamic of the physical systems.

In our daily life, we assume AP and the state-dependent character of physical systems as a kind of tacit metaphysical commitment, as a condition of intelligibility to interpret their changes over time.

2.1.1 The observer and his 'miraculous machine': the case of memory

Every time we appeal to an inexistent (absent) state as if it were determining in some way the current state or behavior of a concrete system, we are treating that system as if it were not a machine but a 'state-independent system', or, as Ashby would say, as if it were a machine with 'miraculous' properties.

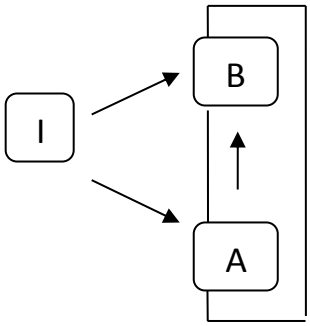
'Miraculous machines', though fictitious from a cybernetic point of view, are not arbitrary creations but well-motivated conceptual constructions. They have to do, essentially, with the kind of epistemological relationship established between the observer and the observed system, and with the tendency in the observer to project into the observed system the

properties of his own observational practices.

In this section we will review the particular case of the concept of 'memory'. Based on Ashby's studies (1960), we will see that 'memory' is a notion that usually emerges as a product of a) the observer's inability to observe the system in all its significant variables, and b) the observer's ability to compensate this ignorance by appealing to non-actual states in the system.

When a system X is not completely observable in all its variables, i.e., when we ignore a part or the totality of its internal mechanisms, it is said that X appears to us as a "black box". When this is the case, what we usually do is to try to infer, on the basis of some observable indicators, the internal mechanisms of the system. This strategy is both legitimate and necessary, but, as Ashby (1960) insists, we have to be very careful in applying it, as we can easily project into the system, without noticing it, properties that belong to our own inferential maneuvers and take them as if they were properties of the internal mechanisms of the system. When we do that, says Ashby, the system may appear to us as endowed with exceptional, abnormal or even miraculous properties. Let's review this idea through some examples provided by Ashby.

Think of a system X composed by two connected parts (A and B) having a common input I.



Suppose the interest of a community of observers is focused on A's behavior; specifically, on whether A exhibits a particular behavior Z. Suppose we are the engineers who have designed and built the system, so that we know that the behavior Z appears in A only under these two

simultaneous conditions: (1) I is at the state *p*, and (2) B is at the state *q*. Suppose additionally that B assumes the value *q* only after I has assumed the value *s*.

Now, let's imagine that two observers are studying A's behavior. One of them has access to I and to every part of the system (A and B); she is a complete observer (CO). The other one, due to certain limitations, has access to I and A, but not to B; he is an incomplete observer (IO). Missing the part B, X appears to IO as a 'black box', i.e., as a system whose internal mechanisms are to a large extent unknown. After following the system for a while, CO arrives (correctly) to the conclusion that Z appears in A whenever I is at *p* and B is at *q*. Thus, if at a given moment the input exhibits the value *p*, she just needs to check the concomitant value of B in order to predict the appearance of Z in A. IO, on the contrary, cannot achieve this predictive accuracy. He observes that A exhibits Z only when I is at *p* (he realizes that I at *p* is a necessary condition for Z), but he also observes that A does not always exhibit Z when I is at *p* (he realizes that I at *p* is not a sufficient condition for Z). Observing that I is at *p*, IO tells us that all that he can assert is that Z might appear in A with a probability of, say, *z*.

IO, nonetheless, does not feel satisfied with this probabilistic prediction. He reviews his records, analyzes his notes, and suddenly he finds something! Attending to earlier states, he realizes that Z appears in A when I is at *p*, but only if the previous state of I has been *s*. His strategy is now different. Observing that I is at *p*, IO relates the occurrence of Z to whether I showed or not the value *s* in a previous moment. Since the sequence $s \rightarrow p$ in I leads invariably to the appearance of Z in A (due to the fact that the transition $s \rightarrow p$ in I leads to the state *q* in B), IO has now the same predictive power as CO. Though IO does not know the internal mechanism that mediates the relation between I and A (the transitions of state in B), he can use the *history* of changes in I to fill in the gap and predict with accuracy the appearance of Z in A.

IO estimates now that the system's behavior in relation to I is regular and fully predictable (he does not need to talk about probabilities any more). But at the same time, he estimates that such regularity has to do with the past states of I, and not only with the current state of the system. Commenting on

their respective findings, IO explains to CO that the system, somehow, exhibits a sort of 'memory' because its behavior is determined, to some extent, by the past states of I. CO, perplexed, cannot see any room for such a fancy feature in the system; to him Z appears in A simply whenever the *current conditions* are I at *p* and B at *q*, without needing the consideration of any past event. At this point IO and CO might engage, if they are quarrelsome, in a long discussion about whether X exhibits memory or not. Yet they could not arrive at any agreement. Why? The reason, says Ashby, is simple. Although they seem to be discussing the same entity, they are actually talking of two quite different systems. When CO talks about the system X he refers to the connected parts A + B, but when IO talks about the system X he refers uniquely to A, as he does not have access to B. CO deals with the system X, while IO deals with the black box X.

The interesting point is that when IO appeals to the notion of 'memory', what he does is substitute his ignorance with respect to X (his inability to observe the part B in X) with a notion that *reflects the way in which he manages to predict X* (i.e., by contemplating past events), but that does not reveal the way in which X actually works. Past events are important for his predictive strategy, no doubt, but as we already saw in the previous section, past events have a null participation in the determination of the current system's behavior. Yet IO does not invent 'memory' in a free and arbitrary way. Given his limited observational conditions and the way in which he manages to predict X's behavior, the idea that certain past events play a role in determining X's behavior is quite reasonable. The notion of memory introduced by IO has to do with an ignorance that is real. That is why Ashby claims that "to invoke 'memory' in a system as an explanation of its behavior is equivalent to declaring that one cannot observe the system completely" (1960, p. 116).

The final (reconstructed, inferred) system for IO is equivalent to I→(A + memory), where 'memory' is a feature of his own epistemic strategy projected into X as a substitute of the missing part B. In this way X becomes a 'miraculous' machine, i.e., a machine whose behavior is determined, in part, by past (inexistent) states. At the same time, though miraculous, X becomes for IO a predictable and regular system.

> *If a determinate system is only partly observable, and thereby becomes (for that observer) not predictable, the observer may be able to restore predictability by taking the system's past history into account, i.e. by assuming the existence within it of some form of "memory".* (Ashby, 1960, p. 115. Original emphasis)

Now, while it is true that IO, by taking the system's past history into account, reaches the same predictive power as OC, it is not the case that he reaches the same explanatory power. CO can not only predict the appearance of Z in A but also point out the mechanisms that generate the appearance of Z. CO can tell us what are, *in the observed system*, the necessary and sufficient conditions that determine the behavior Z in A: (1) I at the state *p*, and (2) B at the state *q*. These conditions are pointed out at an ontological level. That is, they denote the conditions that, once given in the system, trigger the behavior Z in A. IO, not knowing the internal mechanisms in X, can predict Z but not expose the mechanisms that generate it. Of course, he can hypothesize about a 'memory' mechanism, but by doing that he is not revealing any mechanism in X but a feature of his particular observational condition. He also can give us his own version about the necessary and sufficient conditions that determine the appearance of Z in A: The transition $s \rightarrow p$ in I. Yet this particular transition is not the necessary and sufficient condition that determines the occurrence of Z in A, but that which is necessary and sufficient *to know*, for him or for any observer under similar conditions, to be able to predict the occurrence of Z in A. When IO says that $s \rightarrow p$ in I is necessary and sufficient for the occurrence of Z in A, what he means is that, in order to predict Z in A, one needs to know not only that I is at *p* but also that the previous state of I has been *s*. His necessary and sufficient conditions are epistemological conditions, not ontological ones. But why are these conditions not ontological? Why, if both *s* and *p* are states in I, and $s \rightarrow p$ is a transition that takes place in I, not in the mind of IO? The answer is that the only condition in I that plays a role in determining Z in A is *p*, because *p* (not *s*) is the state that in concomitance with *q* in B determines that the following state in A is Z. The state *s* is always the

previous state to the states that determine Z in A; therefore, an *inexistent state* at the moment in which Z is determined. And we already know that that which does not exist cannot determine anything because it cannot be the object of any transition.

What is the moral of the example? Ashby summarizes it as follows:

> Clearly, "memory" is not an objective something that a system either does or does not possess; it is a concept that the observer invokes to fill in the gap caused when part of the system is unobservable. (Ashby, 1960, p. 117)

As I read it, what Ashby is telling us is that 'memory' is more a property of the explanation than a property of the entities being explained.

Now, we have reviewed here an example in which the observed system is extremely simple, composed only by two parts (A and B), and where the internal mechanism mediating between the input (I) and the observed behavior (Z in A) consists just in a transition of state undergone by one element (B). In concrete real systems, nevertheless, the situation is quite different as we face complex mediations that involve hundreds, thousands or millions of components (just think of the brain and its massive connections among millions of neurons), and where the epistemic gap with respect to the internal configuration that determines the observed behavior is so gigantic that the observer, for practical reasons, deliberately ignores it and prefers to explain the observed behavior by making reference to the history of the system and by assuming a sort of internal memory. Let me quote Ashby for a more colloquial illustration.

> [S]uppose I am in a friend's house and, as a car goes past outside, his dog rushes to a corner of the room and cringes. To me the behavior is causeless and inexplicable. Then my friend says, "He was run over by a car six months ago." The behavior is now accounted for by reference to an event of six months ago. If we say that the dog shows "memory" we refer to much the same fact—that his behavior can be

> explained, not by reference to his state now but to what his state was six months ago. (Ashby, 1960, p. 117)

In principle, there is nothing wrong in referring the current behavior of the dog to an event of six months ago. We as observers operating in language can make reference to the past, and that reference is a legitimate tool in the construction of an explanatory proposition. It is a poor explanation (because it does not reveal the internal mechanisms of the organism), but is one that at least satisfies our curiosity. What is the problem then? The problem begins when we assume that the reference to the past is not only a strategy that we, as incomplete observers, use to make sense of the dog's behavior, but also the internal mechanism through which such behavior is generated. Ashby warns us about this point:

> If one is not careful one says that the dog "has" memory, and then thinks of the dog as having *something* […]. One may then be tempted to start looking for the thing; and one may discover that this "thing" has some *very curious properties*. (Ashby, 1960, p. 117. Emphasis added)

"Very curious properties"? Yes, quite curious. Just think about the following point. In which way, if any, could the dog or any other concrete system be in contact with a past event, i.e., with an event that existed in some moment but that does not exist anymore? How could a concrete system recover something from the past? An obvious answer to these questions would be that the past event is not present in the literal sense; the dog does not deal with a *presentation* of the past event but with a *re-presentation* of it. The past is not recovered in a physical dimension (something that looks impossible or metaphysically extravagant) but rather in a referential dimension; something in the present *makes reference* to the past. This answer, although intuitive for most of us, is misleadingly simple. In order to make intelligible the 'memory mechanism', which we have argued—let's not forget—is an explanatory fiction projected by the observer, we have been forced now to invoke nothing less than representational states and

semantic properties. Like the debtor that has to ask for additional money to amortize the original debt, so the observer resorts to more curiosities to explain the already curious properties of his miraculous machine.

We might disregard these questions and leave the issue there, but if the observer is one of those philosophers committed to the 'naturalization of cognition', he will not rest until he finds some way to accommodate these curious phenomena within a naturalistic picture. He will think that the memory mechanism, understood as a representational mechanism, must take place ultimately as a biological mechanism in the organism. He will assume, typically, the existence of internal (neural) representations.

Indeed, this is one of the frequent motivations to invoke (and try to vindicate) the notion of 'internal representation' in cognitive sciences. The observer assumes that the organism's behavior is coordinated or connected with an absent element, temporal or spatially, and since this connection cannot be understood literally in physical terms, he reformulates it in semantic terms positing the existence of internal representations in the organism. Such is, for example, the kind of strategy followed by Andy Clark in his attempt to vindicate the existence of internal representations. He calls attention to those "skills by which some animals [...] are able to maintain *cognitive contact* with [...] absent states of affairs" (Clark, 2001, p. 109). Clark considers the coordination of behavior with non-existent elements a typical instance of what he calls "representation-hungry problems" (Clark and Toribio, 1994); i.e., situations that the organism cannot manage unless it resorts to a certain kind of internal representation.

> The ability to track […] the non-existent requires, prima facie, the use of some inner resource which enables appropriate behavioral co-ordination without constant ambient input […]. Whatever plays that kind of inner role is surely going to count […] as some kind of internal representation. (Clark and Toribio, 1994, p. 419)

Coming back to our example, the story is as follows. The dog, or his brain, has a representation of what happened six months ago, and his

particular behavior in the present has to do with the presence of such internal (neural) representation. The problem with this explanation, although it may be psychologically satisfactory to one or more observers, is that it leaves us in front of an even bigger mystery: how can a portion of matter, in this case a set of neurons, be *about* something or hold referential-semantic relations with certain events of the world? The notion of internal representation, as invoked by Clark's considerations, may well be a way of making the animal's behavior intelligible, and in that sense, a way of explanation. We, who have assumed a Strict Naturalism, are not interested in questioning the psychological goodness of this kind of explanation, its power to make the explananda intelligible or understandable. We are interested in preserving a Strict Naturalistic conception of the neurological mechanisms as biological phenomena, and avoiding any form of over-naturalization, anthropomorphism or observational projection in their description and explanation (we will come back to this point in Chapter 5).

Let us review now, very briefly, the notion of 'closed transition' and the metaphysical principle that grounds it.

2.2 Closed transitions

The Conservation Principle (CP) states that the origin and end of physical unities is always another physical unity or unities. When a new physical unity appears, what appears is a new configuration of a physical reality that was already there; it does not come out of nothing. Similarly, when a physical unity disintegrates or ceases to exist, that means that its physical reality has become something different, not that it has been confined to the absolute nothingness. In cybernetic terms, this can be expressed by saying that the physical universe is a system of 'closed transitions'. A transition is closed if it has both an initial and a final state. A transition is open or undefined if it lacks either an initial or a final state. In the following example we see a trajectory in which the first and the last transitions are open.

Temporal order	...	1	2	3	4	5	6	7
System's states		? →	B →	C →	D →	E →	?	

Recall in the previous chapter we said that if our unity of observation is as generic or basic as 'physical phenomenon', 'physical magnitude' or 'physical unity', then what we have is that everything is conservation through the constant constitution and disintegration of particular entities (be they massless particles, antimatter, dark matter, some kind of energy field, etc.). It is in this sense that physical unities are systems of closed transition. When they arise, they arise as modifications of certain physical reality that was already there. When they disintegrate, their physicality does not disappear but takes form in new unities. In other words, their *physicality* is always conserved.

Under this assumption, open or undefined transitions occur because of the epistemic limitations of the observer, not because of a sudden metaphysical creation or vanishing of the unity under observation. An open transition simply means that its observation is incomplete over time (the observer has missed the antecedent condition of the unity or has ceased to observe it), or that the unity has arisen from, or transformed into, a form of physical existence that the observer ignores or cannot recognize.

Living beings are systems of closed transition simply because that is a trivial condition of any physical unity or system. When living beings appeared on the Earth they appeared as a novel organization of a certain amount of matter and energy that was already there, and the day they extinguish the matter and energy they are will take form in a myriad of inert unities. A living being is a temporary organization of a fundamental physical substratum that is conserved.

2.3 Deterministic systems

The Determinacy Principle (DP) states that physical systems are deterministic systems. Determinism, as a philosophical thesis, may be understood in different ways. In its most austere expression, determinism is a thesis restricted only to the field of physical phenomena (their regularity or law-like nature), without aiming to say anything in particular about the (moral) problem of human free will. In a more ambitious philosophical version, determinism is usually appealed to in order to say something about human free will; for example, that free will does not exist because the world is deterministic. Here I will speak of determinism in an austere sense, and concentrate mainly on three essential sub-principles: 1) Principle of Configurational Determinism (PCD), 2) Principle of Evolutionary Determinism (PED), and 3) Principle of Structural Determinism (PSD).

These three forms of determinism, we will see, seem to operate as basic metaphysical principles in our quotidian experience as spontaneous observers. All or most of the explanatory systems—magic, religious, animist, naturalistic—seem to recognize determinism as a primary metaphysical element, and the violation of determinism as something exceptional, extraordinary or supernatural.

2.3.1 PCD

The Principle of Configurational Determinism states that the configuration of the system (i.e., the total set of its variables) has, at every moment, only a unique value. In cybernetic terms, the configuration of a system is deterministic if its variables are single-valued, i.e., if they assume, at each given moment, a unique value. The configuration of a system is non-deterministic if its variables are multiple-value; i.e., if they assume two or more different values at the same time. Generally, non-determinate machines are expressed in numeric terms, in which case we speak of stochastic or

probabilistic machines. But they can also be conceived of in purely conceptual or propositional terms, in which case we can speak of modal machines.

It is important to note that PCD is a principle restricted only to physical variables. Other kinds of variables are not necessarily subject to PCD. For example, 'nationality', 'institutional affiliation' and other nominal or conventional variables can assume multiple values without any problem. A person can have, at the same time, two or more nationalities, and an academic remain affiliated to two or more universities. With physical variables the issue is different. A person can be a French and Chinese citizen at the same time, but can she physically be, at least for a moment, in those two countries at the same time? An academic can remain affiliated to the University of Edinburgh and to the University of Copenhagen, but can she physically be delivering a lecture in Edinburgh and Copenhagen at the same time? When dealing with physical variables, PCD seems to be the rule.

PCD assumes that any indeterminacy with respect to the configurational value of physical systems is due to some epistemic limitation in the observer, not to an alleged intrinsic indeterminacy in the observed systems. According to PCD, configurational indeterminacy is always epistemic indeterminacy; i.e., uncertainty.

For example, let's say that I want to know the temperature of a system at a given moment, and that my thermometer is such that it can give me only a range of probable values. The instrument says that at the time 1 the temperature of the system is between 33 and 36 degrees Celsius. Must I assume that the system has, *at the same time*, 33, 34, 35 *and* 36 degrees? Well, that is not our most common interpretation. What we usually do is to assume that the temperature of the system at the time 1 has a unique value, a value that may be, disjunctively, 33, *or* 34, *or* 35, *or* 36 degrees, but not all of them at the same time! We assume that the temperature of the system is a determinate property (i.e., single-valued), even when our knowledge of its punctual values remain undetermined.

Now, if I build a mathematical model to describe the thermal trajectory of the system using the data provided by my thermometer, what I get is a stochastic system; a mathematical formalism that works with probabilistic

values. Such a system is a non-deterministic formal system with respect to its configuration (not necessarily with respect to its evolution). Nonetheless, the real system whose temperature I am trying to follow remains a deterministic system. According to PCD, to assign a stochastic (probabilistic) ontology to the real system would be as inappropriate as to assign certain temperature to the mathematical model.

If we assume PCD, as it seems reasonable to do, stochastic or probabilistic machines exist uniquely as formal entities; that is, as abstract systems that we construct to describe, model, explain or predict the behavior of certain concrete systems. Under this assumption, probabilities express, quantitatively, degrees of certainty or uncertainty in the observer, not aspects or features of the concrete observed systems. This is tantamount to assuming that every real physical system is always a single-valued system of variables, i.e., a configuration-determined system. As long as living systems are real physical systems, it seems fair to treat them as configuration-determined systems too.

2.3.2 PED

The Principle of Evolutionary Determinism simply means unique trajectory. The evolution or trajectory of a system is unique if, given exactly the same structural conditions (both in the system and in its surroundings, when the system interacts with its surroundings), the system exhibits exactly the same sequence of states, the same trajectory. The evolution of a system is not unique (i.e., indeterministic) if, given exactly the same structural conditions (both in the system and in its surrounding, when the system interacts with a surrounding), the system exhibits a different trajectory. The classical image to illustrate PED is that of a videotape (Bishop, 2002). Suppose that you are watching a film X and that you want to repeat some scenes. You rewind the tape, run it, and watch the same scenes again. That is unique evolution (normal and familiar stuff). Now suppose you want to do the same with the

film Y. You rewind the tape, run it, and... Surprise! The film does not show the same scenes; it is a different story now. Y is an indeterministic system with respect to its evolution.

Like PCD, PED assumes that any indeterminacy in the evolution of a system is due to some epistemic lack in the observer, not to a metaphysical indeterminacy in the observed system. This is actually the principle that Ashby applies in his analysis of 'memory' reviewed before (The machine appeared as evolutionarily indeterministic only to the incomplete observer).

If the observer sees that a system X exhibits, under what she takes to be exactly the same conditions, different behaviors, and if she subscribes to PED, then she will assume that the behavioral difference has to do with the intervention of some variable that is being ignored (a 'hidden' variable whose existence, in principle, might be discovered at some moment), not with an alleged intrinsic randomness or capricious freedom in the system. This way of reasoning, as Jaynes observes, is the way in which standard biological science has made some of its most valuable discoveries.

> In biology or medicine, if we note that an effect *E* (for example, muscle contraction, phototropism, digestion of protein) does not occur unless a condition C (nerve impulse, light, pepsin) is present, it seems natural to infer that C is a necessary causative agent for E. [...] But suppose that condition C does not always lead to effect E; what further inference should a scientist draw? (…)
>
> In the biological sciences, one takes for granted that in addition to C there must be some other causative factor F, not yet identified. One searches for it, tracking down the assumed cause by a process of elimination of possibilities that is sometimes extremely tedious. But persistence pays off; over and over again, medically important and intellectually impressive success has been achieved, the conjectured unknown causative factor has been finally identified as a definite chemical compound. Most enzymes, vitamins, viruses, and other biologically active substances

> owe their discovery to this reasoning process. (Jaynes, 2003, p. 327)

In this thesis we will apply the same reasoning with respect to living beings, their behavior and their subsystems. If a living being shows, under what we take to be exactly the same circumstances, different behaviors, we will assume that that is because of the intervention of some ignored or not yet identified *structural* factor, not due to an alleged intrinsic freedom of action or ability to choose.

2.3.3 PSD

The principle of Structural Determinism has to do with the source of specification of the structural states of the system. As stated by Maturana, a system is structurally determined if its structural state, at every moment, is specified by its own structure and not by external factors. External factors can interact with the system and *trigger* (or not) some change of state on it, but they cannot specify (define, determine) the structural result of said interaction (Maturana, 1975, 1981, 1987, 2003). PSD states that every time the system receives the action of an external factor, it is the current structural state of the system, and not the external factor, that (i) defines whether this action triggers or not a structural change on it, and, in the positive case, (ii) specifies the concrete structural change that takes place. In cybernetic terms this means that, with respect to the observed system, external factors act only as operative factors (operators), never as instructive factors (instructions), or, equivalently, that the interactions of the system are always operative interactions, never instructive interactions. Let us review these ideas through some examples.

A person presses a button on a laptop with his finger and as a consequence of this the laptop turns on. After a couple of minutes the same finger presses the same button in the same laptop and, contrary to the

previous case, the laptop now turns off. What has happened? We have the same elements interacting in the same way but ending in different results. Well, there is nothing mysterious about that. Although the finger is pressing the same button with the same force, the current structural state of the laptop is different in each case and, consequently, so too the structural change that takes place. The finger encounters the laptop in different structural states, and it is these structural states that specify the nature of the changes in each case. The mechanical interaction with the finger triggers the change of state that is possible at every moment according to the current structural state of the system, but it does not specify or instruct the nature of said change.

The laptop in its turn only reacts in the way in which its structure allows. By pressing the button it may turn on or turn off, but not dance, cry, or cook a pizza. The point is not that it cannot react in these ways but that for doing such things *it would need a different structure*. The structural changes undergone by the laptop are never arbitrary; they are always determined by its structure.

On the other hand, the fact that it is the finger of that specific person and not some other element that triggers certain structural change in the system is, from the point of view of the laptop, absolutely irrelevant. If the laptop is off, it will turn on whenever *something* interacts with the proper button in the proper way. That something may be a finger, a pencil, a stick, a stone, a screwdriver, etc. That is, a disjunctive series of objects. Nonetheless, to the laptop it is all the same; simply a transition from 'off' to 'on' or from 'on' to 'off'. Its structural dynamic is absolutely blind to the distinctions that we as external observers can make concerning the different triggering objects in its environment. The 'true' origin of its structural changes, so to speak, never appears as such for the laptop, and the laptop, on the other hand, does not need to 'know' that to make its transitions.

Even more, for the effects of its structural dynamic, what counts, at every moment, are the structural states in themselves, not the way, whether exogenous or endogenous, in which those states are brought about. When the laptop is 'on', it is 'on' irrespective of whether this state was reached through an internal evolution or through a change triggered by some external factor (e.g., the laptop may have an automatic internal function to

periodically turn on and turn off). We will come back to this point later on when talking about perception in Chapter 5. (It will be argued that the nervous system, from the point of view of PDS, works like the laptop; i.e., its structural states have the operational efficacy that they have independently of the way, endogenous or exogenous, they come about).

The important point is that a structurally determined system always exists in its own domain of structural states and changes, without distinguishing the origin of such structural states and changes.

Now, let us suppose that after six years of use, the laptop suddenly does not turn on anymore. The person insistently presses the button with his finger, but the laptop remains off. What has happened? What has happened is that after six years the structure of the laptop has changed in such a way that now its responses are different. Its current structural state is such that the mechanical interactions through the button do not trigger in it the changes of state that were usual in the past.

The person feels disappointed. Perplexed, he scratches his head and tries to understand the problem: "Wait a moment… That's it! The finger! It must be my finger!" The person, convinced that it is his finger that determines the changes of state in the laptop, runs to the doctor and asks him to check it. "Please, could you check what is wrong with my finger? It is not able to turn on the laptop anymore. I need you to restore its normal functioning!"

If we think that the demand of the person is nonsensical—says Maturana—that is because we assume, as a kind of tacit condition, PSD. We know that here the 'problem' has to do with the structure of the laptop and not with the nature of the external factor (the structural condition of the finger). We know that our fingers are able to turn on only those systems whose structure admits the state 'on' in its domain of states (working TVs, radios, mobiles, laptops, and electronic devices in general). None of us expects that our finger is able to turn on a table, a stone or a glass of water. After six years of use, the laptop has become a system whose structure, like the table, the stone and the glass of water, does not admit the state 'on' within its domain of states. 'Turning on' is not an intrinsic causal power in our fingers, and if it was such, then our finger would act as an instructive factor, not as an operative one. What is the difference between an instructive

factor and an operative factor?

An instructive factor is an external factor that violates PSD. An operative factor is an external factor that respects PSD. Can we see any example of an instructive factor? Yes, we can, but since the principle of structural determinism is valid for all real systems—says Maturana—we can find the example only as a fiction or myth. And the myth exists. There was once a king named Midas who was given the power of transforming whatever he touched into gold. It did not matter the structural nature of the objects that he touched; apples, cups, water, stones, animals, everything was converted into gold. Midas's finger was not an operative factor interacting with the objects; it was a fully intrusive factor dictating, unilaterally, the nature of the structural changes in every one of them. Midas's finger had the power of violating the structural auto-determination of the systems, which, after all, was not a very good thing. According to some versions of the story, Midas forgot the blind and indiscriminate character of his power in the moment when his daughter came to him. Caressing her, he gave her a golden death (Maturana, 1978b).

The fact that we find the violation of PSD in the form of a fiction or myth is telling. If we pause to think about it, we will realize that one of the marks of those phenomena that we usually qualify as 'magic', 'miraculous' or 'supernatural' is that they operate outside PSD. Witches and wizards can, if they want, 'charm' a prince and transform him into a frog. Jesus had the power of transforming water into wine, of healing blind people just with a touch. What is so extraordinary about these beings and their actions? Why do they seem supernatural to us? What seems supernatural to us is that these beings can interact with the systems specifying (instructing) the kind of structural change that they want to effectuate, irrespective of whether these changes are possible or not according to the structural determination of the systems.

PSD is a central assumption for scientific practice. If I want to provide a scientific explanation of a system X, I need to assume X as determined in its own structure, not as subject to instructive interactions. This is the rule too, says Maturana, when the system under consideration is a living being. The living being and its environment are in constant interaction, and this

interaction must be understood as a mutual triggering of structural changes wherein each part acts always as an operative factor for the other, not as an instructive factor.

So far we have presented PSD as a simple metaphysical concept, trying to understand its core meaning through examples and illustrations. In a more formal exposition, PSD is a complex concept that entails, according to Maturana (2003, pp. 61-62), at least four domains of structural determinism:

> [C]omposite unities [i.e., systems] are structure-determined systems in the sense that everything that happens in them is determined by their structure. This can be systematically expressed by saying that the structure of composite unity determines it at every instant:
>
> a) the domain of all the structural changes that it may undergo with conservation of organization (class identity) and adaptation at that instant; I call this domain the instantaneous domain of the possible changes of state of the composite unity.
>
> b) the domain of all the structural changes that it may undergo with loss of organization and adaptation at that instant; I call this domain the instantaneous domain of the possible disintegrations of the composite unity.
>
> c) the domain of all the different structural configurations of the medium that it admits at that instant in interactions that trigger in it changes of state; I call this domain the instantaneous domain of the possible perturbations of the composite unity.
>
> d) the domain of all the different structural configurations of the medium that it admits at that instant in interactions that trigger in it its disintegration; I call this domain the instantaneous domain of the possible destructive interactions of the composite unity.

These four domains are assumed as valid for every interacting dynamic

system. Out of them, here we will concentrate on a) and c).

The domain a), well viewed, corresponds to the adaptive domain of a system. According to Maturana, adaptation, recall, is conservation of organization (class identity) through, or in spite of, structural changes in interaction with the environment. Maturana is now introducing the idea that every phenomenon of adaptation, either passive or active, is subject to PSD. In order to underline this point, i.e., that every phenomenon of adaptation is subject to PSD, Maturana usually uses the expression 'structural coupling' (Maturana, 1980, 2002, 2003). 'Structural coupling' is a process of structural interaction wherein the system and its environment, or the system and other systems, act as mutual operative factors. Thus, to say that a system exists in structural coupling is to say that it exists in adaptation, but emphasizing the idea that said adaptation cannot be explained in terms of instructive interactions.

The domain c) is interesting because it allows us to see that every dynamic system, passive or active, living or not, is always a selective system regarding its interactions. Every system, because of its structural determination, interacts with certain elements or aspects of the environment, with certain structural configurations, and not with others. A piece of marble, in its structural composition, determines a particular domain of perturbations wherein, for example, a subtle touch of my finger cannot trigger any change of state. My mobile phone, in its structural composition, specifies a domain of perturbations wherein a subtle touch of my finger (on its touch screen) does trigger a change of state. In this situation, there is a sense in which we might say that, for the marble, the subtle touch of my finger is something that does not form a part of *its* environment, to the extent that, as an external factor, it never appears in its domain of perturbations. We might also say, complementarily, that the touch of my finger is something that does exist in the mobile's environment, because it appears as a part of its domain of perturbations.

For the marble, the action of a hammer counts as a perturbation that triggers certain structural change. For the mobile, the action of a hammer is excluded from its domain of perturbations because it rather belongs to d), i.e., to the domain of its destructive interactions.

What for X is a perturbation, for Y may be a neutral element, whereas for Z a destructive interaction. Every physical system, in its structural determination, specifies its own domain of interactions, and therefore, its own domain of structural coupling and adaptation. Every physical system, in a certain way, selects an own environment from a broader surrounding. We might use 'niche' to refer to the particular environment that each physical system specifies as its own domain of perturbations, structural coupling and adaptation.

Living beings, as structurally determined systems, constitute a variant of the same theme. They specify a domain of perturbations, and to that extent, a domain of structural coupling and adaptation; i.e., a niche (more about this in Chapter 4).

2.4 Brief final comments

A dynamic system is a machine if its trajectory is state-dependent. A machine is fully deterministic if its configuration, evolution and structural interaction are deterministic. A machine is fully indeterministic if its evolution, configuration and structural interaction are indeterministic. Any condition in between makes the machine partially deterministic (or partially indeterministic).

In this chapter we have tried to show that living beings are fully deterministic machines, or at least, that for the specific purposes of their scientific study there are not compelling reasons to think otherwise. The consequences of assuming determinism in living beings are substantive. Here we will only enumerate some of these consequences. Their impact upon the way in which we understand living beings' behavior will be reviewed in subsequent chapters.

First, a fully deterministic machine has no possibilities or alternatives of action. A fully deterministic machine is where it is and has the structural state that it has, at every moment, as a result of a deterministic chain of transitions of state. Any other possible scenario or state that we can imagine

for such a system is just that, an intellectual construction that belongs to our operation as observers able to conceive of counterfactual situations.

If living beings are assumed to be fully deterministic machines, then they are assumed to be systems that have no freedom of action. They react as they react and do what they do, at every moment, because their deterministic nature allows them no other possibility.

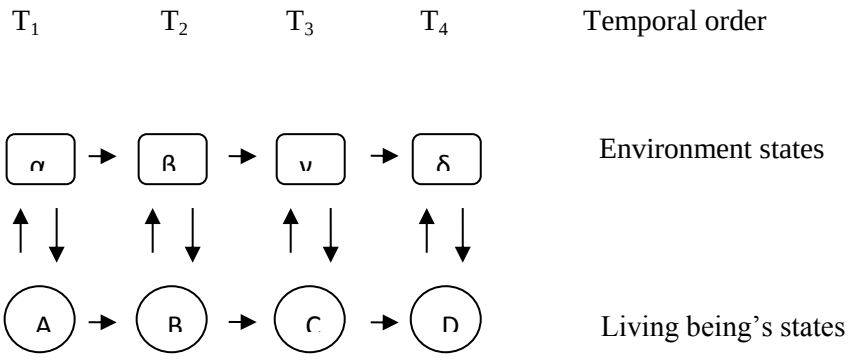
Second, since a fully determinist machine never has alternative ways to act, it cannot exert any control or regulation upon its behavior, basically because there is no metaphysical margin for that. To the extent that, as it has been argued here, natural systems are assumed to be fully deterministic machines, their structural trajectories and behaviors cannot be the result of processes of control or regulation. Planets and rivers do not control or regulate their trajectories, they simply move according to a deterministic sequence of changes of state. Volcanoes cannot control or regulate their eruptions, they simply erupt when their structural conditions so determine it. If living beings are assumed to be fully deterministic machines, then their behavior cannot be the product of processes of control or regulation either. Living beings behave and act without having any control or regulation upon their acts.

Third, a fully deterministic machine is a structurally determined system. A structurally determined system, as was illustrated in the example of the laptop, is a system that entirely exists in its own domain of structural states, without being able to, and without needing to, distinguish the origin of said structural states (whether exogenously triggered or endogenously generated). When a structural change is triggered by an external factor, all that counts and exists for the system is the structural change itself, nothing more. That is, the external factor never appears as what it is for us observers, namely the causative and responsible agent of said structural change. If living beings are assumed to be structurally determined systems, then their interactions with the environment are subject to this metaphysical condition.

Living beings, like any interacting dynamic system, exist in a continuous process of structural coupling with the environment, i.e., in adaptation. This process is a natural phenomenon that, as such, takes place within the metaphysical margins we have delineated in this chapter. That is, living

beings' adaptation is a process of unique evolution (PED) wherein both they and the environment operate as strictly deterministic machines (AP) with respect to their configuration (PCD) and structure (PSD).

The following diagram aims to summarize these ideas, showing a living being in structural coupling with its environment (vertical arrows represent operative interactions):



The main message, for now, is simply that living beings' adaptation is a fully deterministic process (more about this in Chapter 5). But notice the following point.

Suppose we accept, following Ashby and Maturana, the idea that living being's adaptation is a fully deterministic phenomenon. We still would need to explain living beings' adaptation. For adaptation, as a concept, recall, is just a way of naming a certain condition or phenomenon, not an explanatory notion.

In what follows we will examine certain peculiarities of living beings regarding their structural dynamics and functional organization. This will help us to explain and understand, at least in part, their behavior and their adaptation.

Chapter 3

Stability: the apparent teleology of living beings

So far we have characterized living beings as adaptive dynamic systems whose metaphysics, *qua* dynamic systems, is strictly continuous with the rest of dynamic systems in nature. We have said, in the name of a Strict Naturalism, that living beings are physical machines of closed transitions, and that their behavior, whatever the level of "intelligence" or sophistication we want to attribute to it, is generated under ordinary conditions of structural, evolutionary and configurational determinism.

In this chapter we will examine another important property of living beings, namely their stability. Stability is a property that we can find in many systems, not only in living beings, yet its manifestation in biological systems has a special connotation for us. All stable systems—as Ashby will show us—generate to greater or lesser degree a characteristic behavioral pattern that appears to be teleological (i.e., oriented at some goal or purpose). Living beings, as stable systems, are not the exception to this pattern but rather the most representative and strongest case. The main idea of this chapter is that the apparent teleology of living beings is (i) precisely that, an appearance, and (ii) that behind said appearance what exists is a complex form of stability.

Our strategy will be as follows. We are going to start by considering teleology as a valid appearance in our observation of living beings, i.e., as something that, being an appearance, has a real ground in the observed system. Then we are going to try to explain how and why this teleological

appearance emerges in our observation of living beings. My hope is that by understanding living beings as stable systems we will be able to understand, at least in part, the peculiarity of their behaviors without succumbing to a (tempting but not Strictly Naturalistic) teleological interpretation. Ashby puts it, as always, in a clearer way:

> No teleological explanation for behavior will be used. It will be assumed throughout that [...] an animal behaved in a certain way at a certain moment because its physical and chemical nature at that moment allowed it no other action. […] Any [teleological] explanation would […] involve a circular argument; for our purpose is to explain *the origin* of behavior which *appears to be* teleologically directed. (Ashby, 1960, p. 9. Emphasis added)

The way Ashby puts things is interesting. Ashby does not say that there are some behaviors that are teleologically oriented, but that there are some behaviors which *appear to be* teleologically oriented, and that the task is to analyze the *origin* of such an appearance. Accordingly, our question is not "How can living beings behave in a teleological way?" but "Why does the behavior of the living being, a particular class of deterministic physical machine, appear to us *as if it were* driven by certain goals or purposes? What is the origin, the ground, of such an appearance?"

3.1 Appearances

Appearances, in many cases, are not arbitrary errors or caprices of the observer but phenomena that have a real ground in the observed systems. Because of this, some of them may be persistent and difficult to eliminate, even after knowing that they are mere appearances.

Walking in the middle of the desert, suddenly you see a pond surrounded by plants. Whether the pond is really there or is just a mirage you really do

not know at that moment. Let us suppose that a helicopter is flying over the zone where you see the pond located, and that the pilot communicates with you by mobile phone. While you are contemplating the scene, the pilot is telling you that in that area there is no pond. He sends you an aerial picture in which all that you see are some plants. Do you, can you, stop seeing the pond? Well, you cannot stop seeing the pond because the mirage is not an arbitrary mistake of your visual apparatus. On the contrary, given the lighting conditions, the ambient temperature, the angle of the soil with respect to you, the refractive index in the local atmosphere, and the state of your biological machinery, the image of a pond is what you have to see (i.e., the mirage is actually the expected and normal response). There is nothing wrong about that.

If you take the mirage as a grounded and valid phenomenon, then you can ask for an explanation. How does it happen that I see a pond in circumstances that there is no pond? What are the physical and biological processes that generate this experience? This question is both valid and fruitful because you are trying to explain the appearance while keeping the appearance as what it is; a phenomenon that has to do with the particular relation established between your condition as an observer and the state of affairs in the desert, not with the state of affairs in the desert in itself.

Now, suppose you are also curious about the presence of the plants. Why are those plants there in the desert? How do they manage to live there? A good and simple answer would be "because they get water from the pond." Unfortunately, you cannot appeal to the pond as an explanatory element, as plants do not get water from mirages! The pond, as a mirage, is a real and legitimate phenomenon in your experience, a natural result of a well-determined arrangement of physical and biological conditions. Yet that does not mean you can use it to explain other properties or phenomena found in the desert.

In a more quotidian example, we know that the geocentric appearance is a perfectly justified phenomenon for terrestrial inhabitants. While you are contemplating the sunset, the astronomer reminds you that the Sun is not moving around the Earth but that it is the Earth which is rotating on its own axis generating the apparent movement of the Sun. Of course, you know all

that, but... Do you, can you, stop seeing the Sun going down? No you cannot (nor can the astronomer), because the geocentric appearance is something perfectly determined by your particular condition as an observer (your location), and the relative movement of the Earth and the Sun. You can explain the geocentric appearance by appealing to some astronomical knowledge, but you cannot modify your geocentric subjective experience by appealing to such theoretical knowledge. Again, it is an appearance, but not an arbitrary whim. More importantly, you can explain the geocentric appearance by appealing to the heliocentric organization of the solar system and the rotation of the Earth. But it would be a great confusion, after explaining the geocentric appearance in those terms, to explain the duration of days by saying that the Sun takes 24 hours to orbit around the Earth.

The idea, I think, is clear enough. Once we manage to show the mechanisms that generate a determinate appearance in our observation, once we understand its ground, we cannot take said appearance as an operating element in the observed system, nor use it as an explanatory principle. Once explained, appearances can be recognized as valid phenomena of our observation, but not as properties of the observed systems; not even, as some advocates of 'complex systems theory' would like to say, as alleged 'emergent properties'. In our previous example, the apparent geocentric organization of the solar system is not an 'emergent' property of the system; it is just an *apparent* property to the eye of the observer, which is something different.

In what follows we will examine stability as a natural property of certain dynamic systems, including living beings. In particular, we will try to explain the teleological appearance of living beings in terms of their stability as homeostatic systems.

3.2 Stability

Living beings remain alive to the extent that a set of metabolic or physiological variables (usually called critical or essential variables)

maintain their values within certain specific ranges (called physiological or metabolic ranges). Living beings' ability to maintain, in spite of disturbances, their physiological or metabolic condition within these ranges is what is usually known as homeostasis. Living beings' homeostasis is a particular version of stability, which is a relatively common property among dynamic systems. What is stability?

To talk about stability we need basically two things. First, we need a constant condition in the system; a set of variables with invariant trajectories, or a set of variables with bounded variant trajectories that exhibit some constant pattern (e.g., a regular cycle or oscillation). That is, we need the system to be in some kind of stationary or steady state. Second, we need some perturbation (natural or induced by us) that takes the system out of its steady state, i.e., we need a disturbance. (A disturbance is thus a particular kind of perturbation; therefore, and by definition, never a destructive interaction. See again the four domains of structural determinism distinguished in Chapter 2). The constant condition or steady state provides the baseline respect to which we can estimate whether or not the system exhibits stability. If the system, after being disturbed, spontaneously returns to its baseline, i.e., if it recovers its previous condition without external help or assistance, we say the system is stable (with respect to those variables or aspects under consideration). If the system does not return to its baseline, we say the system is unstable (with respect to those variables or aspects under consideration).[7]

The mere observation of a constant condition in a system does not tell us whether or not the system is stable. A system, or some of its aspects, may remain constant just because there is no disturbance; perhaps because the system is not receiving any perturbation at all, or because perturbations do not disturb the constant condition under consideration (in this latter case we say the system is neither stable nor unstable but rather 'neutral'). Whether or not a system is stable, thus, depends on the way in which it responds to a disturbance.

[7] There are cases in which the system has more than one stable point (multiple attractor systems), and whose stability does not necessarily express itself. For the sake of the exposition, we will put aside those more complicated cases.

The crucial point is that every stable system, when displaced from its baseline and then released, exhibits a line of behavior that *returns* to its initial state (Ashby, 1947). It is as if the system, in spite of disturbances, "insisted" on maintaining its steady state. This property of stable systems, as we shall see, can be easily described, if the observer wishes, as revealing an internal drive in the system that leads to a certain final state; i.e., as if the system were trying to reach some goal (to recover its steady state). When the stable system is simple enough, the observer, in general, does not feel a strong motivation to use teleological descriptions or explanations. The case is different, nonetheless, when the stable system is highly complex. The relative simplicity or complexity of a stable system has to do, mainly, though not exclusively, with factors such as its dimensionality (the number of variables or aspects in which the system exhibits stability), its structure and thermodynamic regime (e.g., whether or not the system exhibits an endogenous dynamic, whether the system has a closed or open thermodynamic regime), the presence or absence of feedback mechanisms (both positive and negative), the order or level of stability (e.g., first-order stability, second-order stability), and the system's stability composition (simple or polystable). Living beings, as we shall see, are highly complex stable systems. But before reaching the high complexity of living beings, let us examine more basic examples of stability.

A system that remains in a constant condition or steady state is a system that, in most cases, is in some state of equilibrium (static or dynamic, mechanical or thermodynamic). That is, most steady states are equilibrium steady states. Non-equilibrium steady states are less common, but highly relevant for our purposes. As we shall see, living beings, considered as thermodynamic systems, are special cases of stable non-equilibrium steady states.

Let us start, following Ashby's classical presentation (1960), by reviewing simple cases of equilibrium steady states, or equilibriums, for short. Equilibriums may be stable, unstable, or neutral. To illustrate this distinction, think of three different objects resting on some horizontal surface and disturbed by some external force. A cube resting with one face on a table exemplifies stable equilibrium. A cone balanced on its vertex exemplifies

unstable equilibrium. A ball resting on the same table exemplifies neutral equilibrium. I take the illustration to be quite clear and simple, but notice that these are cases of mechanical static equilibriums wherein the systems have a fixed and limited amount of available energy (i.e., their behavior is restricted to this energetic invariance). Thus, although the cube is a clear example of stability, its behavior does not look too interesting, at least not for us who are trying to understand living beings' functioning.

More interesting, says Ashby, is the case of stable systems with continuous energetic supply. As examples, Ashby offers the Watt governor (fed with a continuous flow of steam), and the thermostat (usually fed with electrical power). These systems are able, in front of transient disturbances, to bring their critical variables (speed of steam flow and temperature, respectively) back to their baseline values, thus demonstrating stability. Yet these systems are different from the cube not only because of their energetic regime, but also due to the way in which they reach stability. They, but not the cube, are organized as functional circuits with inverted polarity (negative feedback), such that the very disturbance produces, through some mediating mechanism, the compensation that brings the system back to its equilibrium.

Living beings, as we shall see, are also systems with continuous energy supply whose physiological stability or homeostasis, in most cases, is preserved by means of a set of servomechanisms or negative feedback loops. There is, of course, a long distance between Watt governors or thermostats and living beings, but for now let us pause on the simpler cases and try to see what Ashby wants to show us about them.

3.3 The appearance of teleology in stable systems

Ashby says that all stable systems, even the simplest ones, may be described and interpreted in teleological terms. Why?

> Every stable system has the property that if displaced from a state of equilibrium and released, the subsequent movement is so matched to the initial displacement that the system is brought back to the state of equilibrium (Ashby, 1960, p. 54).

It is this pairing of the line of return to the initial displacement, says Ashby, what usually motivates in the observer the ascription of teleology to stable systems. A simple example is provided by the ordinary pendulum. When displaced to the right, the pendulum develops a proportional force that tends to move it to the left. When displaced to the left, it develops a proportional force that tends to move it to the right. In both cases, after a couple of oscillations, the pendulum recovers its static equilibrium at the central position, thus showing stability.

> Noticing that the pendulum reacted with forces which though *varied* in direction *always* pointed towards the centre, the mediaeval scientist would have said 'the pendulum seeks the centre'. By this phrase he would have recognized that the behavior of a stable system may be described as 'goal seeking' (Ashby, 1960, p. 54. Emphasis added)

The behavior of a stable system, says Ashby, may be described in teleological terms as 'goal-seeking'. Why? I emphasized the words 'varied' and 'always' in the quotation because they capture, I think, the key point about stable systems' behavior. All stable systems exhibit a typically convergent behavioral pattern; i.e., no matter which way they are displaced from their steady state (the variability of the disturbances), they always return to the same steady state. In the example, regardless of the angle of the displacement, the pendulum will return to the same state of equilibrium (the resting position). It is this combination of variability (by the side of the behavior) with invariance (by the side of the steady state), that gives the idea of 'flexibility' in the system. The system, somehow, seems to have a fixed

goal around which it is able to vary and 'accommodate' its behavior according to the different circumstances. Again, it is as if the pendulum, in spite of disturbances, insisted on maintaining its equilibrium. That is why Ashby says that every stable system *may be* described in teleological terms. But what does that mean? Does that mean that stable systems really try to reach some goal? Is the pendulum really seeking its resting state? Does it move towards the centre because that is its purpose? Perhaps the mediaeval scientist, with an Aristotelian mindset, might have replied "yes". We, instead, who have subscribed to a Strict Naturalism, cannot. Pendulums are simple mechanical systems, absolutely blind to any purpose or goal. The fact that we *can* describe a stable system in teleological terms does not mean that the system *is* teleological; it just means that the teleological description captures an *appearance* that is not arbitrary but that has a real ground in the observed system.

Now, when dealing with pendulums, I guess, most of us do not find a teleological description or explanation terribly attractive, as simple physical variables are enough to explain their behavior. When dealing with living systems, however, the situation seems to change. Why is this so? Why are we, post-Aristotelian observers, so prone to attribute some kind of teleology to living beings? One might say that we humans simply tend to project features of our subjective experience to entities which are close to our genus, and that living beings, without any doubt, are closer to us than pendulums. But that comment, even if true, does not explain the apparent teleology of living beings as a function of living beings themselves; it just expresses, at most, a human bias. The question is "What is peculiar about living beings, among stable systems, such that their behavior appears to be teleological?" The answer, if we follow Ashby, has to do with the complexity of living beings as stable systems.

"Complexity" is a word that needs careful handling, though, as sometimes it is read with too much metaphysical enthusiasm. Some readers, perhaps followers of the so called 'paradigm of complexity', might say "Complexity brings emergent properties, so teleology might be an emergent property of living beings!" That is not the way we talk about complexity here. Living beings' complexity as stable systems has to do simply, as we

shall see, with their interconnected multidimensionality, their nonequilibrium thermodynamic condition, and their ultrastable functioning; all features that enrich or complicate the way in which living beings generate their behavior as stable systems, but that do not introduce any metaphysical novelty or exceptionality, at least not in terms of teleology. An apparent property, recall, is not an emergent property.

Ashby's idea is simple. The structural and thermodynamic properties of a stable system, in combination with the properties of its surrounding, condition the way in which the system recovers its steady state. Sometimes these properties are such that the system, in returning to its steady state, describes a simple and straight line of behavior (i.e., without deviations with respect to its steady state). But sometimes this is not the case, and the system reaches its steady state only after going through more or less large, and more or less long, deviations. For example, a metallic ball resting on a horizontal surface (state R of static mechanical equilibrium), is placed at the top of an inclined plane (state A) and then released. The ball, rolling down, goes back from A to R describing a simple and direct line of behavior. Suppose now that the ball is not a perfect sphere; its surface has some irregularities and its centre of gravity is displaced. Suppose the inclined plane is not smooth, or better, that has some little bars (i.e., "obstacles"). Suppose that flows of air, some of them lateral, some of them upward, and strong enough to move the ball, cross the plane from time to time. The ball, subjected to all these factors, will go back sooner or later to its state of equilibrium R, but this time describing an indirect and complicated line of behavior. We will see the system passing through intermediate states, say B and C, which deviate from its state of equilibrium.

We can complicate the example adding more disturbances, servomechanisms, or endowing the ball with some kind of endogenous dynamism. In any case, says Ashby:

> The fact that the line of behavior does not run straight from A to R must be due to some feature in the 'machine' such that if the machine is to get from state A to state R, states B and C must be passed through of necessity. Thus, if the

> machine contained moving parts, their shapes might prohibit the direct route from A to R; or if the system where chemical the prohibition might be thermodynamic. [I]n either case, if the observer watched the machine work, and thought it alive, he might say; 'How clever! [The system] couldn't get from A to R directly because this bar was in the way; so [it] went to B, [then] from B to C; and once at C, [the system] could get straight back to R. I believe [the system] shows foresight' (Ashby, 1960, p. 69)

A stable system, says Ashby (1960, p. 69), may be regarded both as blindly obeying the laws of its structural dynamic, and also as showing some kind of "intelligent power" (e.g., foresight, anticipation) in reaching its "goals" in spite of "obstacles". Both descriptions are reasonable. The second description is reasonable because it captures an appearance that has a real ground in the system. The first one is reasonable because it captures not the appearance but the very ground of the appearance. Nonetheless, as in the case of the geocentric appearance and the heliocentric organization that grounds it, although both descriptions are reasonable, we have to keep the appearance as what it is, i.e., an appearance, and not to take it as a property of the observed system. The behavioral pattern of a stable system, no matter its degree of complexity, is always a deterministic function of its structural properties, not the expression of an alleged teleological drive.

3.4 Living beings' complexity as stable systems

What are the structural and functional properties that make a stable system a *complex* stable system? What is peculiar about living beings as complex stable systems? There are several properties that are important, but here we will address only a few. One of the main differences between living beings and other kinds of stable system is their multidimensionality. The dimensionality of a stable system expresses the number of variables or

aspects in which the system exhibits stability. A thermostat, for example, operates only in one dimension: temperature. A hyperbaric chamber operates mainly in two dimensions: atmospheric pressure and oxygen concentration. Greenhouses, depending on their technological sophistication, may operate in three or four dimensions: temperature, ventilation, humidity, and luminosity. Living beings are stable systems of high multidimensionality; they operate in many dimensions at the same time (e.g., temperature, blood pressure, hydration, ph, hormone concentration, oxygen concentration, glucose concentration, electrolyte balance, etc.). Whereas the thermostat's behavior has to do solely with temperature stability, the living being's behavior is the complex product of multiple stabilities running at the same time.

More important than the number of stability variables, perhaps, is the degree of connection among them as components of the system. A greenhouse, even managing several dimensions at the same time, may be built so that each subsystem operates separately without affecting the performance of the rest. Temperature and humidity, for example, may go to any value without affecting the dynamics of luminosity and ventilation. The case is quite different when the subsystems are connected, reciprocally conditioning their respective performances. Living beings are stable systems composed of several interconnected stable subsystems; i.e., the imbalance of one of them may bring as a result, sooner or later, directly or indirectly, the imbalance of others. This kind of multidimensionality imposes a complex web of mutual restrictions among the variables, conditioning and complicating the global behavior of the system.

Another important feature is the presence of feedback mechanisms. As we saw in the previous examples of stable systems, cubes and pendulums do not have feedback mechanisms, whereas Watt governors and thermostats do. Living beings' physiology, as it is known from Cannon (1932) onward, is full of feedback mechanisms, yet what matters is not their number but their order or organization. Feedback mechanisms are closed functional circuits that may operate at different levels. When the circuit directly operates upon a determinate variable, we speak of a first-order feedback mechanism. When the circuit operates upon a determinate variable through the mediation of

another feedback mechanism, we speak of a second-order feedback mechanism. A stable system composed by a second-order feedback mechanism is what Ashby called an 'ultrastable system' (1960). Ashby demonstrated the basic principles of ultrastable systems by means of his famous 'homeostat' (an electromagnetic artificial device that exhibited second-order stability). We are not going to review the details of such a demonstration here, but only point out the way ultrastability contributes to the behavioral complexity of the system that has it. Our motto in this chapter, recall, is that the more complex the stable system, the stronger the teleological appearance to the eye of the observer. Ultrastability is one of those features that make living beings complex stable systems.

Very briefly, and taking the examples given by Ashby (1960), we can see the sensorimotor dynamics of an organism as a first-order feedback mechanism. As we shall see in more detail in Chapter 4, sensory and motor surfaces constitute, through the mediation of the environment and the nervous system, a closed functional circuit. Through this sensorimotor circuit organisms may generate, in principle, all kinds of behaviors. But do they exhibit, in equal proportion, all the behaviors allowed by this sensorimotor circuit? No, they do not. Organisms tend to stabilize certain kinds of behaviors and discard others. Why? The reason, explains Ashby, is that the variability of the sensorimotor circuit is limited by another feedback mechanism that, at a second level, operates upon some essential physiological variable (temperature, oxygen concentration, tissue integrity, etc.). In vertebrate animals this second-order mechanism usually runs through some brain structure (generally subcortical nuclei), whose activity connects, at different points, with the sensorimotor mechanisms. In the case of living beings without a nervous system, such as unicellular organisms, the architecture of second-order feedback mechanisms is unclear. Here, for the sake of the exposition, we will assume that unicellular organisms may have, at least in principle, some form of ultrastability.

Taking the example of pain reaction before an external stimulus, and simplifying a bit, every time the second-order circuit finds a level of sensory stimulation that goes beyond a certain physiological threshold, it activates a step-mechanism (i.e., a parametric change) that rearranges the activity of the

sensorimotor circuit (the first-order feedback mechanism), thus generating a variation in the animal's behavior. This process continues until one of the behaviors so generated brings as a result the restoration of the sensory stimulation to its physiological values (e.g., avoiding the source of pain). In other words, the sensorimotor circuit is allowed to generate considerably varied behaviors, under the condition that none of those behaviors displaces the critical variable of the second-order circuits out of its physiological range.

To the extent that living beings are multidimensional systems, one has to assume this ultrastable dynamic at least for every essential physiological variable, and also, as we already saw, a certain degree of connection among them. Ashby called the combination of several ultrastable systems a 'multistable system' (1960), which seems a fair characterization of living beings as stable systems.

The coexistence of several interconnected ultrastable dynamics confers a considerable degree of complexity, yet there is still something missing about the peculiarity of living beings' stability. Ashby demonstrated the phenomenon of ultrastability with his homeostat, and, in principle, one might obtain a multistable system by joining several homeostats. Nonetheless, the activity of a homeostat, which basically consists in changes of potentiometers and electrical current, looks still quite distant from any living being's behavior (in a full working homeostat all you see is four absolutely immobile boxes; the "behavior" of the system reduces to needle movements, buttons turning on and off, and things like that).

To understand the peculiarity of living beings as stable systems we need to address one last and very important aspect; the thermodynamic regime. From a thermodynamic point of view, living beings belong to a special group of physicochemical systems called dissipative structures (Prigogine & Stengers, 1984). Examples of these structures include Benard cells, lasers, flames, stars, tornadoes, and whirlpools (Ji, 2012; Ulanowicz & Hannon, 1987). The peculiarity of these systems, as opposed to the so called equilibrium structures (or near-equilibrium structures), is that they exist and conserve their organization in far-from-thermodynamic equilibrium conditions; i.e., they are nonequilibrium stable steady states. These systems

are thermodynamically open, and maintain integrity through the constant exchange of energy and matter with the environment. In other words, they disintegrate if this exchange is cut off.

Where is the difference then between homeostats and living beings, both of them complex stable systems? The difference, if we go back to Chapter 1, has to do with their respective dynamic regimes. Homeostats, like any electronic device, are dynamic systems of variable regime; they can alternate, without loss of organization, between passive and active dynamics. Living beings are active dynamic systems of fixed regime; they cannot change to a passive regime without loss of organization. Let us look at this point more closely.

The homeostat is a system fed with free energy supply (electrical power). If you cut the supply the system ceases to operate, but does not disintegrate. Later on you can turn it on again by restoring the energy supply. That is because the homeostat exists and maintains its organization in near thermodynamic equilibrium conditions; i.e., its integrity does not depend on a continuous material and energetic exchange with the environment. Living beings, instead, are nonequilibrium structures, and if they go to thermodynamic equilibrium they not only cease to operate, they die.

> What is most crucial to the case of living systems is that they are far from equilibrium not only with respect to their "operation", but with respect to their existence. The thermodynamics involved in the ontology of the kinds of system that they exemplify is an irreversible thermodynamics — they cannot be "restored", because of that irreversibility. They die if they go to equilibrium, and that is final. (Bickhard, 2007, p. 582)

Living beings are a particular version of dissipative structures; their peculiarity, as we shall see in the next chapter, lies in the kind of endogenous dynamic that constitutes them as physicochemical systems (i.e., the autopoietic dynamic). The important point here is that living beings, like any other dissipative structure, are systems whose region of physicochemical

stability is far-from-thermodynamic equilibrium. This means that, when disturbed, they move not to equilibrium but to the specific far-from-equilibrium region in which they conserve integrity, which is a remarkable feature. As an image, it is as if a volume of water were suspended in the middle of a slope, and after being displaced downward by an external force, it moved upward returning to its initial location. The stability exhibited by dissipative structures is a highly complex form of stability, in fact, a form that for a long time was thought of as violating natural laws, specifically the second law of thermodynamics. (That that is not the case has been demonstrated by many people, but the works of Prigogine are surely the best known. See Prigogine, 1980; Nicolis & Prigogine, 1977; Prigogine & Stengers, 1984).

Every dissipative structure, at different scales, exhibits the same behavioral pattern of stability. Disturb a candle flame in different ways (without being destructive, of course), and you will see how the flame reconstitutes as such. Disturb a maelstrom in the sea, and you will see how the maelstrom returns and conserves its integrity. Once a nonequilibrium steady state stabilizes as such, it is able to exhibit a considerable degree of stability in spite of disturbances. Sure, the stability that a system X can reach in far-from-equilibrium conditions is more precarious than the stability that it might reach in equilibrium conditions (sooner or later, stars disintegrate and living beings die), yet it is still a quite strong stability.

As in any case of stability, dissipative structures seem to "insist", despite disturbances, in retaining their organization, and so are susceptible to teleological descriptions. In the case of the living being, the typical image is that of a system "struggling" to survive and maintain its integrity; e.g., the classical Spinozan "conatus", or what enactivists would call 'motor intentionality' (Gallagher and Miyahara, 2012). Being dissipative structures, and therefore thermodynamically open systems, living beings are stabilized as energy and matter flow systems. Like the pendulum that always returns to its equilibrium steady state, so the living being always returns to its non-equilibrium steady state; i.e., it constantly restores the exchange of energy and matter with the environment.

For example, if in an animal the intake of energy and matter is

diminished to some degree (e.g., it has not eaten or drunk anything for a relatively long period), and the system, because of this, is displaced from its steady state as a thermodynamic system (i.e., essential variables such as hydration or glucose concentration deviate from their optimal values), at some moment the system will automatically operate, through one or more ultrastable mechanisms, the reactivation of its sensorimotor circuit. From that moment, the ultrastable dynamics held between the second-order feedback mechanism (which operates upon essential variables) and the first-order feedback mechanism (which operates upon the correlations of sensorimotor activity), will iterate until the behavior of the system brings as a result an intake of matter and energy such that the essential variables return to their optimal values. In this complex process, what the observer will see from the outside is that the cat gets up, walks to the kitchen and drinks some milk, the frog targets the fly, shoots its tongue and eats it, the lab rat, trapped in a maze, searches a way out, finds it and eats the food; all this with the alleged "purpose" of recovering their respective physiological levels of energy.

Living beings face a continuous flow of disturbances, both internal and external, and their behavior as dissipative structures is a constant return to the far-from-equilibrium condition where they exist. That is why we see them constantly renewing the exchange of energy and matter with the environment. Are there here purposes, goals, telos? Not really, though, as with any stable system, their behavior may be reasonably described, if the observer wants to, in teleological terms. The temptation is especially strong here, since living beings, among dissipative structures, are the only ones endowed with ultrastable mechanisms. The teleological description of living beings captures an appearance that is valid and justified, in the sense that it is grounded in stability mechanisms that are real and operative in living beings. Nonetheless, being an appearance, we cannot use it as an explanatory element in our cognitive theories; not, at least, if what we want is a Strictly Naturalistic cognitive science.

Chapter 4

Organizational closure: autopoiesis and senso-effector systems

So far we have characterized living beings as adaptive dynamic systems (Chapter 1), deterministic machines of closed transitions (Chapter 2), and multistable dissipative systems (Chapter 3). In this chapter we will analyze one last and very interesting property of living beings; their organizational closure. More specifically, we will examine living beings' organizational closure with respect to their autopoietic constitution and sensorimotor dynamic.

As we shall see, organizational closure is a relatively common property in certain natural and artificial systems, not an exclusive mark of living beings. Nonetheless, there is one specific domain in which, apparently, *only* living beings exhibit organizational closure; that of processes of molecular synthesis. This is the central claim of Maturana's autopoietic theory of living beings.

Maturana's autopoietic theory comprises two related but logically independent ideas: 1) all living beings discovered in Nature hitherto are, or are composed by, physical autopoietic systems, and 2) any physical autopoietic system, irrespective of its concrete molecular realization (carbon based or not) and origin (natural or artificial), can be considered a living system. I will argue that, out of these claims, number 1 is almost trivially correct, whereas number 2 might or should be left open to discussion. Claim number 1 is simply a descriptive abstraction, and, in my view, a correct one. As we shall see, the notion of autopoiesis is nothing more than the formal

expression of the well-known cyclic character of cell metabolism, and does not offer by itself much space for the debate. Claim number 2, instead, is stipulative. It says that the presence of autopoiesis in any physical system, discovered in Nature or artificially created, and whatever its molecular base, should be enough for us to qualify it as a living being.

Although I think that this stipulation is basically correct, I would like, at least for the purposes of this thesis, to leave it open to discussion. Some people think that the mere presence of autopoiesis in a physical system does not make that system, immediately and without ambiguity, a living being. Some think that the presence of autopoiesis is a necessary but not sufficient condition for a given physical unity to be considered as a living being; that something more is needed (Bitbol and Luisi, 2004; Bourgine and Stewart, 2004). The objection, roughly speaking, is that as a demarcating criterion, autopoiesis is too abstract and general; that we can conceive of, or artificially create, systems whose dynamic is basically autopoietic but whose structural complexity, behavior or mode of interaction with their environment is too poor, fragile or simple to be considered genuinely 'alive'.

I take it that this discussion is, to a large extent, a matter of convention. Some people, like Maturana and myself, would not have any problem in calling an autopoietic physical system made of entirely 'atypical' kinds of molecules, without any resemblance to DNA or protein molecules a 'living being'. Others, instead, think that a chemical system without DNA and its associated ribosomal system, even if autopoietic, could not be considered a genuine living being (Barbieri, 2012, 2008). Some people, like Maturana and myself, would not have any problem in calling an extremely fragile and simple autopoietic physical system, one that cannot survive minimal perturbations or easily disintegrates, a 'living being'. Others, instead, may think that a system that cannot survive minimal perturbations cannot be a living being (Damiano and Luisi, 2010). This discussion I want to leave open because, ultimately, it is not relevant for the essential purposes of this thesis.

Strictly speaking, the general theory of cognition that I am defending here does not depend on the particular fortune of autopoietic theory. Maturana, and I as a follower, might be entirely wrong in thinking that autopoiesis is an

exclusive mark of living beings, or that its presence in a given system is sufficient for that system to be considered alive. Yet that would not change anything regarding our Strict Naturalistic conception of living beings, which is the point that really matters to us. The crux, for the specific purposes of a Strict Naturalistic cognitive science, is not whether autopoiesis is or is not a distinctive mark of living beings but rather whether living beings, whatever their organization, are or are not metaphysically ordinary dynamic systems subjected to the same constraints that rule any physical system in general, and whether or not they must be conceived of in Strictly Naturalistic terms. In this sense, the main message of autopoietic theory, as we shall see, is that the only peculiarity of living beings within the physical universe is their autopoietic constitution, and that there is nothing metaphysically extraordinary about that!

Now, although the theory of cognition presented in this thesis is independent of the autopoietic theory, the notion of autopoiesis is important in its own right because it helps us to understand, in part, the origin of the behavioral distinctiveness of living beings as dynamic systems. We have said in previous chapters that living beings' adaptation is simply one more version of a universal phenomenon among interacting dynamic systems, namely structural coupling; that all physical systems, as long as they exist in a given environment, exhibit some form of adaptation. Adaptation, recall, is conservation of organization (class identity) through, or in spite of, structural changes in interaction with a medium. What defines, to a large extent, the different forms of adaptation found in different systems is precisely the kind of organization (class identity) conserved in each case. We have seen in the previous chapter that living beings belong to the natural kind of dissipative structures, and that within that group they are the only ones endowed with ultrastability mechanisms. This has helped us to understand a good part of living beings' behavioral distinctiveness, but we still have not specified the particular organization around which living beings maintain adaptation. The notion of autopoiesis will help us to fill this gap. Autopoietic theory will help us to understand a bit more why, despite their ordinary metaphysical constitution, living beings behave in ways that are peculiar within physical systems. Consequently, we will dedicate a part of this chapter to the

autopoietic theory of living beings.

In the context of our cognitive theory, the most important aspect of living beings' organizational closure is not their autopoietic constitution but their senso-effector dynamic. All living beings, in normal conditions, have one or more senso-effector systems incorporated in their structure, some of them internal (e.g., the endocrine system), others in functional contact with the external environment (e.g., the sensorimotor division of the nervous system). The structural implementation of these systems varies according to the complexity of the organisms. In unicellular organisms, for example, senso-effector systems are usually composed by a set of macromolecular structures (e.g., membrane chemoreceptors, microtubules, etc.) and organelles (e.g., flagella, cilia, etc.). In multicellular organisms, with the only exception of sponges and a few other similar species, the senso-effector systems are invariably composed of sets of specialized cells (e.g., neurons in the nervous system, secretory cells in endocrine system). The most common senso-effector systems in multicellular organisms are the nervous system, the endocrine and exocrine system, and the immune system. Out of these systems, here we will concentrate on the nervous system (most of the times only on its sensorimotor division).

The nervous system interests us for obvious reasons, yet what interests us are not its anatomical details but its functional organization, the 'logic' of its operational dynamic as a network of senso-effector correlations. In this chapter Maturana will tell us that the nervous system, from the point of view of its functional organization, is a closed system, and that as such it does not have inputs and outputs, inside and outside. We will try to understand this idea, evaluate its philosophical cogency, and explore some (only some) of its consequences. Our guiding questions will be; what does it mean to say that the nervous system is an organizationally closed system? Is this a proper characterization of its senso-effector dynamic? If so, what implications does this have for our understanding of living beings' behavior? What can 'perception' be for a closed system (like the nervous system)? What can 'action' be for a closed system (like the nervous system)?

4.1 Organizational closure

In Chapter 1 we said that in some active dynamic systems we may distinguish not only changes of state but processes, operations or mechanisms that exhibit a certain organization. We called these systems, generically (and lacking a better name), 'systems of processes'.

Systems of processes can be characterized in terms of the kind of events, operations, processes or mechanisms that constitute them, and/or in terms of the way in which these events, operations, processes or mechanisms are organized. For example, to say that X is a system of power generation is to say that the kind of processes that compose X are processes of energy generation, or that, taken as a whole, these processes bring as a result the generation of energy. In saying that, let us notice, we are not saying anything about the organization of these processes. Perhaps X's processes run in parallel, perhaps serially. Perhaps they are arranged as a distributed network, perhaps as a centralized flow.

The events, operations, processes or mechanisms that constitute a determinate system (concrete or abstract, natural or artificial) can be organized in different ways. There is, nonetheless, one kind of organization that is especially relevant to us, namely the circular (cyclic) or closed organization. An organization of processes is circular or closed if the result of its processes (re)enters and participates as a constituent element of the organization of processes itself. This organization contrasts with the open or linear organization, in which the result of the processes does not enter and participate as a constituent of the organization itself.

Systems that exhibit closed or circular organizations are, for example, natural cyclic systems such as the hydrologic system (the so called 'water cycle') or the geological system of rock formation (known as the 'rock cycle'). Closer to our topic, all feedback systems constitute examples of organizationally closed systems too. In the previous chapter we saw several of them, but at that moment, focused on the topic of stability, we did not emphasize their circular organization. Thermostats, Watt governors, automatic pilots, but also homeostatic mechanisms in living beings, are all

instances of closed functional circuits.

In this chapter we will concentrate on two aspects in which living beings exhibit organizational closure: metabolism and sensorimotor activity.

4.2 Autopoiesis and living beings

The word ‘autopoiesis’—literally ‘self-production’—denotes a system of productive processes organized in a circular way.

Maturana’s autopoietic theory is focused on the basic (or minimal) living unity, the cell. The individual cell, either as a unicellular organism (bacteria, amoebas, paramecia, etc.) or as a composing cell of a multicellular organism, is, according to Maturana, an autopoietic molecular machine. An autopoietic machine is a subclass of poietic machines. A poietic machine is a machine that produces or fabricates something. Two basic subtypes of productive machines can be distinguished: allopoietic and autopoietic machines. Allopoietic machines are systems that produce something distinct from themselves (e.g., a car factory), while autopoietic machines are machines that produce themselves (Maturana, 1975).

Living cells are also allopoietic machines; they constantly produce elements that do not form part of their own productive system (e.g., hormones, neurotransmitters, etc.). Maturana contends that living cells are distinctively autopoietic and trivially allopoietic machines. That is, that among allopoietic physical systems, only living cells are at the same time autopoietic.

The word ‘autopoiesis’, without further specifications, denotes a particular kind of processes (production processes) and a particular kind of organization (circular organization), not a particular kind of physical entity. It is worth reminding the reader here that ‘organization’ is a formal notion, an abstraction. It refers to the set of relations that define the class identity of a system, not to the concrete conditions under which such relations are satisfied or conserved in a particular domain of existence. Thus, the “notion of autopoiesis [...] says nothing about the nature of the components that

realize the system as a network of productions" (Maturana, 1981, p. 22). When we say that "Y" is an autopoietic system what we mean is that "Y" is *organized* as a self-producing network. A self-producing network of what? Well, that is not relevant for the identification of "Y" as an autopoietic system. "Y" may be a formal (ideal) system of purely abstract elements, a bidimensional model in a virtual space, or a system made of physical components in a concrete tridimensional space. This latter case is the case of living systems: "a living system is an autopoietic system in *physical* space" (Maturana, 1981, p. 22. Emphasis added). In other words, while all living cells are autopoietic systems, not all autopoietic systems are necessarily living cells.

The formal status of the notion of autopoiesis is usually overlooked among commentators, as most of them tend to include in the notion of autopoiesis the material constitution of living beings, as if 'autopoietic systems' and 'living systems' were coextensive categories. For example, Evan Thompson claims that an autopoietic system is a "self-producing bounded molecular system" (2007, p. 44), while Michael Wheeler thinks that "autopoiesis is autonomy plus materiality [, and that] materiality is definitional of autopoiesis" (2011, p. 151). This is not entirely correct. Maturana explicitly points out that "[t]here is no restriction on the space in which an autopoietic system may exist [, and that the] physical space in which living systems exist is only one of many" (Maturana, 1981, pp. 22-23). The materiality alluded to by Thompson and Wheeler may well be an essential element of living beings as particular instantiations of autopoietic systems (as the quotation below clearly shows), but not of autopoietic systems as a general class.

> A living system is a discrete self-contained molecular dynamic system that produces itself as a closed network of productions of molecules that in their interactions produce the same network of molecular productions that produce them as a stationary dynamics sustained in a *continuous flow of matter and energy through it* (Maturana, 2011, p. 145)

I have highlighted the final words of this paragraph simply to emphasize something that should be obvious, but that sometimes seems to be overlooked. In the physical space of molecular dynamics, any productive process entails the consumption of energy and the utilization of some 'raw' material.

It is also worth noting that in the concept of autopoiesis the suffix 'poiesis' captures *a particular sense* of the original Greek *poiēsis*. *Poiēsis* is a broad Greek concept that means, in its most embracing sense, 'to bring forth', 'to bring into presence' or 'to bring into appearance', and that may be equally applied to artistic creation, handicraft manufacture, or the spontaneous (non-manmade) arising of natural formations. Many senses are involved in this primitive concept of *poiēsis*: 'to create', 'to conceive', 'to generate or beget', 'to produce', 'to make up', 'to fabricate', 'to bloom', 'to sprout or burst forth', etc. Out of these multiple meanings, Maturana alludes only to a process of production or fabrication; more specifically, to a process of 'synthesis' or 'composition' whereby a set of elements are assembled (combined under certain organization) to form a complex whole. Maturana wants to capture, in formal terms, the permanent dynamic of *molecular synthesis* (formation of molecular compounds, generally organic polymers, by means of one or more chemical reactions) that takes place in the cell metabolism. In this sense, if a car factory is an allopoietic machine, *a cell is an autopoietic machine as long as it is a molecular factory that synthesizes the molecules that constitute it as such*.

This notion of 'production' as 'synthesis' must be differentiated from the notion of production used in certain philosophical theories of causation, where it is said that an event *E* has as a cause the event *C* if *E* is the effect *produced* by *C* (Hall, 2004). In this case 'to produce' means simply to bring about or generate a certain state of affairs. For instance, we may say that the increase in temperature produces (causes) the melting of snow, that the friction of bodies produces heat, or that earthquakes produce structural damage in bridges. None of these causal relations, however, involves the assembling of parts or elements to build a complex whole, which is the sense in which 'poiesis' means 'production' (synthesis, composition) in autopoietic theory (see also below the example of the burning candle).

To be a poietic machine in general (allo or autopoietic) a system X has to produce or fabricate something. If X merely repairs, restores or renovates components without synthesizing anything, then X is not a poietic machine. There are several physical systems that keep constant their organization through a permanent renovation of their material components. We saw some of them in the previous chapter when talking about dissipative structures. For example, a turbulence in the current of a river or a tornado are natural systems that remain constant in their configuration through (or in spite of) the renovation of their material components. A cellular system is a system that renovates its material components too, of course, (it needs nutrients and it evacuates chemical waste), and that consumes and dissipates energy (it is a dissipative system), *but that is not what defines its class identity as a living system*. The difference is that the cell is organized as a productive network, not just as a system through which some components come and go. A tornado or a turbulence, to be an autopoietic system, should be constituted as a network of productive processes, as a factory; they should assemble elements and build the very compounds that constitute them as systems.

There are also self-sustaining dissipative structures, where what one sees is a chain of physicochemical events that maintains itself. The burning candle is a typical example: the heat of the flame melts and liquefies the wax (combustible solid), this liquefied combustible ascends through the wick (by capillarity) where it is vaporized to finally burn in the flame, whose heat melts and liquefies the wax..., and so on and so forth. At this point the reader, noting the evident circularity of the process, could describe the sequence in causal terms by saying "the heat of the flame causes or produces the melting of the wax, the capillary action produces the ascent of the combustible liquid [...] the combustion produces the flame, whose heat causes or produces the melting...", and then ask "Is this not a productive circular network, and therefore a self-producing or autopoietic system?"

The confusion vanishes if we keep in mind the distinction between the 'poietic' notion of production and the 'causal' notion of production mentioned before. While it is true that the burning candle constitutes a causal circle, it is not the case that this circle is an assembling network that synthesizes the material compounds that constitute the burning candle as

such. The combustion process consists in a series of exothermic chemical reactions that release some products, but these products (combustion gases) do not participate in the synthesis of the constituent elements of the chain of reactions. For example, the released CO_2 does not enter in any chemical reaction to produce wax (combustible).

Living beings are self-sustaining dissipative structures too, but, again, that is not what confers their class identity. What is peculiar to living beings among dissipative structures is their internal dynamic of molecular self-production.

In the same way, we must distinguish the specific notion of 'production' as 'synthesis' from the more general idea of *poiēsis* as 'bringing into presence'. Heidegger, for example, in the context of his theory of truth as *alethēia* (truth as 'unhiddenness'), speaks of *poiēsis* in the broad sense of 'passing from a state of concealment to a state of unconcealment'. Quoting the Plato of *Symposium*, Heidegger says that "[e]very occasion for whatever passes over and goes forward into presencing from that which is not presencing is *poiēsis*, is bringing-forth" (1977, p. 10. Original emphasis), and refers to the organic Nature (*physis*) as the primary source of *poiēsis*. Why? The reason is that, unlike manmade artifacts, natural phenomena would have their poietic principle in themselves:

> *Physis* is indeed *poiēsis* in the highest sense. For what presences by means of *physis* has the bursting open belonging to bringing-forth, e.g., the bursting of a blossom into bloom, in itself (*en heautōi*). In contrast, what is brought forth by the artisan or the artist, e.g., the silver chalice, has the bursting open belonging to bringing-forth not in itself, but in another (*en allōi*), in the craftsman or artist. (Heidegger, 1977, pp. 10-11. Original emphasis)

Some commentators have believed that they have found here a direct antecedent of Maturana's autopoietic theory (Dicks, 2011; Di Paolo, 2009; Ilharco, 2003; Mingers, 1995). Heidegger characterizes 'organic nature' (i.e., biosphere or biology, broadly construed) in terms of 'poiesis in itself'

(*poiēsis en heautōi*), and the artificial world of human creations as 'poiesis in something else' (*poiēsis en allōi*), which seems to resemble the Maturanian distinction between 'auto-poiesis' (in biological systems) and 'allo-poiesis' (in manmade systems). The similarity, nonetheless, is only superficial. Heidegger's main concern is to recover, through the notion of *poiēsis* as 'bringing-forth', the primitive (Pre-Socratic) ontological sense of 'truth' as *alethēia* or 'unhiddeness', i.e., 'truth' as a property of Being rather than as a property of our beliefs or linguistic expressions. His philosophical project has not to do with defining the class identity of living beings but with reviving a forgotten and, to his eyes, more fundamental sense of 'truth':

> Through bringing-forth, the growing things of nature as well as whatever is completed through the crafts and the arts come at any given time to their appearance […] Bringing-forth brings hither out of concealment forth into unconcealment. Bringing-forth comes to pass only insofar as something concealed comes into unconcealment. This coming rests and moves freely within what we call revealing […] The Greeks have the word *alethēia* for revealing. The Romans translate this with *veritas* [, and w]e say "truth" […]" (Heidegger, 1977, p. 11. Original emphasis)

This splendid and deep Heideggerian reflection about the notion of 'truth' as an ontological category does not correspond to the modest and metabolically inspired sense in which Maturana talks about 'autopoiesis', and therefore it does not seem particularly useful to explain or clarify its meaning.[8]

Another apparent connection, sometimes mentioned in the secondary literature, is Canguilhem's philosophy of biology. For example, Di Paolo points out that Canguilhem, before Maturana, had "already in 1951 used the

[8] This is not to say that both notions are incompatible. One can build, actually, a Heideggerian interpretation of Maturana's theory (or vice versa) in a relatively easy way (see Mingers, 1995). A more harmonic connection would be between Heidegger and the enactive approach, which explicitly refers to the notion of 'bringing forth' as a way of characterizing cognition.

term *autopoetique* to define the character of living organisms" (2009, pp. 43-44. Original emphasis). This association, tempting at first glance, is risky in more than one sense. First, it suggests that Canguilhem and Maturana, in using similar *terms*, are using similar *concepts*, which as we shall see is not the case, and second, it overlooks the notable disparity (and perhaps incompatibility) that exists between the normative approach of Canguilhem and the strictly non-normative stance of Maturana. In Canguilhem the concept *autopoétique* does not refer to the self-productive dynamics of living beings but, to (what he takes to be) the biological normativity of their internal physiology (their *inner milieu*). Canguilhem opposes the internal autonomy of living beings to the *hétéropoétique* character of manmade instruments and artifacts. The idea, roughly speaking, is that while human artifacts respond to the external demands of their users, living beings impose their own physiological demands in an autonomous way; that while human artifacts adjust to an external functional environment, living beings adapt, also and primarily, to their own internal functional environment. Canguilhem's distinction between *autopoétique* and *hétéropoétique* does not take place in the context of the question about the defining organization of living beings but in the context of a methodological reflection about the particularities of the 'experimental method' in biology (see Canguilhem, 1965, Part 1 "La méthode biologique"). The notion of 'poiesis' as production—that in French corresponds to *poïèse* or *poïètique* (rather than to *poétique*)—is completely absent in Canguilhem's analysis, whose main concern is to highlight the alleged self-normative (auto-nomic) character of living beings. And nothing could be more distant from Maturana's autopoietic machines than the idea of an alleged 'intrinsic normativity' in living beings. Maturana is emphatic in saying that "what [we] call normative activities *are not aspects of* [...] *autopoiesis* [but only] commentaries or explanatory propositions that [we] make about what [we] may think that should occur in the […] organism" (2011, pp. 149-150. Emphasis added).

In this section we have tried to clarify the notion of autopoiesis and identify the precise sense in which it applies to living beings. We have also tried to warn of some conceptual associations or similarities that may be misleading. But the most important message, in fact the message that

Maturana wants to transmit with his theory, is that once revealed as autopoietic systems, i.e., as cyclic or circular metabolic systems of molecular synthesis, living beings are revealed as metaphysically trivial physical machines whose only peculiarity, as dissipative thermodynamic structures, lies in their self-productive dynamic. What the autopoietic theory tells us, ultimately, is that among natural systems, living beings simply constitute a *version*, not an *exception*.

Autopoietic theory is, indeed, a way of justifying the Strict Naturalistic view of living beings. If the cell, the basic unity of living beings, is ultimately a set of chemical reactions *organized* in a certain way and sustained under certain thermodynamic conditions, then there seems to be no reason to treat living beings as metaphysically exceptional entities. If in standard natural sciences rivers, volcanoes, stars, comets, or any other kind of natural systems, are not conceived of or treated as semantic, intentional, teleological, agential, epistemic or normative systems, then living beings should not be so conceived of or treated either.

Let us examine now the organizational closure of living beings with respect to their sensorimotor dynamic.

4.3 Organizational closure in senso-effector systems

All structural interactions in the physical world involve some transfer or exchange of energy. In some cases, this transfer is purely quantitative, i.e., without energy conversion (e.g., thermal conduction between two or more systems, elastic collisions). In other cases, the transferred energy is not only received but also transformed into another type of energy. The structures that receive an amount of energy and convert it into another type of energy are usually called ‘transducers’.

A transducer, in a broad sense, is any structure that reacts to certain structural configurations of its environment, converting the type of energy associated with the said configurations into the type of energy associated

with its *own* structural configuration. Transduction phenomena are, thus, a subtype of *structural* interaction, subject, therefore, to the same metaphysical principles valid for any structural interaction. Their only difference with the rest of structural interactions is that in them the transferred energy is converted into another kind of energy; conversion that is realized, invariably, according to the *structural determination* of the recipient system (this metaphysical point will be relevant for our discussion on perception later on).

Senso-effector systems are, essentially, functional systems composed by two or more transducers whose dynamics are connected or are functionally dependent in one way or another. Being functional systems, senso-effector systems can be found in a great variety of structural realizations, biological or not. A living being's sensorimotor system is an instance of a senso-effector system. For example, pressure sensors in the skin receive mechanical energy and convert it into electrical impulses, thus acting as transducers. Photoreceptor cells in the retina (cones and rods) receive electromagnetic radiation and convert the energy associated with photons into electrical impulses, thus acting as transducers. On the other side, muscle fibers receive electrical neural impulses and convert them, with the aid of their own energetic resources, into mechanical energy (movement), thus acting as transducers too. Artificial examples of senso-effector systems can be found in infrared sensors, photovoltaic cells in power generation plants, or bimetallic strips in thermostats.

From a structural point of view, senso-effector systems are almost always incorporated as subsystems in some larger system that contains them (an organism, an airplane, a power generation plant, a thermostat, etc.), and with respect to which, generally, the observer distinguishes or assigns the roles of sensor and effector between otherwise functionally equivalent transducer elements. For example, previously we saw that both photoreceptor cells and muscle fibers operate as transducers. If they are linked through the nervous system, thus forming a system, which one of them is the sensor and which one the effector? Well, none, or both; it depends on the chosen point of reference. Let us see.

In the case of photoreceptor cells, and considered only from the point of

view of transduction, their environment or domain of perturbations has to do essentially with photons. In the case of muscle fibers, their environment or domain of perturbations has to do essentially with electrical impulses. Taking as a point of reference the organism in which these transducer structures are incorporated, what one sees is that photons come from the outside (ambient light) and the electrical impulses received by muscle fibers from the inside (nervous system). According to this perspective, one says that the organism senses the environment through its photoreceptor cells (sensor element), and that acts or responds through its muscle movements (effector element). However, since both photoreceptor cells and muscle fibers are transducers that have their own environment or domain of perturbations, we might validly change our perspective and see the muscle fibers as sensing the nervous system (sensor element) and the photoreceptor cells as acting upon the nervous system (effector element). We do not do this because our descriptive and explanatory purposes with respect to living beings usually have a specific orientation, i.e., we want to understand living beings' adaptation with respect to what we see as their external environment. But it is worth noting, when we talk about senso-effector systems, the conventional status of the categories 'sensor' and 'effector'. (The point is not as trivial as it seems, especially when we take into account, as we shall see soon, the closed nature of the sensorimotor system in living beings).

A senso-effector system may be organized in different ways. A thermostat, for example, may be assembled so that the sensor component (e.g., the bimetallic strip), through intermediate mechanisms, affects the effector component (the heater), but the effector, located in a distant and separate room (e.g., in a different house), does not affect the sensor component. That would be a case of an open or linear senso-effector system. In normal conditions, thermostats are always assembled as closed circuits, allowing the heater to affect, through the environment, the sensor component. That is because in designing and using thermostats what we want is precisely the correlated and coordinated activity of their sensor and effector components (in this case, the negative feedback loop).

The sensorimotor system of living beings is a senso-effector system that, in normal conditions, is organized as a closed circuit too; i.e., as a feedback

mechanism. What happens at the level of the sensory surface affects, through some mediating biological mechanism (that may or may not include a nervous system), the activity of the effector surface, and the activity of the effector surface affects, through the mediation of environmental factors, what happens at the level of the sensory surface. Ashby illustrates the idea with the example of a kitten approaching fire:

> The various stimuli from the fire, working through the nervous system, evoke some reaction from the kitten's muscles; equally the kitten's movements, by altering the position of its body in relation to the fire, will cause changes to occur in the pattern of stimuli which falls on the kitten's sense-organs. The receptors therefore affect the muscles (by effects transmitted through the nervous system), and the muscles affect the receptors (by effects transmitted through the environment). The action is two-way and the system possesses feedback. (Ashby, 1960, p. 38)

Maturana, with amoebas and protozoa, provides a more basic example (a sensorimotor dynamic without a nervous system):

> The presence of the protozoan generates a concentration of substances in the environment. These substances are capable of interacting with the amoeba membrane, triggering changes in the consistency of the protoplasm which results in the formation of a pseudopod. The pseudopod, in turn, causes changes in the position of the moving animal, thus modifying the number of molecules in the environment which interact with its membrane. This cycle is repeated and the sequence of movements of the amoeba is therefore produced through the maintenance of an *internal correlation* between the degree of change of its membrane and those protoplasmic changes we see as pseudopods. (Maturana and Varela, 1987, pp. 147-148).

In both cases, what we have is a closed functional circuit wherein the sensor and effector components maintain a set of dynamics correlations between them. Following these observations, Maturana draws the idea that from the point of view of its sensorimotor dynamic, the "living being is [a] system *closed on itself*" (Maturana, 1970, reprinted in Maturana and Varela, 1980, p. 50. Emphasis added). What does this mean?

4.3.1 Organizational closure and sensorimotor system

Recall in Chapter 1 we said that, when observing living beings, we tend to see them as having an external world before them, a world toward which their existence is open and oriented. Living beings appear to us as endowed with sensory 'windows' that pick up stimuli from the external world (i.e., as capable of perception), and effector organs that act upon said external world (i.e., as capable of action). Could we be mistaken about that? According to Maturana's theory, we not only could be but actually are mistaken about that (Maturana, 1970/1980, 1975, 2003). As in the case of the geocentric appearance, which is grounded in the heliocentric organization of the solar system and our relative position as observers, the sensorimotor openness and directedness of living beings would be a normal and expectable appearance that has to do, on the one hand, with the circular organization of the observed system, and on the other, with our particular position as observers in relationship to the said system.

To properly interpret Maturana's idea, it is important to understand the *functional* notion of closure at play here. When we say that living beings' sensorimotor dynamic is a closed system, someone might ask: "Closed to what?" This question presupposes a physical-spatial interpretation of closure, as when we close a door and leave someone outside. In this case, it sounds as if the sensorimotor system, in its closure, were blocking the entrance to something external (e.g., the environment, the external world). Yet this is not the sense in which Maturana talks about sensorimotor closure.

> The sensory and the effector surfaces that an observer can describe in an actual organism, do not make the […] system an open […] network because the environment (in which the observer stands) acts only as an intervening element *through which* the effector and sensory [surfaces] interact completing the closure of the system (Maturana, 1975, p. 318. Emphasis added)

The functional organization of the sensorimotor system is closed, but this is not because the environment is left outside the system, but rather because the environment is always incorporated as a functional step *within* the system. The sensorimotor system is not closed *to* the environment; it closes on itself *through* the environment. What Maturana means is that the section (functional gap) that we, from our position as external observers identify as the environment or the outside of the system is, for the system, one more of its functional links.

If this is the case, Maturana goes on, if the environment is always included as a functional component of its sensorimotor dynamic, then the living being cannot have the environment as something external to it. That is, if in the closed dynamic of the sensorimotor system the environment works as one more functional link, then what we see as 'the environment' or 'the external world' of the system must be, for the system, and from the functional point of view, rather a 'transparent' element; something too intimate and inner, so to speak, to be distinguished as a separate object.

The idea is that, considered as senso-effector systems, living beings form with the environment a functional unity, a continuum in which inputs and outputs, *understood as intrinsic properties of the system*, do not exist. It is the observer who, for her own descriptive or explanatory purposes, may 'open' the system and consider the environmental stimuli as 'inputs', or the motor activities as 'outputs'. Nonetheless, argues Maturana, this distinction, as far as the dynamic/functional organization of the system is concerned, is arbitrary, and does not reveal any intrinsic property of the system.

In saying this Maturana is not denying that, from the structural-

topological point of view, living beings have physical boundaries that separate them from the environment. Living beings are physically discrete unities with more or less clear boundaries (membrane, skin, exoskeleton, etc.). Maturana recognizes this point when he affirms that every living being is a "*discrete* [...] molecular dynamic system" (2011, p. 145. Emphasis added). The idea is rather that said physical boundary, though structurally real, does not have operational presence (i.e., it is functionally irrelevant) in the domain of the sensorimotor system as a closed circuit.

Let me insist, to prevent misunderstandings, and before going deeper with the analysis, on the strictly functional and sensorimotor dimension of this notion of closure. We have seen that living beings, from the material and thermodynamic point of view, are open systems that conserve integrity through a constant exchange of matter and energy with the environment. Maturana, as any standard biologist, acknowledges this point when he says that a living being is a "system […] sustained in a continuous flow of matter and energy through it" (Maturana, 2011, p. 145). Once again, the closed character of living beings pointed out by Maturana refers to their *functional organization* as sensorimotor systems, not to their material existence as dissipative structures.

Having said that, if we acknowledge, as Maturana and I want to, the circularity and closure of the functional organization of the sensorimotor system (in the relevant sense), then any directedness or openness will appear as an observer-relative ascription, not as an intrinsic property of the system. This is because the sensorimotor system, as a closed network, is always interacting with itself, no matter, from the functional point of view, whether this interaction takes place through a functional node that is inside or outside the organism as a discrete physical unity. An observer may, out of the many functional nodes of the system, pick out one in particular as a point of reference and describe the system as oriented or open to it. Any node might, in principle, be chosen. However, since we as observers are always located in one of the functional nodes of the network, namely the gap between the sensory and motor surfaces of the organism, we, tacitly (as a sort of 'natural attitude'), always take that node as a point of reference. We see the organism's functioning as directed precisely at the node in which we operate

as observers; i.e., to the environment. This directedness, however, is entirely observer-relative. Were we placed in a different functional node, we might validly treat that node as *the* 'environment' toward which the system is open and directed.

Let us take, to visualize the point, the example of the nervous system as a subcomponent of the sensorimotor system. The nervous system connects the sensor and effector surfaces through a complex network of neurons that interact in electrochemical terms. In each one of these interactions, called synapses, what we see, basically, is a presynaptic surface that affects, through chemical mediators, a postsynaptic surface. This interaction, from the strictly *functional* point of view, has no difference with the interaction that we see between the effector and the sensor surfaces of the organism. Presynaptic and postsynaptic surfaces act as minitransducers that convert electrical impulses into chemical energy (release of neurotransmitters at the presynaptic surface), and complementarily, chemical energy into electrical impulses (membrane depolarization at the postsynaptic surface). The only difference, one that is not relevant for the closed dynamic of the system, is that in one case the mediators are intra organismic chemical compounds, while in the other they are environmental factors.

As a metaphor, *and only as a metaphor*, it is as if the sensorimotor system was always 'talking' to itself, without noticing whether the transmitter vehicles are bodily chemicals or environmental factors. All there is for the sensorimotor system, including the nervous system, is a circular dynamic of correlations; i.e., a constant 'monologue' (more about this soon).

Maturana's relativistic point is the claim that if we as observers were placed within the nervous system, we might validly treat one or more of its synaptic gaps as the 'environment' toward which the system is open, directed or adapted.

The situation, if you will, is similar to the description of a moving object as approaching or moving away. What is the "right" description? What is the description that reveals the intrinsic dynamic of the object? Well, neither of them—neither reveals an intrinsic (non-relative) property in the moving object. The description of the object as approaching, or as moving away, is entirely dependent on the point of reference adopted by the observer. The

approaching of the object (or its moving away) is a property that we fix by adopting a descriptive convention, not something that we "find" in the object.

It is important to remark that, as in the case of the geocentric appearance, the openness and directedness that we see in the sensorimotor system is not a capricious error or an arbitrary misconception. It is, given the circular organization of the system *and* our particular location as observers, rather the normal and to-be-expected interpretation. Thus, as in the case of the teleological appearance reviewed in the previous chapter, our tendency to ascribe openness and directedness to the sensorimotor system is, to a large extent, justified. The problem only arises when we assume that such an appearance reveals an intrinsic directedness or openness in the system.

Ancient sailors did not find big problems in orientating their sailings using their own nautical charts, charts that were constructed under the assumption that the Sun orbited around the Earth. The assumption, based on an appearance, was wrong, but that did not render the ancient nautical charts entirely useless instruments. Similarly, I think, cognitive science can treat living beings as sensorimotor systems open to the environment, with inputs and outputs (e.g., information processing theories), or as systems intentionally directed at the external world (e.g., enactivist approaches). Cognitive science can, with relative success, initiate some, perhaps many, explorations using the theoretical maps so constructed. We, who have subscribed to a Strict Naturalism, are not interested in questioning the relative usefulness of such a strategy. Our interest is rather to remain loyal to the primary ontology of living beings as natural systems, avoiding, whenever we can, any observational projection into them.

Suppose we accept the idea that the sensorimotor system of a living being is a closed system (in the specific sense examined above). What might the consequences be for our understanding of perception and action in living beings? What might perception and action be in the context of a closed and directionless dynamic?

4.4 Action and perception

In the previous section we have seen that, in a sort of Copernican turn, Maturana calls for a radical change of perspective with respect to living beings' sensorimotor dynamic. We are invited to pass from a descriptive framework centered on our own observational angle to a broader descriptive framework in which our observational angle appears just as a particular viewpoint, and which does not necessarily reveal the way in which the observed system operates itself.

The way in which the sensorimotor system operates, according to Maturana, is one in which there is no direction or openness, and where what we take to be the system's environment proves to be, for the system, a functionally transparent element, or even more, an entirely inexistent object:

> An observer that sees an effector/sensor correlation as an adequate behavior does so because he or she beholds the organism in the domain of structural coupling in which the distinguished behavior takes place in the flow of its conservation of adaptation. The organism in its operation [, nonetheless,] does not act upon an environment; *the environment exists only for an observer* (Maturana, 2003, pp. 102-103. Emphasis added)

How might this move, if accepted, impact our conception of perception and action in living beings? If the sensorimotor system, as Maturana argues, is a closed system, and if, because of this, the environment exists only for an observer, what does the organism perceive when it perceives? Perception, traditionally, is understood as a kind of 'openness to the world'. The world, with all its entities and properties, is out there, and living beings access that world, directly or indirectly, through their perceptual mechanisms. If, as Maturana thinks, something like the environment or the external world never appears as such for living beings, how can we make sense of the idea of perception? What is the process of perception, then, as a biological

mechanism?

What about action? Action, traditionally, is understood as a specific kind of behavior. Action is the kind of behavior which is *directed* (or intended) in at least two senses: 1) Teleological (directed at some end or goal), and 2) Intentional (directed at the world). In Chapter 2 we said that a living being's behavior is the product of strictly deterministic mechanisms, and in Chapter 3 Ashby tried to show us that the teleological character of said behavior is nothing more than an appearance grounded in complex stability processes. In doing so, we have partially deconstructed the notion of action. Maturana now seems to want to finish the deconstructive task, questioning the intentional directedness of living beings' behavior. According to him, such an intentional behavior, understood as an action executed upon the world, would be also an observer-relative appearance. We have been told that "[t]he organism […] does not act upon an environment; [because] the environment exists only for an observer" (Maturana, 2003, pp. 102-103).

This interpretation strikes us (including myself) as deeply counterintuitive. If we, from our position as observers, see an organism climbing a tree, drinking water, escaping fire, etc., what is wrong in describing what we see precisely in terms of actions; i.e., 'climbing a tree', 'drinking water', 'escaping fire'? If the organism—as Maturana thinks—is not doing any of these things, what on earth is it really doing?

To better understand the reach of Maturana's ideas about perception and action, let us help ourselves, once more, with the geocentric analogy.

If watching a sunset you see the Sun going down, it would not make sense to ask you, in the name of Copernicus, not to see the Sun going down. The fact that you see the Sun going down does not contradict Copernicus' theory. On the contrary, it supports it. Copernicus' theory wants to expose the primary ontology of the solar system, its intrinsic organization, not to eliminate the geocentric picture that appears before your eyes. The picture that appears before your eyes is, in fact, one of the things that the theory aims to explain. "What is going on in the solar system while I see the Sun going down?" That is the kind of question that the theory aims to address. "While you see the Sun going down, what is happening is that the Earth is simultaneously rotating upon its axis and orbiting around the Sun." That is

the kind of answer the theory provides.

Similarly, if observing a living being in its environment, you see it perceiving and avoiding obstacles, finding and eating food, detecting and escaping predators, it would not make sense to ask you, in the name of Maturana, not to see the organism perceiving and avoiding obstacles, finding and eating food, detecting and escaping predators. The fact that you see the organism perceiving the environment and doing different things does not contradict Maturana's theory; in a certain sense, as we shall see, it rather supports it. Maturana simply wants to expose the intrinsic (non observer-relative) functional organization of the sensorimotor system, and explain, not eradicate, the fact that you see the living being as perceiving things in the environment and acting according to such perceptions.

The questions to be formulated would be the following: "What is going on in the living being while I see it perceiving such and such thing and doing such and such thing?" Or "What kind of mechanisms and processes generate what I, placed here in a particular functional node of the system, see as the organism's perception and action?"

The basic elements to answer these questions were already provided in the previous chapters. It just happens that at that moment, when talking about structural determinism or stability, we did not focus our attention on the problem of perception and action. Let us start by reviewing the notion of action.

4.4.1 Action

In Chapter 3 we saw, with Ashby, that the sensorimotor system is basically a first-order feedback mechanism, a closed functional circuit. We also saw that, in some cases (e.g., vertebrate animals), this system works coupled to second-order feedback mechanisms which maintain certain invariance upon physiologically essential variables. In these conditions, what the sensorimotor system is doing, all the time, as any feedback system, is

maintaining a certain correspondence or coordination between its sensor and effector components, i.e., an internal correlation of activity.

What is going on in the living being while we see it, say, detecting and eating food? Let us consider, once more, Maturana's example of the amoeba engulfing a protozoon, but now let us pay attention to the way in which the 'eating' action (i.e., the formation of pseudopods) appears as a sort of 'unintended' result of the amoeba's feedback mechanisms, and only in the descriptive domain of the external observer:

> The presence of the protozoan generates a concentration of substances in the environment. These substances are capable of interacting with the amoeba membrane, triggering changes in the consistency of the protoplasm which results in the formation of a pseudopod. The pseudopod, in turn, causes changes in the position of the moving animal, thus modifying the number of molecules in the environment which interact with its membrane. This cycle is repeated and the sequence of movements of the amoeba is therefore produced through the maintenance of an *internal correlation* between the degree of change of its membrane and those protoplasmic changes *we* see as pseudopods. That is, a recurrent or invariable correlation is established between a perturbed or sensory surface and an area capable of producing movement (motor surface) which maintains unchanged a set of internal relations in the amoeba. (Maturana and Varela, 1987, pp. 147-148. Second emphasis mine).

If we take the point of view of the amoeba and its sensorimotor system, all that we see is the dynamic maintenance of a strict correlation or correspondence between the activity of the sensor and motor surfaces; a dynamic that, as it unfolds, brings as a result a series of structural changes in its protoplasm. That is, its protoplasmic changes appear as a function of the dynamic correlation established between the sensor and effector

components. The correlation, in this case, is positive (a positive feedback mechanism). The virtuous circle is that the higher the number of molecules interacting with the membrane, the more pronounced the structural changes in the protoplasm; changes that, in turn, bring as a result an even higher number of molecules interacting with the membrane, and so on. That is the whole picture in terms of the amoeba.

Now, let us change point of view. How would this sequence of structural changes look to an observer placed outside the amoeba? Since the established correlation is positive, i.e., the higher the activity at the sensor surface, the more pronounced the protoplasmic deformations (and vice versa), we might predict that to an external observer these structural changes would look, more or less, like progressive protoplasmic prolongations (perhaps pseudopods) going toward the source of the chemical stimulation in the environment (which might be a high concentration of certain substances, or perhaps some microorganism). That is, such an observer would see something like pseudopods extending toward some element or unity in the amoeba's surroundings.

Let us have a look now. What do we see? Well, what we see is basically an amoeba extending its pseudopods toward a protozoon. We see a living being executing an action upon the environment.

The case of the amoeba, according to Maturana, is the general case for all sensorimotor systems, with or without a nervous system. What we see as the living beings' action would be a phenomenon that appears only in our descriptive domain as external observers, not something that living beings intrinsically realize.

> In the organization of the living systems the role of the effector surfaces is only to maintain constant the set states of the receptor surfaces, not to act upon an environment, no matter how adequate such a description may seem to be for the analysis of adaptation (Maturana, 1970/1980, p. 51).

We will come back to this point in the next section when analyzing the case of the frog.

4.4.2 Perception

What might perception be for a closed and directionless functional system like the sensorimotor system? If, as Maturana thinks, the environment exists only for an external observer, what is perceived, if anything, during an act of perception? Can we say, from the biological point of view, that there is something like an 'object' of perception?

The answers to these questions are not easy to elaborate, in part because what seems to be at play, if we follow Maturana's analysis, is the very meaning of the concept of perception. Here I will provide only some elements to *address* these questions, without aiming to answer them. A sketch of a response will be presented in the next chapter.

There are two aspects that are important for addressing the problem of perception: 1) The functional closure of the sensorimotor system, and 2) The structurally determined character of the sensory surfaces and of the nervous system in general. Let us start by reviewing 1).

We have seen that the sensorimotor system is a feedback circuit whose circular dynamic basically consists in the maintenance of certain internal correlations of activity. In this context, what we have said with respect to action above is equally valid for the case of perception. Perception, understood as 'openness' or 'access' to an external world, is something that only appears in the descriptive domain of an external observer, but that does not reflect the intrinsic dynamic of the sensorimotor system. What we see as 'objects' of perception for the organism do not appear as such for it, basically because, in terms of its sensorimotor organization, such 'objects' are always included as internal variables of the system, not as separate entities.

This point needs careful handling. When Maturana says that living beings are closed sensory systems, he does not mean that they are blind, deaf or insensitive to the environmental factors. Ashby's cat certainly reacts to the stimuli from fire; i.e., the heat, the light and the smoke perturb its different sensory surfaces and trigger in them specific structural changes. The point is that for the cat's sensorimotor dynamic, none of those stimuli appear as

something external to its activity. Fire, as a source of stimuli, works as one more link within the sensorimotor system, indistinguishable, *from the functional point of view*, from any synapse occurring within the nervous system.

From the material point of view, of course, fire is very different from the neurotransmitters that act at the level of the synaptic gaps, and we, as external observers, clearly see that fire is located outside the cat's physical boundaries, whereas neurotransmitters are inside. However, these distinctions are functionally transparent for the sensorimotor system as a closed circuit. Fire, as an object that is external to the cat's sensorimotor system, is something that only appears for us as observers.

All this may sound counterintuitive, but let us try to go on with the analysis. Suppose we accept, just to do the philosophical exercise, the idea that perception, from the point of view of the sensorimotor system, does not correspond to an 'openness' to the external world, that what we take to be the 'objects' of perception do not have operational presence in the sensory dynamic of the organisms. There still remains the fact that the sensory surfaces, as we saw at the beginning of this section, react as transducers before specific configurations of the environment. Perhaps the overemphasis put on the *functional* point of view, somehow, masks the real fact that the sensory surfaces, after all, do make contact, directly or indirectly, with something that is *structurally* external to them, that is, with the environment.

It may be that the cat's sensorimotor system, in its constant 'monologue', so to speak, never 'realizes' that certain stimuli come from a world of objects that is outside the cat as a physical system, and that is entirely different from the synaptic 'world' of the nervous system. Yet that does not eliminate the fact that the energy from the fire, which we know is out there in front of the cat, is received or detected, in terms of transduction, by the different sensory surfaces of the cat. So it appears there is still a sense, a real one, in which what we call perception, even if it does not have the kind of directionality or openness we think it has, is not entirely in the eye of the beholder. But is this appearance misleading?

Perception and structural determinism

Effectively, we have said that the idea of functional closure does not aim to deny that living beings, through their sensory surfaces, react before, and to that extent 'sense', the different stimuli coming from the environment. But there is an important point that has not been considered yet in the analysis. Sensory surfaces and the nervous system in general, like any physical system, are structurally determined systems.

In a structurally determined system, as we saw in Chapter 2, all that counts are the structural states of the system, irrespective of the way in which these states are brought about. This is because the states of the system, whatever their nature, are always determined by the system itself. Recall the example of the laptop presented in Chapter 2. Laptops are full of transducers, generally artificial tactile sensors. All the buttons, or, in some cases, the icons of the touchscreen, operate as transducers that convert mechanical energy (mechanical pressure) into electrical charges. By pressing a button we turn on the laptop. When the laptop is on, the structural state 'on' exists and has the operational presence that it has, irrespective of whether it is reached thanks to the triggering action of my finger, another person's finger, a screwdriver, the impact of a stone, etc. That is, there is a disjunctive series of external factors that can trigger the same structural state in the system, and the system, in responding the way it does, does not seem to be able to distinguish among the different external factors.

Even more, the structural state 'on' exists and has the operational presence that it has, irrespective of whether it is reached thanks to the triggering action of some external factor operating upon the transducer button or thanks to some internal dynamic of the laptop (e.g., a self-programmed function). Once the state 'on' is present in the system, it is simply there, regardless its origin. From the point of view of the laptop, a structural state *X* reached through the action of external factors is indistinguishable from a structural state *X* reached through an endogenous dynamic.

Who distinguishes the origin of the structural states of the laptop in each case, whether external or internal, is the observer, not the laptop. For the

laptop and its structural dynamic there is nothing external; all that exists is the presence of its own structural dynamic. We observers see the laptop and see it surrounded by a whole world. However, the laptop, from the point of view of its structural determinism, operates in a sort of self-contained metaphysical space where 'the external' has no place.

Living beings and their sensorimotor system, including the nervous system (when they have one), are structurally determined systems. Sensory organs, as transducers, exactly behave like the laptop's buttons. Every time some external factor triggers a structural change in a sensory organ, all that appears and exists for the system is its own structural change associated with its own energetic modality. Once this structural state is reached, all that counts, for the structural dynamic of the nervous system, is that the sensory organ is effectively in that state and not in another. The way in which this state has arisen in the sensory organ does not have any operational relevance for the system, which operates, moment after moment, only according to its structural present. As in the case of the laptop, once a determinate structural state is present in the system, it is simply there, and its origin is not relevant.

The situation does not change even if the sensory organ reaches a determinate structural state, systematically, only in presence of a specific kind of perturbation, and never as a result of a spontaneous internal dynamic. Even so, it remains the case that the nervous system operates with the structural states found at each moment, just as they are found, without making distinctions regarding their origin. The external factors, for the sensory organ and for the nervous system in general, never appear as what they are for us observers, namely the causative and responsible agents of the system's structural change.

It is the observer, not the nervous system, who can distinguish the different origins of the structural states of the sensory system. The sensory system, as any structurally determined system, exists and operates in its structural present; whether its present structural state is endogenously generated or externally triggered is a distinction that has no operational presence in its dynamic of states. In such circumstances, there is no way, so to speak, in which the sensory structure can 'communicate' to the rest of the nervous system the origin of its structural states. The origin of the structural

states disappears in the structural determination of the system.

This metaphysical condition, which is no more or less than PSD (see the metaphysical principles examined in Chapter 2), easily explains, as we shall see, phenomena such as illusion and hallucination. But these phenomena, in turn, open the door to a complex and old philosophical problem with respect to perception.

4.4.3 The problem of perception: on frogs and drifting ships

The problem alluded to above can be set in a very simple way. Let us present a fly in the visual field of a frog. Then, let us conserve the structural state of its retina and of all its nervous system (disregard the technological plausibility of the procedure). While we do this, let us remove the fly from the frog's visual field. Is the frog still seeing the fly?

If we take as a point of reference our position as external observers, probably we would answer "No, the frog is not seeing the fly; the frog is having a hallucination whose hallucinatory content is a fly". But if we take the point of view of the frog's nervous system, what would we say? Someone might think that the answer in this case should be "Yes, the frog is seeing the fly, because the structural state of its nervous system is exactly the same structural state that takes place when the fly is actually there in its visual field". But, if we take seriously the position of the frog's nervous system as a structurally determined system, could we really even answer the question?

If the frog's nervous system is a structurally determined system and if, because of this, all that counts for its operational dynamic are its own structural states, irrespective of whether or not they are brought about thanks to the triggering action of some external factor, can there be for such a system any possible distinction between seeing (perceiving) and having a hallucination? Maturana, and I with him, think that such a distinction is not

possible for the frog's nervous system. What is more, we argue that such a distinction is not necessary for its functioning as a biological system (this last point will be discussed in the next chapter).

The distinction between perception and hallucination, or between perception and illusion, is a distinction about the *origin* of the structural states of the sensory system, not about the structural states themselves. We as observers can identify such and such a structural configuration in the frog's environment (e.g., the presence or absence of a fly) and contrast it with such and such a structural condition in the frog's nervous system. But this is precisely the kind of distinction that the frog's nervous system, operating entirely in its own space of structural states, cannot make. The frog's nervous system cannot make such a distinction because, like the laptop in the previous example, its structural dynamic unfolds in an entirely self-contained metaphysical space, without reference to the outside. For the frog's nervous system, and this time from the point of view of structural determinism, there is nothing 'external' to its operations.

But how can this be? Are we implying that the frog's nervous system, somehow, operates in a sort of vacuum, in the absolute darkness of its self-contained metaphysical space? Well, yes, that is more or less what we are saying. But, is the frog's nervous system *really* operating in a vacuum? Well, in a sense, no, of course not. There is a whole world around the frog, with plants, flies, etc. So what is the point? The point is that those who can distinguish the existence of such a world are we external observers with respect to the frog, not the frog's nervous system. The frog's nervous system does not 'have in view' the external world that we behold (strictly speaking, no world in particular). The frog's nervous system operates upon its own states and according to its own rules, only maintaining certain internal correlations of electrochemical activity.

As an illustration, slightly modifying Maturana's classic example (1970/1980), let us think of a man who has lived all his life below the decks of a ship, without knowing anything about the world outside the cabin. He has been conditioned, through hypnosis sessions after which he only remembers the conditioning rules, to maintain, by moving a series of buttons and levers, a certain pattern of correlations in the values displayed on a

screen. He is ignorant about the meaning of these values, and about the fact that they are causally related to some external factors. (The ship is full of very sophisticated sensor devices, but he does not know that). He is also ignorant about the overall result of his maneuvers and performs his task in a ritualistic way, mechanically, as after so many hypnosis sessions he really does not know why he is maintaining a certain correlation of values on the screen.

The man's actions are such that, despite his absolute ignorance, the ship gracefully navigates through the seas, avoiding reefs, getting away from storms, and doing all the things that a ship driven by a helmsman would do.

Is the ship being driven by a helmsman, a pilot that controls its course in light of having in view reefs, waves, winds, storms? Or perhaps, if not having in view these external objects as such, at least trying to infer, guess or predict, from his position, the causal structure of the world around the ship? Well, not really. When we open the physiology of the ship, so to speak, all we find is someone obsessed with an internal game of meaningless values, playing the game without knowing why. That man is not driving any ship, nor trying to know, infer or predict anything about the external world.

The amazing fact is that, in spite of operating in the dark, i.e., without any reference or concern toward the external world, his actions result in the graceful displacement of the ship through the seas. The amazing fact is that the ship, although from the outside appearing to be driven by a helmsman, is indeed drifting. It is gracefully drifting through the seas.

The frog's nervous system, says Maturana, operates more or less like that man, i.e., with no access to what we external observers distinguish as 'the external world', with 'reefs', 'storms', 'plants' and 'flies', and without performing any role that might be compared to the role of a helmsman or pilot who controls the frog's navigation through the world. The frog's nervous system, strictly speaking, is not 'trying' to do anything. It does not have any 'task' upon its shoulders, any 'duty' to fulfill. It is not trying to solve any epistemological problem with respect to the external world either, deciphering or inferring its causal structure, predicting or guessing future events.

When we see a frog catching a fly, we assume the frog perceives the fly,

targets it, and shoots its tongue to trap it. But if the frog, its sensorimotor system, and its nervous system as a whole are a structurally determined system, such a description can hardly reveal what is going on in the frog, "no matter how adequate such a description may seem to be for the analysis of [the frog's] adaptation" (Maturana, 1970/1980, p. 51). What is going on in the frog is an internal game of sensorimotor correlations focused on themselves, whose 'blind' and unintended result corresponds to what we see as the action of 'catching a fly'.

But, the reader might insist, if the frog's nervous system is operationally equivalent to the man in the ship, if it operates in the absolute darkness, how does the frog manage to target the fly and catch it? Well, that is the whole point, we reply. The frog never targets the fly. Right, the reader might concede, the frog does not target a 'fly' but something more unspecific or abstract than that, perhaps something like 'flying black spot', or simply 'food'. No, we reply, the frog does not target anything at all. It is the very action of targeting that is absent in the dynamic of the frog's nervous system.

Wait a minute. What are we talking about?

Perhaps it is important, at this point of the discussion, to remind ourselves of the Strict Naturalistic frame we have assumed with respect to living beings in this thesis. The so called 'problem of perception', at least in philosophy, is mainly, though not exclusively, an epistemological problem. We want to account for the fact that the animal, in this case the frog, successfully adapts to the environment, and to that extent, somehow, 'knows' the world in which it lives. In general, we see that animals adapt to their environment, and this ability reveals to us some form, perhaps very basic, of knowledge or cognition. But in assuming a Strict Naturalistic stance with respect to living beings, what we have done is precisely to remove this kind of question. Or, more or less equivalently, to expose the conventional and arbitrary character of its exclusive association to living beings, showing that, after all, every physical system exists in structural coupling and adaptation.

So perhaps, before going on with the discussion, it is worth clarifying our respective expectations. If the reader is hoping we will, at some moment,

somehow, do the trick of baking an 'epistemological pie' of perception by using only 'structural ingredients', then we are expecting different things. This thesis, recall, is about the biological "roots" of cognition. Our purpose is to examine the primary ontology of biological systems, and show that their behavior and adaptation, to which we tend to attribute an epistemic or cognitive dimension, appear as natural results of said ontology. We have assumed that living beings, like all other natural systems, relate to the world in strictly structural terms, not in epistemic, informational, intentional or semantic terms. That is why we remarked, at the beginning of this section, that every sensory process, i.e., every transduction phenomenon, is a subtype of structural interaction and nothing more than that. (See again in chapter 1 section 1.2.1 'The cognitive construction of living beings').

That is what we are talking about.

Now let us go back to the frog. The frog never targets something in the world, we were saying. The amazing thing, as in the case of the ship, is that when the frog shoots its tongue, there is a fly right there available to be trapped. How does that happen?

Notice that once we take a Strict Naturalistic stance, the nature of the question about action and perception starts to change. We do not ask "How does the frog's nervous system target the fly and drive the tongue to trap it?" Our problem now is different. The question that we have to face is: "How does it happen that when the frog's nervous system, in the vacuum, without targeting anything in the world, and without any purpose in particular, shoots the frog's tongue, there is a fly right there in the world available to be caught?"

Our question is about the coherences of the structural coupling between the frog and its environment, or better put, between the frog and its niche. We need to explain the congruence that we observe between what the frog does and the particular structural configurations of its niche, the apparent synchronicity, so to speak, between the fly and the frog's tongue. And we have to do it acknowledging or assuming that the frog's niche cannot specify or instruct the structural dynamic of the frog, and that the frog's nervous system, both in its functional organization and structural determinism, operates as a closed system. As in the case of the ship, we have to explain

the effective navigation of the frog through the world, the conservation of its adaptation, assuming that there is no ‘pilot’ in charge of the navigation.

We are, it seems, placed in front of a new kind of problem. It is a big one, mainly because our resources to face it are quite limited. For many, perhaps, explaining the behavior of a simple frog may seem an easy and basic task, not a problem. But recall that within our framework there are several restrictions. We cannot appeal to epistemic notions, neither open nor dressed up (e.g., information), we cannot use representations or semantic relations of any kind, intentionality (with or without content), control mechanisms, agency, normative aspects, or teleology. All we have is a bunch of metaphysical principles, the notion of stability, some thermodynamic considerations, and the closed character of living beings.

Perhaps—we have to admit the possibility—with this class of resources, the problem we have to face, even if modest, turns out to be an unsolvable problem, at least for now.

The challenge is there, anyway, and I think we should go for it. Even if after trying our best the reader remains unconvinced, I am sure the intellectual effort, or the mere exploration in itself, will have been worth it. Next is a brief chapter where we will try to articulate and consolidate all we have said so far about living beings. We will try to test, partially, the explanatory force of the ontological characterization of living beings offered so far.

Chapter 5

Structural drift: adaptation, intelligent behavior and cognition

In this thesis, following the requirements of a Strict Naturalistic approach, we have characterized living beings as (i) adaptive dynamic systems, (ii) deterministic machines of closed transitions, (iii) multistable dissipative systems, and (iv) organizationally closed systems with respect to their autopoietic and sensorimotor dynamics. Now it is time to apply these concepts and see whether we can explain and understand the behavior of living beings on the basis of said ontological characterization.

In the previous chapter we argued that action and perception, as understood in their traditional sense, correspond to distinctions made by the observer in the interactional domain of living beings, not to intrinsic processes of living beings. We took the representative case of a frog catching a fly, and characterized the frog through an analogy to a drifting ship that, in spite of lacking a helmsman, gracefully navigates through the world conserving its adaptation. Now we have to explain the operational coherences that we observe between the living being and its environment, or in our example, between the frog and its specific niche.

Our question was: “How does it happen that when the frog’s nervous system, operating in the dark, without targeting anything in the world, and without any purpose in particular, shoots the frog’s tongue, there is a fly right there in the world available to be caught?”

5.1 The frog and its circumstance

The previous question can be split into at least two sub-questions. First, if the frog's nervous system is unable to distinguish the origin of its stimulation, how does it "know" that the triggering factor is a fly and not another thing? How does it "know" that this is not an illusion, or even a hallucination? Second, even assuming that the frog's nervous system is able to identify the presence of the fly, how does it manage to shoot the frog's tongue at the right moment in the right direction? If the frog's nervous system operates in the dark and lacks internal maps or models of the external world, how can we explain the accuracy of the frog's action?

The first answer is that the frog's nervous system actually does not "know" that there is a fly out there in the world. The frog's nervous system does not react to the presence of a fly but to a certain stimulation pattern on the frog's retina. That is why, as is well known, it is relatively easy to "fool" a frog by moving a little piece of black paper in the air. Any object able to produce the "right" stimulation pattern on the frog's retina is able to trigger the same reaction. So the question should be another one. How does it happen that, in natural conditions, the frog's tongue finds a fly and not something else that would produce a structurally similar retinal stimulation pattern (e.g., a piece of black paper), out there in the world? This, notice, is not an epistemological question. We are not asking about an epistemological correspondence but about an ecological congruence. We are asking about the existence and availability of flies in the frog's natural niche. The answer to this question, without going into details, is relatively simple, and has to do with the fact that frogs and flies share a common evolutionary history in the biosphere; a history in which they have each become part of their respective ecological niches. Frogs find flies in their environment with the same naturalness they find ponds and muddy lands, and flies find frogs in their environment with the same naturalness they find feces, decaying meat and rotting fruits. This is not a matter of providential harmony or an 'invisible hand' that arranges the things for the convenience of everyone. It is matter of a shared natural history in which different species and their respective niches

have jointly evolved, resulting in a certain ecological coherence in which they conserve their adaptation.

The answer to the second question has to do with the structure and functional organization of the frog's sensorimotor system. The sensorimotor system is essentially a feedback mechanism, a closed functional circuit. In such a system, as we saw in the previous chapter, the environment and its triggering factors work as functional nodes of the circuit, not as external elements. The environment and its triggering factors constitute elements through which the sensorimotor network coordinates its own activity. In other words, when we ask about the congruence between the frog's action and its circumstance, we are asking, ultimately, about the congruence of a feedback system with itself. Let us see the point through a simple example of closed senso-effector system.

In a room X, the decrease of the temperature to 18 degrees triggers a structural change in the sensor component of a thermostat. This structural change triggers in turn a cascade of structural changes in the thermostat such that the result is the activation of a heater (the effector component). The heater raises the ambient temperature to 20 degrees, at which point the sensor component suffers a new structural change that triggers a subsequent cascade of structural changes in the thermostat. The result now is the deactivation of the heater. The overall situation is that the ambient temperature in the room X constantly oscillates between 18 and 20 degrees, without going out of this range. Taking into account that the thermostat is a structurally determined system, and that, consequently, it absolutely works within the limits of its own structural states, without having in view anything like rooms or ambient temperatures, we can formulate the following question. How does it happen that when the heater initiates its activity there is a room X out there in the world with an ambient temperature of 18 degrees? The answer is that this happens, first, simply because the system is assembled as a closed circuit. If the heater was located in a different house, it would not act upon the room X but upon another one. Second, the fact that when the heater initiates its activity the room X is at a temperature of 18 degrees, and not of 15 or 23 degrees, is because the thermostat has been set to be activated by an ambient temperature of 18 degrees, and not by another

one. Why does not the heater, once activated, go on working and raising the ambient temperature more and more? Why does it cease to work at 20 degrees? This happens because of the negative polarity of the thermostat's wiring. If the thermostat's wiring was inverted to a positive feedback circuit, the heater would go on working and raising the temperature until breaking down.

The thermostat always works within the metaphysical boundaries of its structural determinism establishing internal correlations of activity. What explains its 'successful' behavior with respect to the room X is its particular organization and structure (its wiring, its polarity), and the way in which it is coupled to the ambient temperature of the room X.

In the case of the frog the explanation is essentially the same. First, the frog shoots the tongue at the right moment, i.e., only when the fly is present in its visual field (neither before nor after that), because it is the very fly, with its appearance in the frog's visual field, that triggers and initiates the response in the frog's sensorimotor system. The frog's sensorimotor system has been set, by evolution, to react in that way only before the stimulation pattern caused by that kind of flying object (which, in its specific ecological niche, is usually a bug).

Second, the frog shoots the tongue in the right direction thanks to the particular anatomical and functional organization of its sensorimotor system as a closed circuit. As with the thermostat, if we change the wiring of the system, the spatial accuracy of the response will change too. That this is the case is dramatically revealed by the classical experiments of Roger Sperry (1943, 1945). Sperry changed the anatomical configuration of the frog's sensorimotor system by rotating the retina of one of the frog's eyes 180 degrees. Once recovered from the surgery, the frog was able to catch its prey without any problem if the rotated eye was covered. When the rotated eye was uncovered (and the normal one covered), however, the frog reacted by shooting its tongue with an exact deviation of 180 degrees with respect to the position of the prey. Sperry reported that the frog never could change its response, persisting in its 'error' until the end. That is, the frog's sensorimotor system, under this new anatomical organization, continued running its internal correlations strictly following its own patterns of

operation, irrespective of the dramatic changes produced at the level of the frog's external behavior.

When Maturana saw this experiment, he asked "Is this an error in targeting or an expression of a new internal [sensorimotor] correlation?" (Maturana and Varela, 1987, p. 126).

The frog's sensorimotor system always operates within the metaphysical boundaries of its structural determinism, without having access to the spatial coordinates of the external world. The spatial accuracy of the frog's behavior has to do simply with the specific way in which its sensorimotor system is anatomically and functionally organized, not with the possession of alleged 'internal maps' of the world. Change the wiring of the system and you will change the response. Spatial categories such as 'up' and 'down', 'front' or 'back' only exist for the observer, not for the frog's sensorimotor system (Maturana and Varela, 1987).

Properly viewed, says Maturana, the frog's sensorimotor system, after the surgery, continues running its internal correlations with strict accuracy (that is why the observed deviation is of exactly 180 degrees!). It is only from the viewpoint of the external observer that the behavioral result of its new correlation counts as an 'error'. The sensorimotor system, strictly speaking, like any natural system, never makes mistakes.

Now, what is important in the example is not that the frog cannot 'correct' the behavior, but that the sensorimotor system, whatever its organization, always operates under its own structural determination. What validates the idea of structural determination is not the rigidity of the frog's behavior, i.e., the fact that it never 'learns' to trap a fly under the new wiring, but that the change from one kind of behavior to the other depends on the structural changes suffered by the frog's sensorimotor system. Otherwise, any instance of behavioral flexibility—what we usually call 'learning' or 'intelligence'—would speak against the structural determination of the sensorimotor system. Let us examine this point more closely.

5.2 Learning and intelligent behavior

In a cognitive observational context, i.e., when living beings are constructed as cognitive systems, it is usually said that an organism is intelligent if it is able to learn, if it is able to modify its behavior in order to achieve some goal. In such an interpretative context, as a general rule, and broadly speaking, behavioral rigidity is an indicator of lack of intelligence and behavioral flexibility of intelligence. What matters, to talk about intelligence, is not the achievement of the goal itself, the success in the task, but whether the organism, put in a different scenario, is able to modify its behavioral strategy and find an alternative way to achieve the goal.

A typical example of lack of intelligence, within this cognitive descriptive frame, is the rigid behavioral pattern exhibited by the dung beetle. The beetle digs a nest and lays its eggs. Then, it goes for a ball of dung and carries it toward the nest. The observer assumes that the beetle's purpose is to supply enough food for the future larvae. Then, while the beetle is dragging the ball, the observer removes the ball from its grasp and sees what happens. An intelligent reaction, to her eyes, would be reinitiate the action and go for another ball. The beetle, however, goes on with its routine and pantomimes plugging its nest as if the ball was still there. The observer repeats the experience several times and the beetle never 'learns' the lesson, thus evidencing, to her eyes, a very low level of intelligence. The (cognitive) observer comments: "Evolution has built *an assumption* into the beetle's behavior, and when it is violated, *unsuccessful* behavior results", or "[the organism] is unable to learn that its innate *plan* is *failing*, and thus will not change it" (Russell and Norvig, 2010, p. 39. Emphasis added).

A dung beetle, in normal conditions, is successful in its behavior and manages to plug its nest with the ball. But that, for the cognitive observer, is not genuine intelligence. The beetle is successful just because it happens that all the environmental conditions it 'assumes' as given (according to its innate 'plan'), are in place. If the 'assumption' is 'violated', the beetle is revealed as an extremely fragile cognitive agent, unable to learn and modify its behavior.

In light of the previous example offered by Sperry, wherein the frog never changes its behavior and, in a sense, 'pantomimes' catching a fly where there is no fly at all, it might seem that assuming structural determinism in living beings might explain, at most, cases of 'stupid' behavioral rigidity, or cases where there is successful behavior but only thanks to that all the 'assumed' conditions are given.

Like Sperry's frog, the dung beetle persists in a sensorimotor routine that, from the adaptive point of view, goes nowhere. It is as if the beetle's nervous system operated locked up in its own internal affairs, entirely disconnected from the behavioral consequences of its sensorimotor routines; all of which, admittedly, is consistent with the idea that the nervous system is a structurally determined system that operates in the dark without having the external world in view. Is this the kind of behavior that PSD aims to explain?

In the previous section we explained why, in spite of the structurally determined character of its sensorimotor system, the frog, in normal conditions and within its natural niche, succeeds in catching a fly. However, the action of catching a fly, someone might say, tells us almost nothing about learning and intelligence. The frog's action is successful, true, but only because, as in the case of the dung beetle, all the 'assumed' ecological conditions are in place. Is this the explanatory scope of PSD?

Sperry's study is an extreme case of sensorimotor rewiring. Under less violent or extreme manipulations, the frog's nervous system is able to 'correct', after some time, its patterns of activity and produce what the observer appraises as the 'right' kind of behavior. Actually, in natural conditions, the animal's nervous system is constantly, at minimal scales, changing its structural composition and system of connections. There is, so to speak, a continuous process of 'smooth and fine rewiring' which is contingent to the history of the organism's structural coupling, and that is the base of those phenomena that we call 'learning'.

Does all this contradict the structural determination of the sensorimotor system? No, it does not. That a system is structurally determined does not mean that its structure cannot change; it just means that all its structural changes, either externally triggered or internally generated, are specified by its own structure. But then, the reader might wonder, how can we explain

learning processes in living beings?

Learning, understood in the cognitive sense of 'intelligence', is not merely behavioral modification, but behavioral modification in the 'right' direction. A sensorimotor system is able, in principle, to undergo many structural changes and rewiring processes, generating many forms of internal correlation and behavior in the organism. If the sensorimotor system operates in a vacuum, only 'concerned' with its own structural states, without having in view the external results of its dynamic, how does it happen that only certain correlations are consolidated, and that they are precisely those that produce the 'right' kind of behavior for the organism? How can a system that lacks a pilot, that is drifting through the world without any purpose or control, that has no idea about what is going on outside, 'learn' and generate the right kind of behavior?

Contrary to the case of the dung beetle, a paradigmatic case of intelligent behavior in animals is that of rats' spatial orientation. Let us take Tolman's classical studies (Tolman, 1948; Tolman & Honzik, 1930). Greatly simplified, in these studies a rat is put in a maze for a couple of days and left to freely explore the space. The maze has several alternative ways out (A, B, C). Then, the rat is deprived food and hungry, is put back in the maze. At the exit of the maze, this time, there is food. The rat finds, after a couple of 'errors' but without much difficulty, a way out, say A, and eats the food. Just luck? Next time the experimenter blocks the way out A. The rat goes toward it, finds it blocked, then reorients its movements in different directions, finds an alternative way out, again without much difficulty, and eats the food.

How can the rat do that? Tolman's classical answer, an early precursor of cognitivism, was basically that 1) the rat, during the exploratory period, had built an 'internal map' or spatial representation of the maze, and 2) that the rat's behavior was not ruled by simple stimulus-response mechanisms but by goals and purposes (Tolman, 1948, 1932).

How might a structurally determined system, with all the restrictions we have detailed, be able to generate that kind of intelligent behavior?

This is a good question. In fact, it is a question that exceeds the limits of the sensorimotor system. To properly address it we need to contextualize the functioning of the sensorimotor system within the nervous system as an

integrated unity, and ultimately, within the organism as a thermodynamic system. That is, we need to refer to the ontology of the living being as a whole.

5.3 Back to the biological 'roots'

The sensorimotor system is one division of the nervous system. It is a first-order feedback mechanism that closes on itself through the environment. But the nervous system has other senso-effector divisions (e.g., the autonomic division), some of which operate upon essential physiological variables (in mammals, generally through connections with the endocrine system at the level of basal ganglia or equivalent structures). These subsystems operate, with respect to the sensorimotor system, as second-order feedback mechanisms. That is, they maintain a set of essential physiological variables within specific ranges, in part, and indirectly, through operations upon the sensorimotor system (constituting what Ashby called 'ultrastable systems').

Now, the so called essential variables are 'essential' because they have to do, ultimately, with the thermodynamic stability of the organism as a dissipative system; i.e., with the availability of matter and energy for the tissues and cells of the organism, and for their respective metabolic processes (e.g., chemical energy through meals, oxygen concentration, hydration, heat conduction and dissipation).

Animals in general (rats, frogs, dogs, dolphins, etc.) are complex dissipative structures composed by billions of autopoietic unities, each one a dissipative structure itself. As with any dissipative structure, animals are thermodynamic systems that exhibit *stability* in far-from-equilibrium conditions. That is, they are constantly compensating for the different disturbances that affect their respective thermodynamic balances. Their peculiarity among dissipative structures lies in the possession of specific stability mechanisms, among which the feedback mechanisms embodied in their nervous system are a key element.

Second-order feedback mechanisms maintain stability around essential

physiological variables, in part, operating upon the sensorimotor division of the animal's nervous system; activating and deactivating changes in the sensorimotor correlations. The sensorimotor system, which in principle may generate many different forms of internal correlations, and thus a great variety of behaviors, is restricted in its operations by these second-order feedback mechanisms.

Importantly, the presence and influence of these second-order mechanisms over the sensorimotor system is variable among the different animal species, and is limited in line with the specific behavior and the concrete anatomical and physiological conditions of the animal (as it is evident in the experiments of Sperry or in the case of the dung beetle). But in normal conditions, and every time the behavior is immediately critical regarding the thermodynamic stability of the organism, these mechanisms always modulate the activity of the sensorimotor system so that the metabolic integrity of the animal is conserved.

Let us go back to the example of the dung beetle. The behavior of rolling a dung ball with the legs towards the nest is, for the observer, a very important behavior, but not an immediately critical behavior for the beetle's thermodynamic balance. If the ball is removed, there seems to be no special second-order mechanism ready to activate new sensorimotor correlations and modify the behavior. What we see is a rigid behavioral pattern without correction capability. However, if we remove the dung ball not when the beetle is rolling it with its legs but eating it with its mandibles, the result is different. The beetle does not continue 'biting air' and pantomiming chewing movements. Why? In this case, the sensorimotor routine that generates the action of eating, which is immediately critical for the animal's thermodynamic balance, is penetrated by second-order mechanisms able to activate new correlations and modify the behavior (e.g., go for another ball). The (cognitive) observer, witnessing this correction capability, perhaps would comment: "I knew the beetle could not be that stupid".

The sensorimotor system generates certain variability of behaviors, but most of the time within the margins imposed by the second-order feedback mechanisms. That is, under the selective pressure of the thermodynamic demands of the organism. That is why the sensorimotor system, in spite of

operating within the strict confines of its structural determinism, systematically stabilizes internal correlations that result in what the observer sees as ‘adaptive’ or ‘successful’ behaviors.

When Tolman’s rat is deprived of food, its energetic balance is displaced from the normal physiological values. Since the rat is a *stable* thermodynamic system, this very condition activates second-order feedback mechanisms which, indirectly through the activation of the sensorimotor system, operate as compensatory thermodynamic forces. The rat’s sensorimotor correlations, under the pressure of the second-order feedback mechanisms, generate the series of movements that we see as the purposeful behavior of ‘looking for food’. This sensorimotor activation, which operates without ‘having in view’ the maze or the future consequences of its dynamic, does not stop until the rat’s energetic values are brought back to their physiological values and the second-order mechanisms are deactivated. What we see as a result of this dynamic is that the rat, in spite of obstacles (e.g., blocked paths), persists in its search and varies its behavior until it finds a way out. (There are more elements to take into account in the explanation of the rat’s behavior, but for now these considerations are enough to make our point. We will come back to Tolman’s rat in the next section).

Let us go back to our questions. If the sensorimotor system operates in a vacuum, only ‘concerned’ with its own structural states, without having in view the external results of its dynamic, how does it happen that only certain correlations are consolidated, and that they are precisely those that produce the ‘right’ kind of behavior for the organism? How can a system that lacks a pilot, that is drifting through the world without any purpose or control, that has no idea about what is going on outside, ‘learn’ and generate the right kind of behavior?

The answer, I think, is that learning phenomena have to do with the fact that the sensorimotor system operates under the stabilizing force of ultrastable mechanisms linked to the thermodynamic demands of the organism as a dissipative structure. It is this particular functional organization of the organism and its nervous system that explains, in good part, the organism’s ability to modify its behavior in the ‘right’ direction and

conserve its adaptation.

The exact anatomical and functional details of these complex mechanisms still need to be clarified, but their operational logic, at least in abstract terms, is relatively clear. More importantly, their functioning is entirely intelligible within the deterministic metaphysical principles assumed in this thesis to be valid for living beings as natural systems.

Animals, as thermodynamic autopoietic machines, and their nervous systems, in all their divisions, are fully deterministic machines. Everything that is valid, in terms of structural determinism, for the sensorimotor system as a first-order feedback mechanism is also valid for second-order feedback mechanisms and for the nervous system in general. All the divisions of the nervous system are linked in one way or another, but none of them 'has in view' what is going on in the others, or estimates the consequences of its activity for the organism. There is no general plan or purpose in their functioning.

Second-order feedback mechanisms do not work, as it might seem, in order to ensure the survival of the animal (we have examined this point in detail in Chapter 3). They are just stability mechanisms operating upon certain physiological variables, able to trigger variations in the sensorimotor correlations of the nervous system.

The behavioral flexibility exhibited by animals, usually interpreted as intelligent and purposeful, is the result of a complex but strictly deterministic form of thermodynamic ultrastability in far-from-equilibrium conditions. No need for helmsmen or pilots that control the animal's behavior according to internal maps, plans or purposes. The notion of structural drift, although counterintuitive at first glance, now appears as a valid alternative for explaining the behavior of living beings.

5.4 The importance of good 'logical accounting'

The emphasis Maturana, in discussions of perception and action, puts on the structurally determined character of living beings and their nervous system

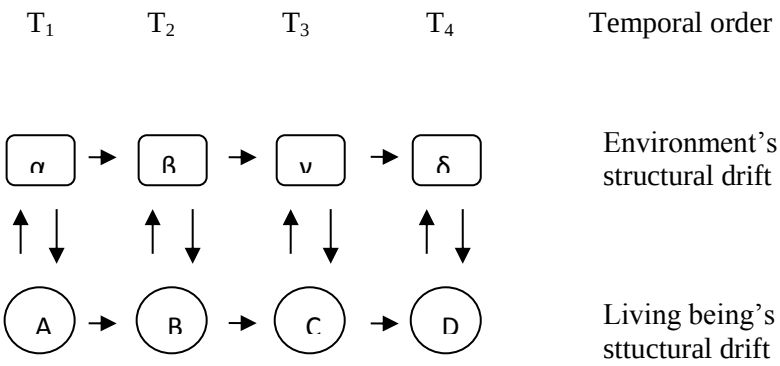
has been criticized, quite often, as excessively internalist (Godfrey-Smith, 1996), idealist (Johnson, 1991) or frankly solipsist and anti-realist (Searle, 1995; Zolo, 1991).

Initially, when hearing Maturana talking about living beings (with the metaphor of the nervous system as a man locked up in a ship), it would seem that the external world is completely absent and plays no role in a living being's behavior; that its presence is entirely irrelevant, that all that counts is the internal structure of organisms.

However, on a second reading, trying to properly grasp the metaphysical meaning of PSD, one can see that this is not the case. To avoid misinterpretations it is crucial, as Maturana recommends, to maintain clear 'logical accounting' with respect to our observational domain and explanatory practices.

PSD aims to clarify the metaphysical nature of the interactions that a system holds with its environment, not to deny the existence of such interactions or declare them as dispensable for our explanatory practices. When Maturana assumes PSD as a metaphysical condition of living beings, he never draws the conclusion that, since living beings are structurally determined systems, all we need to explain living beings' adaptation and behavior is the study of their internal structure. It is one thing to say that the environment cannot specify or instruct the structural changes of a system, and another to say that the environment is therefore irrelevant.

Maturana *starts,* as we did in Chapter 1, by presenting living beings as dynamic systems that exist in continuous structural coupling with the environment. That is, as adaptive systems:



As we saw in Chapter 1, all interacting dynamic systems endowed with structural plasticity, natural or artificial, living or not, are adaptive systems. Some of them exhibit passive adaptation, some of them active adaptation. Living beings have interesting peculiarities among active adaptive systems (ultrastability, far-from-equilibrium thermodynamic regime, autopoiesis), but none of them, as we have seen in this thesis, place them outside PSD.

PSD does not claim that the vertical arrows in the diagram above are a myth invented by the observer (that such interactions do not exist). What PSD claims, ultimately, is that those arrows must be understood as strictly *structural* interactions (not as instructive, informational, semantic, epistemic or intentional interactions).

This is quite different from saying that the environment does not exist, or that external factors are irrelevant to the structural drift of living beings.

That living beings exist in structural coupling implies that the environment is constantly triggering structural changes in them, and that, consequently, the environment forms part of their history of structural drift. Every living being, at each moment, is where it is and has the structural state that it has as a result of its particular structural drift; a structural drift that is shaped, moment after moment, by its own internal dynamic *and* by the structural changes triggered by the environment (Maturana, 2003).

In this sense, the structural present of the living being, as with any adaptive system, is always the embodiment of its history of structural coupling. And in this sense too, *none of the interactions of the living being with the environment is trivial for its structural drift* (Maturana, 2003, 1987). None of its encounters with the environment, so to speak, vanishes into the air; everything is embodied in its structure. But here is where we face a dangerous bend in our analysis, and where Maturana asks us to reduce our speed, to pause, and maintain as clearly as we can the 'logical accounting' of our descriptive and explanatory exercises (Maturana and Varela, 1987).

In Chapter 2 we saw that living beings are machines (i.e., state-dependent systems), and that as such, they exist as actuality (AP). Ashby, with his discussion of the concept of memory, showed us that a machine strictly operates in its structural present, and that its past or history has no operational presence in the generation of its behavior. This is the case even

when, as we are seeing now, the structural present of the machine is effectively the result or embodiment of its history. How can we coordinate these two aspects?

When Ashby and Maturana emphasize that living beings are machines, what they mean is that the structural states that explain the behavior of a living being at a given moment are just those which are present at that moment, irrespective of the way in which they have come about. A dog, to return to Ashby's example, runs away when the noise of a car engine is heard. Its owner explains "He was run over by a car six months ago". With this comment we understand that the particular behavior of the dog is not an innate response but a (dramatically) learnt response; i.e., the product of a particular history of structural coupling. What happened six months ago has, to some extent, changed or rewired the structure of the dog's nervous system in such a way that now the noise of a car engine is an external factor that triggers this particular sensorimotor response. The dog's history, thus, is crucial to understanding its behavior. Nonetheless, we also know that the dog's behavior in the present is a direct function of the structure of its nervous system in the present, not of his history. If we build, molecule by molecule, an instant copy of the dog, its response to the engine noise will be exactly the same, no matter the total absence of a history of structural coupling.

Machines, living or not, always operate in their structural present, and their history, when there is one, is just that, history. It is the observer, not the machine, who can take the machine's history and put it in relation to the observed behavior.

> [H]istory becomes embodied both in the structure of the living system and in the structure of the medium, even though both systems necessarily, as structure-determined systems, always operate in the present through locally determined processes. Therefore, although from the cognitive point of view adequate behavior as a case of adaptation cannot be understood without reference to history and context, from the operational point of view adequate

> behavior is only an expression of a structural matching in the present between organism and medium, in which history does not participate as an operative component. History is necessary to explain how a given system or phenomenon came to be, but it does not participate in the explanation of the operation of the system or phenomenon in the present. (Maturana, 1978a, p. 39)

'Logical accounting', in Maturana's jargon (Maturana and Varela, 1987), means keeping the structural plane of the operational effectiveness of the observed system separate from the descriptive/explanatory plane belonging to the observer. When we keep these two planes separate, we can recognize their respective validity without confusing them.

We do not need to deny that the structural present of the system is the result or embodiment of its history of structural coupling, and that in said history the participation of the environment has been crucial. On the other hand, we do not need to deny that even when the structural present of the system is the result of its particular history of structural coupling, the system as such only operates according to its structural present, regardless its history of structural coupling. The living being, as a machine, does not 'take into account' its history of structural coupling in its behavior; it rather finds itself, moment after moment, with a given structural configuration, and reacts or operates from there. It is the observer who can, and many times needs to, in order to explain the living beings' behavior, take into account the history of its structural coupling.

When the observer misses this point, as Ashby's incomplete observer did in the case of the dog (see section 2.1.1), she puts herself into a trap. She puts herself before the pseudo-problem of explaining how the organism, or its nervous system, handles or manipulates something that happened six months ago, something that belongs to the past. She raises the pseudo-problem of explaining how the organism's nervous system, as Clark (2001) says, 'makes contact' with an absent state of affairs. The trap is that, since this 'contact' cannot be explained in physical or structural terms, the observer feels the need to appeal to non-structural (i.e., epistemic, semantic,

intentional) factors or dimensions. The observer, thus, ends by building what Clark calls a 'representation-hungry problem' (Clark, 2001). This 'representation-hungry problem', according to all we have said in this thesis about the functioning of living beings as structurally determined machines, is in reality a pseudo-problem; a problem that only arises thanks to the observer's bad logical accounting.

The same runs for the topic of perception. If we keep our logical accounting clear, we can affirm the structurally determined character of the sensory system without denying the existence of external factors impinging on its structural dynamic. We can recognize that such and such structural state in the sensory system is consistently triggered by such and such environmental configuration, and at the same time understand that the sensory system, for the effects of its structural dynamic, never distinguishes or 'takes into account' the external origin of said structural state. (It is not a bad exercise, if the reader has the energy, to revisit the previous chapter and try to apply, to the case of the frog, the 'logical accounting' recommended by Maturana).

In general, we fail in our logical accounting every time we project into the living being features or elements that belong to our descriptive/explanatory domain. When we do this we end, sooner or later, by transgressing the boundaries of Strict Naturalism, positing non-structural (i.e., intentional, semantic, representational, epistemic) properties in living beings' internal dynamic.

Tolman's rat is another case in point. The notion of 'internal map' aims to explain what Tolman takes to be a sample of purposeful and intelligent behavior that involves some kind of memory mechanism or learning process; a process that Tolman understands in terms of 'internalization' of the environment. How does the rat 'internalize' and 'keep in its memory' the spatial layout of the maze? How does the rat manage to orient its behavior, in spite of obstacles, towards a future (i.e., absent) state of affairs (the presence of food at the exit of the maze)? Tolman faces, without doubt, a clear case of a 'representation-hungry problem'. Yet the problem, again, is just a pseudo-problem.

The rat's nervous system, as a structurally determined machine, does not

have any ‘internal map’ of the maze, nor does it ‘make contact’ with some absent state of affairs (its previous visits to the maze, or the future presence of food at the exit of the maze). The maze, as a spatial object that constitutes the ecological context of the rat, is something that Tolman clearly sees and distinguishes in his observation, but not, as we have argued before, something that appears for the rat’s nervous system. The rat’s nervous system, as a structurally determined system, only operates in the space of its own structural states, without distinguishing (let alone ‘internalizing’) the presence of an external maze.

At the same time, although the rat’s nervous system has no idea about the existence of the maze, it is a fact that its structural present is the embodiment of a history of structural coupling wherein the maze, as a recurrent source of perturbations, has played a key role. The rat’s nervous system has not internalized an image, map or model of the maze, but its present sensorimotor configurations have been partially shaped, in a non-trivial way, by the recurrent encounters with the maze. Since the structural present of the rat and its nervous system is the product of a structural drift in which the maze has participated as a modulator element, it is not a surprise that, when put again in the maze, the rat reacts and behaves in ways that express and manifest this structural familiarity. That is what we call ‘learning’.

We know, on the other hand, that the rat’s nervous system, as a biological machine, always operates from its structural present, without establishing any ‘contact’ with non-actual states of affairs. The temporal horizon, past and future, in which Tolman frames or contextualizes the present behavior of the rat, is a useful and legitimate descriptive dimension that belongs to his own observational domain, but not an operating element in the rat’s behavior as a biological machine. The rat’s history of structural coupling, including its previous encounters with the maze, has no operational presence in the generation of its behavior. This is so even when we know that, without that precise history of interactions with the maze, the rat’s nervous system would not have the structure and sensorimotor configurations that it has in the present, and would not be able to generate the kind of behavior that we appraise as intelligent.

Similarly, the coherences that Tolman sees between the present behavior

of the rat and some future state of affairs, and which he assumes reveal the presence of purposes or internal representations of goals, correspond to features and possibilities that he finds in his observational domain, not to operating elements in the structural dynamic of the rat. The goal-directed appearance of the rat's behavior, we know, is a function of its particular and complex stability dynamic.

Both the history and the ecological context are real aspects in the life of living beings, and we observers can, and need to, take them into account if we want to explain and understand their behavior. However, a certain discipline and epistemological vigilance is required. It is important, as Maturana says, to keep good logical accounting and remind ourselves that, although both history and ecological context are available for us as explanatory resources, none of them exists or has operational presence in the internal structural dynamic that generates living beings' behavior.

Chapter 6

Social phenomena, communicative behaviors and language: the social origins of mind

What is the origin of mental phenomena? How do they appear in the natural world? The final chapter of this thesis explores these questions and tries to answer them by appealing to a sociolinguistic theory of mind. It is argued that some of the essential marks of the mental, such as representational content and intentionality, have their origin in language.

In previous chapters we have reviewed the biological roots of animal cognition and showed that behind the intelligent and adaptive behavior of living beings, all that we find is a deterministic process of structural drift; a process wherein notions such as information transferring, intentionality and internal representations do not have a place. In general, and in a way that is relevant for our purposes here, we have argued that such notions, when applied to the internal structural dynamic of living beings, are nothing more than projections or logical accounting errors introduced by the observer.

This non-representational and non-intentional picture of living beings leaves us, nonetheless, in front of a big question. If the intelligent behavior of living beings is just a matter of structural drift, if there are no such things as internal (neural) representations, where do mental representations come from? If there is nothing within the anatomical confines of the organism, nothing in its nervous system or brain that could be considered an intentional phenomenon, where do the intentional phenomena come from?

So far we have analyzed the behavior of living beings considering them as individual unities in interaction with their medium. Now it is time to expand the focus. In order to explore the origins of mental representations and their intentional properties we have to turn our attention to what happens when two or more living systems start to interact in a recurrent manner constituting new domains of structural coupling. In this chapter we are going to review the emergence of social phenomena and the nature of the communicative dynamics established by living beings. The purpose is to understand the peculiarities of language as a communicative system and the special kind of phenomena that it inaugurates in the natural world. The central hypothesis is that the emergence of language and the emergence of the intentional-representational ability in living systems is, essentially, one and the same phenomenon. A complementary hypothesis is that the mental experience emerges as a transformation or metamorphosis of language.

In Chapter 1 we said that only creatures capable of language operate as observers, and that many of the explanatory elements traditionally used in relation to living beings' behavior (intentionality, teleology, representation, control, normativity, agency) are nothing more that anthropomorphic projections or logical accounting errors introduced by the observer. Since we have identified the ability to observe with the ability to operate in language, this Chapter, in offering an explanatory hypothesis about language, offers at the same time an explanatory hypothesis about the observer. This Chapter may be viewed, thus, as an attempt to explain both the emergence of mental phenomena and the constitution of the observer.

6.1 Origins and minds

To say anything about the origins of something we need first to have at hand a more or less clear idea about that something whose origins we want to identify, and also a more or less clear idea about what it means to identify the origins of something. In our case, we need to agree to some criteria by which we can identify certain phenomena as being mental phenomena, and

also some criteria for deciding, after attempting a theory, whether or not we have successfully reconstructed the origins of such phenomena.

So the first thing that we have to do in this chapter is to offer some criteria of individuation for mental phenomena. There are several criteria that philosophers use for this purpose, and in this chapter certainly we could not embrace all of them with the same detail. So I propose to take a more or less standard approach and to consider some traditional "marks" of the mental.

We are asking for the origins of mind. But, what kind of "things" are minds? What is it to have a mind? We say, for example, that while John has a mind, the stone that he has in his hand does not have a mind. When we say that, what do we mean? The expression "having a mind" sounds like "having a heart", "having a liver", "having a brain", etc. This kind of expression tends to promote a substantival view of mentality. The substantival view supposes that mind is more or less like an object in the world (like a car, a tree, a chair, etc.), a kind of substance with particular properties (e.g., being immaterial). Although popular some centuries ago, today almost nobody endorses this view. The reason is that the noun "mind" seems to be nothing more than the product of a linguistic turn known as nominalization. This linguistic move transforms verbs (actions) or adjectives (properties, attributes) into nouns (things, objects). For example, if a horse runs fast and gracefully, we say that the horse "has a good gallop". If an athlete, a runner, is able to accelerate her velocity in a short period of time, we say that she "has a good acceleration". The same with some adjectives or attributes: if a metal is hard, we can talk about "the hardness of the metal". Interestingly, we can do the same with some relational attributes. For example, if two things are different, we can talk about "the difference between them". We will see that the noun "mind" constitutes a general nominalization of certain attributes.

To reject this substantival view is not to deny that minds are real, but only that they are substances. The idea is that expressions like "he has a great mind", "he lost his mind", "he is out of his mind", being nominalizations, should not be taken literally in the sense that there are objects in the world called "minds" that people "have", "lose", or "are out

of" (Kim, 2011).

But if mind is not a substance, what is it? We could say that mind is basically an *attribute* of certain systems. When we say that an entity has a mind we mean that such entity has the *property* of undergoing certain *states* (uncertainty, wonder, frustration) or that *is able to do* certain things (believe, think, remember). "Having a mind" may be understood as having a set of mental capacities, like when we say that a horse has a good gallop or that an athlete has a good acceleration (Kim, 2011). Gallop and acceleration are not things or organs inside the horse or the athlete, though both the horse and the athlete have organs that, working in a certain way, participate in the generation of the movements that we call gallop and acceleration respectively. In the same way, mind is not a thing or an organ inside the persons. This is so even when we know that there are certain organs inside the persons (e.g., the brain) whose working is essential for generating the kind of phenomena that we identify as 'mental' (Ryle, 1949).

In order to avoid falling into the trap of the substantival mind, every time we use the term "mind" the reader should translate it immediately into "the attribute of undergoing or performing mental phenomena". But, what kind of phenomena are mental phenomena? Is there any "mark" of the mental?

First, from an epistemic point of view, it is usually claimed that mental phenomena (states and events) are special because we have a privileged access to them from a first person perspective. This epistemic mark is articulated in different ways and degrees, and philosophers talk about immediacy of knowledge, privacy or first person privilege, transparency, infallibility and similar notions. But in general, and putting aside more complex technicalities, the idea is that the content of our mental states is private in the sense that, for example, we can entertain a desire that, at least in principle, if we decide never to communicate it, may remain forever unknown to anyone else. Suppose that a woman has a secret desire that she does not want to confess. She is aware of this desire and knows perfectly well that it is not socially accepted. Suppose she maintains the secret until her death. What has happened with the desire and its content? The desire was real and was well known by the woman, but nobody else knew about it and nobody will ever know about it either. The content of that mental state

was, and will remain forever, private for the knowledge of the rest of the people.

We often take this privacy as if it were related to some physical specification. For example, we tend to think that this privacy has to do with the fact that our thoughts are located inside our skull, and that the other people cannot "see" through our cranium and "look at" them. Phenomenologically speaking, from the point of view of our subjective experience, when I think something without saying it loud, almost unavoidably the feeling is that the thought remains in a certain way "inside" me (e.g., in my head). If I decide to communicate the idea, to make it public, the feeling is that the idea is "going out" to the external world (e.g., through the mouth). Roughly stated, the intuition is that my thoughts remain private as long as they remain *internal*, while they become public as long as they are externalised.

Following this experiential datum, mental phenomena are traditionally conceived as something essentially *internal*. Nonetheless, this inner character should not be identified with any spatial determination. Philosophers do not (or should not) interpret the inner character of mind in a literal concrete sense, but rather in an abstract metaphorical sense. We say that mental phenomena are inner just in terms of *epistemic accessibility*. A mental phenomenon is inside my mind in the sense that I, and only I, can access it by means of an introspective examination. Other people can, of course, have access to my mental contents too, *but not by means of an introspective examination*. They have to infer my mental states by observing my behavior, or they have to trust in me when I declare that I am in such and such mental state (they have to believe in my report). If they perform an introspective examination, what they are going to find is not my mind but their own mental experience!

As van Gelder (2005a) says, my mind is something internal in the sense that its contents are always inside a *boundary of epistemic accessibility* that excludes all observers but one; I. Thus, in order to avoid possible misunderstandings the reader has to recall that the inner character of the mental concerns an epistemic space, not a physical one. Mental phenomena are essentially intra-epistemic (epistemically private), not essentially

intracranial.

Second, it is widely agreed that most of mental phenomena exhibit intentionality and representational content (Crane, 2001, 1998; Dretske, 1995). Technically, intentionality and representational content are not exactly the same, but for our present purpose we can treat them as two aspects of the same thing.

That a mental state is intentional means, basically, that it *refers* to something or that it is *about* something. Intentionality here does not have to do with the psychological distinction between intended (voluntary, purposive) and unintended (involuntary, non-purposive) actions. An event or act is mental not because it is, or it seems, oriented to some purpose or goal, but because it is *about* something (Brentano, 1874/1973). For example, if I believe that Santa Claus does not exist, my belief is *about* the existence of Santa Claus. If I think that Martin Scorsese is a good director, the *contents* of my thought are the qualities I attribute to Martin Scorsese as a film director. Every mental state seems to be *directed* to something; something that is the object of the intentional act (i.e. the object of my happiness, the object of my belief, the object of my desire, etc.). This directionality, to be clear, is not the conventional directionality that we distinguish in the physical world. We use the term "directionality" in a quite abstract and metaphorical sense, and it is not an uncommon error to forget this point (even among philosophers). This directionality is neither spatial nor temporal; it is not a property for which we can specify values in terms of space-time. If I think of the Middle East, my thought is not pointing, like a compass, toward the East. If I remember that yesterday was a sunny day, my mental state is not pointing, like a clock in reverse, to the past. The same runs for the present and the future as temporal categories of our experience. We say that mental states are directed rather in a *semantic* sense; that they *make reference* to some intentional object (that may or may not have a real correlate in the world). Conversely, if an event or phenomenon is directed to some spatial location (e.g., the clouds are going toward the coast, the dog is going to the door) that does not make it a mental phenomenon. If an event or phenomenon exhibits certain temporal directionality (e.g., this chemical reaction is irreversible, the system tends to the increase of entropy) that does not make it a mental phenomenon. It is

important to keep in mind these distinctions and to avoid identifying the intentionality of mental acts either with purposive (intended) actions or spatiotemporal directionalities.[9]

Another way of putting this is to say that mental states exhibit a certain content; a content that, again, is not physical (like the water *contained* in a glass) but rather *referential* or *semantic*. This is indeed what people understand as the *representational content* of mental states. Typically, mental events are viewed as phenomena that, accurately or inaccurately, correctly or wrongly, represent something. The specific nature or format of this representational function may be, sometimes, an object of debate among philosophers. For example, some people think that mental representations are symbolic (Fodor, 1975; Pylyshyn, 1984; Fodor and Pylyshyn, 1988) while others think that they are subsymbolic or based on a different architecture (e.g. connectionist patterns; Clark, 1989; Smolensky, 1987). Some people think that the content of mental phenomena is exclusively conceptual, while others think that there are also non conceptual contents (Peacocke, 1998, 1992; Crane, 1992). Some philosophers think that the content of the representations is determined by causal-informational relations (e.g., Fodor, 1981; Dretske, 1981), while others think that what matters is the purpose for which such representations are recruited (teleological theories; Millikan, 1993, 1984). These are just some of the many discussions about the specific nature of mental representations. What is almost universally accorded, nevertheless, is that mental phenomena are representational phenomena. But just what is a representation? When can we say that such and such event is representing, or is a representation of, something else?

There is a long discussion about what exactly a representation is and what its ontological properties are. Nonetheless, a more or less accepted view says that every representational phenomenon must exhibit at least three elements: (X) that which is represented (the 'object', concrete or abstract, real or fictitious, of the representation), (Y) that which is taken or interpreted as

[9] The philosophical notion of intentionality is a bit more complex than as presented here. There are additional features associated with intentionality such as 'aspectual shape' (Crane, 2001) or the capacity to 'misrepresent' (Dretske, 1995). Here, nonetheless, 'aboutness' and 'semantic directedness' are enough for our purposes.

standing for (X) (the vehicle of the representation), and (Z) that which takes, uses or interprets (Y) as standing for (X) (the user of the representation) (Von Eckardt, 1995; Menary, 2007; Bechtel, 1998). This triadic composition constitutes basically a relational phenomenon. That is, a phenomenon that only exists as long as this triadic relation takes place (i.e., no one of the elements can be absent) or, equivalently, as long as these three components X, Y, and Z are effectively present and related in the adequate way. One could say, in the broadest sense, that something Y represents something else X, if there is something Z for which such Y designates, stands for, or means X.

The important point is that, although these elements are all equally indispensable, only one of them is that one that defines the triadic relation as a representational relation: Z, the user. The dyadic relationship between X and Y, considered in itself, is never a representational relation. It is Z that establishes the representational relation between X and Y, that "takes" Y *as standing for* X and that constitutes Y as a representational vehicle. Y, by itself, is never a representational vehicle; it becomes one as soon as some Z uses it to designate some X.

For example, the smoke is caused by the fire. The relation between these events is purely causal. Nonetheless, if some Z takes the smoke as an indicator of fire, then, and only then, the smoke emerges as a representational vehicle (an index) and the fire as an object of representation (an X). Previous to the constitution of the triad (previous to the incorporation of Z), smoke and fire are just two physical events causally linked, nothing more. The same occurs when the shape of a cloud resembles the shape of, say, an animal. The cloud, by itself, is not a representation of any animal; it becomes such (an icon) when it evokes in some Z the idea of an animal.

The general idea is that a representation is a triadic relation that is constituted when a Z takes something (Y) as standing for something else (X).

So if mental phenomena are representational phenomena, we should be able to identify in them, with certain clarity, the triadic structure aforementioned. And especially, due to its important constituent role, we should be able to distinguish with precision what in the triad works as Z, the

representational user.

As it is easy to note, intentionality and representation are intimately linked. After all, every representation is always *about* something that is being represented, and the referential relation entailed in intentionality is, certainly, a *semantic like* relation. Some philosophers recognize this fact by using two different senses of intentionality: *referential* intentionality and *content* intentionality (Kim, 2011). The first one concerns the *aboutness* of our mental states and the second one emphasizes the fact that our mental states have *meanings* or *semantic contents*.

Intentionality and representation are not bad candidates for individuating mental phenomena because, apart from language, no other type of phenomenon (state or event) seems to possess such properties *as intrinsic aspects*. Physical states are what they are without being themselves, in the technical philosophical sense aforementioned, *about* other things (Brentano, 1874/1973). Physical states do not relate in referential or semantic terms; they just connect in causal terms. If the ambient temperature goes down to minus 3 C° degrees, water will pass from a liquid state to a solid state. We can appreciate a causal connection between these events, but neither is the ambient temperature about the water nor the solid state of water about the ambient temperature. The ambient temperature is not the referential content of the solid state of water and the solid state of water is not the referential content of the ambient temperature. They, as physical events, simply do not entertain semantic contents at all.

Another completely different thing is that I, as a representing system, may observe the solid state of water and take it as an indicator, a sign, or evidence that allows me to infer other facts, as for example, that the ambient temperature must have descended at least to zero degrees. But the one who establishes the referential connection here is me, not the physical states. The freezing of water is not, by itself, a representation of the ambient temperature; it is just one of the many natural effects produced by the falling of the ambient temperature. Yet I can treat the causal connection as a semantic relation and say "The water is frozen. That *means* that the ambient temperature must have fallen down at least to zero degrees". I can interpret a causal relation as a meaning relation, certainly. And why can I do that? The

answer has to do with the very topic of our enquiry. I can do that because I have a mind, and having a mind is, at least for the most part, just to be able to assign certain meaning to things, events or phenomena.

If the ideas of intentionality and representation are clear enough, then we could take them as criteria for individuating mental phenomena. That is, we could say that identifying the origins of mental phenomena consists in identifying the origins of intentionality and representational phenomena. The question could be formulated as: How do intentional relations emerge in a world of causal relations? How do representational phenomena emerge in the physical world?

Our second requirement is about the notion of origin. When and under what circumstances can we say that we have exposed the origins of something? I propose as a criterion the following idea: to identify the origins of a certain phenomenon is to expose the conditions under which the phenomenon in question appears as a natural result. That is, one has to show those conditions that, once given, bring as a result the phenomenon under consideration. For example, if I ask "How do earthquakes originate?" a proper answer has to give the geomechanic and energetic conditions under which earthquakes appear as a natural result. A stronger formulation of this idea would say that one has exposed the origins of a phenomenon when one has given the necessary and sufficient conditions for the occurrence of such phenomena.

In our case, identifying the origins of mental phenomena would be equivalent to exposing the necessary and sufficient conditions for the occurrence of intentional and representational phenomena.

In what follows, I shall try to offer a hypothesis about what has to happen in the natural world such that the result is the emergence of intentional and representational phenomena. No doubt, here I will not commit myself to giving the exhaustive set of necessary and sufficient conditions for the emergence of mind. My purpose is naturally much more modest. I just want to suggest some preliminary ideas for orienting further explorations on the origins of mind.

The structure of the hypothesis is as follows. I consider the social dynamics established by certain living beings and fix the attention on

communicative behaviors. I try to show that language is a special kind of communicative behavior; a recursive linguistic behavior. I do this in order to show that it is the recursive nature of language (understood in a cybernetic sense rather than in the classic Chomskyan generative sense) which inaugurates the semantic and representational phenomena in the natural world. Finally, I offer some general ideas about the way in which language, originally social, becomes an individual and more or less private experiential domain, constituting properly what we usually call *our* mind.

6.2 Social coupling and communicative behaviors

That language is, or works as, a representational system is something that cognitive scientists and philosophers hardly could deny. Language is typically considered, with good reason, as the paradigmatic example of every semantic or referential system. Using language we make *reference* to different aspects of our experience, we *represent* the world as being in such and such way, we talk *about* different things, we denote, designate, mean, etc.

I will assume that the reader, like the cognitive experts, really does not need to be persuaded about this point. The semantic and representational power of language appears as something more or less evident in our quotidian life. Our task, accordingly, is not to discuss whether language works or not as a semantic system, but try to understand how such a semantic system appears as a natural phenomenon in the animal realm.

Now, since mind is basically a representational system, it is not surprising that many philosophers have taken language as *the* model for understanding mental activity. A traditional line in philosophy of mind, championed by authors like Fodor, has conceived the mind basically as a *language like* system. And I would say that this intuition is essentially correct, insofar as it detects the strong ontological continuity that exists between language and mind. Nonetheless, the way in which these theories

(e.g., Fodor's LOT, 1975) develop this intuition seems to me, according to the autopoietic framework here defended, basically wrong. Although here we will not set an open debate against this kind of theory, the reader will easily appreciate the points of disagreement. We will see that: 1) taking as a constitutive rule what is just a feature of our quotidian psychological experience, these theories assume that mental representations are ontologically prior to linguistic representations. The autopoietic hypothesis, on the contrary, assumes that linguistic representations are a precondition for mental representations; 2) taking as a spatiotemporal specification what is just an epistemic distinction, these theories assume that mental activity takes place inside the head, i.e., in the brain. As opposed to this, the autopoietic hypothesis assumes that the mind is not in the head but in the recursive communicative dynamics of linguistic organisms. We will review these points soon.

Before starting our analysis, some terminological specifications are needed. In the autopoietic theory, the structural coupling that a first-order autopoietic unity (a unicellular organism) holds with its environment is called first-order coupling (e.g., a bacterium in its aquatic medium). The structural coupling that a first-order autopoietic unity holds with one or more first-order autopoietic unities is called second-order coupling (e.g., a colony of bacteria). This second-order coupling may lead sometimes, although not always, to the constitution of second-order autopoietic unities. That is, to the constitution of multicellular organisms (e.g., hydras, ants, mice, gorillas). Finally, the structural coupling that a second-order autopoietic unity holds with one or more second-order autopoietic unities is called third-order coupling. For example, a colony of insects, a pack of wolves, a community of gorillas, are all systems constituted by third-order couplings. Now, a colony of insects (termites, bees, ants), a family of wolves, a community of gorillas, are also typical examples of social systems. The autopoietic theory, accordingly, defines social systems as third-order coupling systems, or, equivalently, as third-order biological unities (Maturana and Varela, 1987).

As a general rule, in a social system organisms coexist and communicate modulating their behaviors reciprocally. Communication, from this point of view, is basically a process in which two or more organisms coordinate their

behaviors in a mutual way. Yet this behavioral coordination, according to all we have said before, cannot be understood as the result of information transmission processes or transference of semantic contents (Maturana, 1978a). We face here another dangerous bend.

We tend to associate the communicative phenomenon with a process in which "something" (a message, a content, a piece of information) "travels" from one place (the emitting source) to another (the receptor) through a kind of "channel" or conduit. This is a conventional metaphor that, though familiar to us, does not illuminate the biological phenomenon that we are addressing here. We must recall that communication, understood as an interactional phenomenon that takes place between second-order autopoietic systems, is just a particular version of structural co-drift; a process of structural change wherein, as we have seen before, there is no room for informational or semantic contents. Thus, in order to avoid misunderstandings in what follows, the reader should try to put aside, at least for a while, the familiar and traditional conceptions about communication. The idea is to keep in mind that when we talk about communication, we are talking about third-order structural coupling processes between living systems, nothing more. Nonetheless, it is better to make it explicit. Do I mean communication without information? Yes, I do. Do I mean communication without any semantic content or message to transmit? That is exactly what I mean.

6.3 Communication and recursion

A wide spectrum of living beings exhibit one or another way of communication or behavioral coordination. The concrete mechanism of structural coupling used by the organisms varies according to the different species. For example, the majority of the so called social insects coordinate their behavior through the interchange of chemical substances (e.g., throphallaxis in ants), but others prefer to use patterns of sound (e.g., crickets), and there are even insects that seem to "dance" and coordinate

their behavior through certain specific patterns of movements (the classical example of bee dances). Birds and whales elaborate quite sophisticated songs, and superior mammals exhibit very complex communicative patterns mixing all kind of sounds and movements.

The way in which organisms coordinate their behaviors may be strongly specified by their phylogenic history, or may be the result of certain particular ontogenic co-drift. In the first case we assume that the communicative patterns are basically innate, while in the second case we say that the organisms have developed or acquired (learned) certain communicative patterns as a result of their particular history of structural coupling. In many cases the communicative behavior of the organisms is certainly a mix of these two conditions.

Whatever the case, innate or learned, we tend to treat these communicative behaviors as semantic interactions. That is, as if the course of the interactions was determined by the meaning of certain "messages" and not by the dynamic of structural coupling of the interacting organisms. Just like in the fables, where animals speak and act in a human like way, we usually describe animal communication in semantic and intentional terms. These semantic descriptions, according to what we already know about structural coupling processes, are basically incorrect, yet they have a well founded origin: the fact that *animal communication constitutes the prelude of human language*. Human language and animal communication are indeed very close relatives, and in that sense, it is not an arbitrary error to treat animal communication as if it were a semantic phenomenon. The point is that, though somehow justified, an error is always an error.

Now, if animal communication and human language are close relatives, where is the difference between them? We have said that certain organisms communicate in a very sophisticated way; why do not such communicative behaviors (e.g., the "language" of bees, the "language" of birds, the "songs" of whales) count as genuine language? Many linguists and philosophers have tried to provide one or another demarcation criterion with respect to this point. The autopoietic theory offers its own hypothesis: while communication is behavioral coordination, language is basically behavioral coordination of behavioral coordination, or, what is the same, *recursion* in

the domain of communicative behaviors (Maturana, 2000, 1978a). What does this mean?

To understand the peculiarity of language as a communicative system we have to understand in a proper way the notion of recursion. The cybernetic notion of recursion used by Maturana has little to do with the syntactic notion of recursion used in linguistic theory. In linguistic theory (especially in Chomskyan generative grammar) it is said that a distinctive property of human language is its recursive power. That is, the ability of producing an infinite number of sentences by inserting new phrases (clauses) once and again within the same sentence. For example, if we have the sentence 'Mary met a new boyfriend', we can insert the sentence 'Mary went to Italy' to form a new one: 'Mary, who went to Italy, met a new boyfriend'. We can go on inserting, for example, the sentence 'Mary's mother was born in Italy' and obtain: 'Mary, who went to Italy where her mother was borne, met a new boyfriend'. And so on and so forth. Recursion here is appealed to for accounting for the productivity of human language, which is usually viewed by the majority of linguists (not all of them) as one of its distinctive marks.

Maturana's cybernetic notion of recursion, by contrast, does not point to any property of language (i.e., productivity, compositionality) but to the process by which language emerges as a special kind of communicative pattern. Recursion here means the recurrence of an operation upon its own result. The idea is easy to see if we differentiate between recursion and repetition. For example, in the operation '$\sqrt{a^2} = a$' we apply an operator ($\sqrt{}$ square root) upon the operand (a^2) and we obtain the result (a). To repeat this operation is simply to replicate it:

$$\sqrt{a^2} = a$$
$$\sqrt{a^2} = a$$
$$\sqrt{a^2} = a$$

In a recursive process, instead, we have to incorporate the result of the operation as the operand of the next operation:

$$\sqrt{a^2} = a$$

$$\sqrt{a} = a^{½}$$
$$\sqrt{a^{½}} = a^{¼}$$

In a process of repetition there is no possibility of obtaining any new result. In a process of recursion, on the contrary, there exists the possibility of obtaining something new. In a sequence of repetition there is no historical connection among the operations. In a recursive sequence the operations are connected because they take as their operand the result of the previous operation. How novel may the results of this kind of recursive processes be? That depends on the particular domain in which the recursion takes place. Sometimes the novelty may be trivial or merely quantitative (in our example, 'a' and '$a^{½}$' are different though not in an interesting way), and sometimes it can be notorious and qualitative (see some illustrations below). What is important to clarify, in order to avoid misunderstandings, is that this recursive novelty is not the recursive novelty of the linguistic productivity. The productive recursion makes a proposition grow only by inserting new elements in the same chain, say, always in a horizontal sense.

To take our previous linguistic example, a cybernetic recursion of the sentence 'Mary met a new boyfriend' would be equivalent to applying to the sentence the same referential function that the sentence applies to a certain state of affairs in the world. The very sentence should be taken as the object of a new referential function. For example: 'The sentence "Mary met a new boyfriend" contains two nouns and one verb'. In this case what we see is the emergence of a second-order referential level (a metalanguage) wherein the original sentence assumes the role of what philosophers call 'object language'. This recursive transformation is qualitatively different from the productive syntactic one. The sentences 'Mary met a new boyfriend' and 'Mary, who went to Italy, met a new boyfriend', though different, both remain in the same referential level (language→world). Yet the proposition 'The sentence "Mary met a new boyfriend" contains two nouns and one verb' is not only different from the sentence 'Mary met a new boyfriend', but also operates in a different referential level (metalanguage→language).

Now, the notion of recursion that Maturana has in mind is not restricted to this kind of formal domain. In natural contexts some recursive processes

may exhibit interesting results. A quotidian illustration is to take a mobile mechanical toy in our hands, say a little robot, and turn it on. The robot starts to move its legs, but if we keep it suspended in the air, all that we see is a mere repetition of movements. In contrast, if we put the robot in contact with the ground, those very movements start to operate in a recursive manner. Every movement is applied over the result of the previous movement giving place to a historical sequence of displacements that we call "steps", and whose overall outcome is the "walking" of the robot (Maturana and Pöersken, 2004).

Interestingly, in this example we note that the difference between repetition and recursion does not lie in the system (the robot) *per se*. Either suspended in the air or in contact with the ground, the robot performs basically the same movements using the same physical engine. Nonetheless, the results in one case and the other are qualitatively different. The difference does not lie in the structural composition of the robot. The difference does not lie in the pattern of movements either. What has changed is *the way in which the movements relate*. What is new is the *recursive* relation established between them.

Let us review this example more in detail. It will help us to address the more abstract issue about the emergence of language and mind.

Is the walking robot the same as the robot who moves its legs in the air? In one sense, yes, it is the same. We disassemble the toy and we cannot find anything new in its internal composition. In another sense, nonetheless, it is different. The walking robot is certainly *doing* something new. Yet, where does this new activity take place? Does it take place in some component of the robot, some internal engine, some structure? No. The walking takes place in the relation between the robot and the ground, and that which walks is the robot as a whole, not some of its parts. Does this mean that the physical structure of the robot is irrelevant? No, it does not. If the robot loses one of its legs it cannot walk. If the robot loses its motor mechanism it cannot walk either. Does that mean that the "centre" of the walking function is located in the leg, or in the internal motor of the robot? No, it does not. Walking is an activity that takes place in the behavioral domain of the robot, not in its internal dynamic (although such internal dynamic is crucial for generating

the pattern of movements involved in that walking).

What is important is to note that the robot cannot bring about the activity of walking by itself. It needs a ground, a certain surface to transform its repetitive movements into a recursive walking. Without interaction, without establishing certain relations, the system cannot create the recursive phenomenon. At the same time, once the robot is put in contact with the ground, the recursive phenomenon of walking emerges as a necessary outcome. The interaction between the movements of the robot and the surface of the ground cannot but produce the phenomenon of walking. Strictly speaking, there are no alternative possibilities. We also know that there are no purposes in this phenomenon either. The robot does not have any goal (it does not "want" to walk). The recursive phenomenon takes place simply as a natural result; as a deterministic phenomenon.

Nonetheless, we have to consider that if the structural conditions of the robot and the surface were different, the recursive phenomenon might be less neat and obvious. What if there is a robot that has weak legs made of thin and feeble pieces of wood? What if its feet are too heavy (made of granite)? Could this robot walk? How far? Or, what if we put our robot in contact with a different surface? What if we put the robot in contact with a layer of thick honey or a gelatinous surface? Could it still walk? How far? We can imagine different situations in which the phenomenon of recursion might appear attenuated or just partially realized. We can imagine different situations in front of which we could not estimate with certainty the status of the robot: "is it really walking or not?"

Whatever the case, when the robot is effectively walking, the robot *is* a new system. Not in terms of its physical constitution, but in terms of its doing. The walking robot opens a new range of phenomena because there is a set of events that can happen only when one is walking (bump into things, fall, get stuck), things that could not happen if the robot remained moving its legs in the air. The walking robot and the non-walking robot exist in different behavioral domains.

Among the things that may happen only when the robot is walking, there is one that is especially interesting for our analysis: to leave a footprint on the soil. If the ground is covered with a fine layer of dust, we will see that

the robot leaves a trace. On walking, the robot draws a path; a path that does not pre-date the walking of the robot, but that is created by it. The robot, of course, does not have any purpose in doing that; it does not have as a goal to draw any path. The path is just the result of the structural coupling between the robot and the land. What is more, the robot cannot avoid leaving that trace. Given its structural constitution and the structural constitution of the dusty ground, the marks are unavoidable and the final result will be a path drawn on the land.

6.4 Language, representation and original intentionality

Having all these elements in mind, we can come back now to the autopoietic hypothesis about language. We have said that language emerges basically as a recursive phenomenon in the domain of communicative behaviors among living beings, and we have previously defined communication as behavioral coordination. In other words, the autopoietic hypothesis contends that language arises when a group of organisms start to coordinate their behavioral coordination establishing a second-order communicative pattern (Maturana, 1978a). As in the case of the robot, this process of recursion is nothing mysterious (recursion is a natural outcome under certain conditions), but we certainly can expect some qualitatively novel results. The autopoietic hypothesis contends that when organisms start to operate recursively upon their communicative behaviors: a) the semantic-intentional domain arises as a new domain of social coupling, and b) organisms acquire the social ability to represent or establish referential relations between distinct aspects of their experience. In other words, the organisms become interpreters or "Zs" (the key element in our representational triad "X – Y – Z"). They start to "walk", like the toy robot, but instead of drawing a path on the land, they start to create a world of shared (socially constructed) meanings.

We have said in previous sections that a representation is always a triadic relation wherein something (Y) is interpreted by some entity (Z) as standing

for something else (X). We have also said that in this structure the "semantic motor" is Z; that which is able to establish semantic relations between non-semantic things or events. At the same time, we have reviewed the question about the ontological status of Z: What exactly is Z? What kind of thing or system in the world may be considered a Z? Our answer now, after having discarded some popular candidates for assuming the role of Z (e.g., the nervous system and the brain), is that Z is a system that operates recursively in the domain of communicative behaviors; Z is essentially a linguistic system. To say this is tantamount to saying that the emergence of the representational phenomena in the natural world coincides with the emergence of language. Let us examine this idea.

We have said that when the organisms start to operate in language, they become Zs or interpreters. What does that mean? Does it mean that they acquire any new structural component, some new organ in their body? No, it does not. Like the toy robot, the difference is not about any physical constitution but about a *new doing*, a new behavioral domain. It is the new recursive character of the social communicative relations that makes the difference. Does this mean that the structural constitution of the organism does not matter, that it is an irrelevant point? No, it does not. We know that if we remove the brain from the organism, it will not be able to participate in any dynamic of social coupling, recursive or not (assuming, for the sake of the philosophical argument, that the organism could remain alive without its brain). Certainly, it will not be able to develop the ability to use language at all. Does this mean then that the "centre" of language is in the brain? Does this mean that there is some region in the brain performing or "controlling" the key recursive mechanism involved in language? No, it does not. We saw in our example that the recursive phenomenon of 'walking" took place in the interaction between the robot and the ground, not in any of the robot's structures. The "walking" was not in its legs, though without its legs the robot could not walk at all. In the same way, we see that *language is a phenomenon that arises and takes place in the behavioral interactions among living beings*, not in their respective internal dynamics. This is the case even when we know perfectly well that these internal dynamics are crucial (as is dramatically revealed in many cases of brain damage) for

generating the behavioral patterns involved in such communicative interactions. But from the fact that a healthy nervous system is a necessary condition for the recursive phenomenon of language, it does not follow that the ontological place of such recursive phenomenon *is* in the brain.

Z is usually called "user" or "consumer" of representations, but now we start to see that these denominations probably are not the best ones. The notion of user or consumer presupposes the pre-existence of that which that is used or consumed. The table is there as something already fabricated, and I can use it in such and such way. The energy is there in the world, and we humans consume it in such and such way. Yet the action of Z does not look like the action of a consumer, but rather like the action of a creator, or, to be more precise, of a social co-creator. In the same way that the path was not there on the land waiting for the robot, the representations (representational vehicles or representational relations) are not there in the world waiting for consumers or users. It is not the case that Z encounters the meanings out there in the world, say, like the animal encounters fruits hanging from trees. Like the robot that creates a path in walking, so the representing organisms create meaning by interacting recursively in their social coupling dynamics.

Now, we should note that this creative (representing) activity is not an optional hobby for Z. Once the systems start to operate in language, once they become consolidated interpreters (Zs), they actually cannot avoid establishing semantic relations. We saw that the walking robot could not avoid leaving a trace on the land, that it could not avoid drawing a path. In the same way, representing systems cannot avoid interpreting the world, and everything that is contained in it, in a meaningful way. The 'objects' appear as linguistic distinctions within the experiential flow. The smoke is perceived as meaning "fire" or "danger", the clouds are perceived as meaning "rain" or something else. What is more, representing systems tend to make sense of natural phenomena in semantic and intentional terms. All of the mythological systems in primitive and ancient civilizations have to do with this projective tendency. Today few people interpret natural forces in semantic-intentional terms, but the tendency persists in the interpretation of animal communication: the dog is telling me "take me out"; the bee is reporting "nectar is 500 meters northeast"; the lion is roaring and saying "get

out of here!" and so on and so forth.

The overall result is that the whole experience becomes a semantic network of meanings. We said that when the robot starts to walk, it starts to exist in a "walkable" world. In the same way, when the organisms start to create meaning, they start to create and inhabit a meaningful world. They create a semantic niche.

We have to underline that this representational activity emerges as a particular way of social coupling. When the organisms start to operate recursively in the domain of their communicative behaviors, they start to constitute, *in a collective dynamic*, the meanings of such interactions (Maturana, 1978). We saw that the robot cannot create the recursive phenomenon by itself; that it needs to establish certain interaction with a proper surface. In the same way, no single individual organism can produce by itself (alone on an island) the recursive dynamics that generate the semantic-intentional domains. It is only as a participant in the domain of social interactions that an organism can become a Z. This is why the representational phenomena are, in their origin, social phenomena.

So finally, what exactly is a Z? What has changed in a system once it has become Z? We say that what has changed is its relational doing; its condition as a communicative system. A little piece of wood is nothing more than a piece of wood, but if we put it on the second row of a chessboard, it becomes a pawn. It is the logic of the game, and not some intrinsic property in the piece, that transforms a mere piece of wood into a pawn. In a similar way, an organism becomes a Z as soon as it starts to participate in (if the linguistic community is already constituted, as it has been for every one of us), or co-create with its peers (if the linguistic community does not exist yet, as was the case with primitive communities of hominids), the game of recursive communications that constitute language. Without participating in this particular game the organism cannot become a Z, no matter how healthy and powerful its brain may be. As some extraordinary cases of feral children have shown, a life outside of every linguistic community (for instance, in a wolf pack) produces organisms who are able to communicate in an effective way, but that cannot establish recursive relations upon such communicative interactions. To answer our question: a Z, a representing system, is a

communicative system that is able to establish recursive communications within a network of social couplings.

Although the emergence of the representational (semantic) phenomena associated with a community of Zs is something really novel and conspicuous, we have to keep in mind that, being a version of social coupling, language is, ultimately, just another form of coexistence among living beings. Language is a form of behavioral coordination through which certain living beings conserve their adaptation constituting a third-order biological unity. In other words, language is a biological phenomenon and cannot be conceived, in spite of its novel properties, as something separate (discontinuous) from life. Strictly speaking, nothing that living beings do is independent from their condition as living systems. To be clear, social phenomena, as long as they are realized through the behaviors of living beings, are biological phenomena. In that sense, as language is something that we, human living beings, do, language is a biological phenomenon too. Language and its semantic properties are not a biological anomaly but a natural phenomenon that takes place within certain dynamics of social coupling. Language does not emerge from nowhere, but from the communicative practices of social animals as a recursive phenomenon within such practices.

That this is the case seems more or less clear if we see the kind of communicative interaction that we can build with certain superior primates under specific training conditions. When higher primates interact closely with humans maintaining a recurrent communicative dynamic for a relatively long time, they are able to manage a certain kind of communicative behavior that, though very limited compared to human language, exhibits some recursive features (Gardner & Gardner, 1971; Savage-Rumbaugh, 1986). This is a more or less predictable result since, first, higher primates are our closest evolutionary relatives (we humans are in fact just a subclass of primates), and second, we know that the phenomenon of recursion may appear, under certain conditions, in attenuated or transient forms. In our previous toy robot example we saw that there were certain conditions, such as those brought about by changing the robustness of the robot's legs or the viscosity of the surface, under which the

phenomenon of recursion could appear in a very limited way. Certain higher primates exhibit, after a period of training, communicative behaviors that seem recursive. Nonetheless, like the robot that had to walk on a surface of sticky honey, these animals can perform only a limited range of recursive interactions and cannot develop, at any rate, a full and systematic linguistic system. But beyond these considerations, perhaps the key difference between the communicative achievements of these animals and we human beings is that they cannot generate, spontaneously and by themselves (by means of their own dynamics of social coupling), any recursive system of communications. They can participate in some limited recursive communicative interactions only as a result of a prolonged interaction with humans (humans specially dedicated to promote linguistic interactions with them). Without a human to constantly induce linguistic interactions, primates and animals in general do not develop recursive communications in their natural niches.

In our lives language is a powerful tool, yet as a natural phenomenon language arises in the evolutionary history without any specific purpose. Semantic and representational phenomena are part of language, but they are not its *raison d'être*. Mother Nature did not create language for endowing certain animals with representational and semantic capacities (such capacities were never necessary for the survival of any animal, social or not). In our previous toy robot example we saw that both the recursive process of walking and the path drawn by it emerged as natural results of the interaction between the movements of the robot and the ground. That is, as purposeless outcomes. In the same way, language emerges in the natural history of living beings as a peculiar version of coexistence and social coupling, without any evolutionary purpose.

6.5 Mind and language: a dialectical internalization

In the previous section, and having as a background the non-representational

autopoietic theory of cognition, we formulated the question about the origins of mental representations. We asked: if the intelligent behavior of living beings is just a matter of structural drift, if there are no such things as internal (neural) representations, where do mental representations come from? If there is nothing within the anatomical confines of the organism, nothing in its nervous system or brain that could be considered an intentional phenomenon, where do the intentional phenomena come from?

Our answer was: intentional and representational phenomena, the core features of mental activity, arise and come from language.

We have also said at the beginning of this chapter that to point out the origin of mental phenomena is tantamount to pointing out the origin of intentional and representational phenomena. Well, in the previous section we have just presented a socio-linguistic hypothesis about such an origin. If the reader has taken the hypothesis at least as a plausible one, then we have provided a plausible hypothesis about the origins of mind. But, have we really provided a reconstruction of the mind? It looks like we still have to face the following question: if language seems to have all that is essential for having a mind, namely intentional and representational properties, where is the difference between language and mind? Are representations and intentionality all that we need for having a mind?

Here is where the epistemic mark of the mental, its inner character, plays a relevant role. Certainly, language and mind share intentional and representational properties, but while language is a phenomenon that takes place in the public domain of social communicative behavior, mental activity takes place in the private domain of the individual life. Language is essentially an *inter*-personal dynamic, while mind is an *intra*-personal one. Mind, but not language, appears to us as something essentially *private.*

In this last section we are going to outline some general ideas about the process by which language, a social public phenomenon, may give rise to the inner domain of mental experience. Our guide here will be the Marxist psychology of Lev Vygotsky and its "sociogenetic view of human cognition" (Valsiner and Van der Veer, 1988), interpreted within the broader framework of the autopoietic theory.

How does the mind emerge from language? How does this process

occur? Our Vygotskyan hypothesis says that mind emerges as an "internalization" of language. To properly set this hypothesis we need to consider the following points with respect to the notion of internalization. First, we need to remind ourselves of the precise sense in which we say that the mind is something inner. Second, we need to take into account the Marxist framework of Vygotsky's theory and interpret the process of internalization as a *dialectical* phenomenon.

The reader must recall that the inner character of our minds concerns the distinctive form of epistemic access that every one of us has toward her or his own mental contents. When we say that language is internalized, we do not mean that language, coming from the outside, traverses the skull of persons and embeds itself in specific regions of their brain. We allude rather to the distinction between the public and the private domains as spheres of epistemic accessibility.

Vygotsky's work is well known, among other things, by his famous study on the relationship between language (speech, in his words) and thought (problem solving ability). Vygotsky's account (1962) argues that language, being initially a social communicative phenomenon (inter-psychological) becomes an inner phenomenon (intra-psychological) through a process of internalization. But the notion of "internalization", in a Marxist thinker like Vygotsky, cannot but mean a dialectical process, and we have to understand that meaning.

From a non-dialectical point of view, internalization means simply that something moves from one place (the external space) to another place (the internal space). On the one hand, this movement is just a displacement, a change of location, and on the other hand, the internal space toward which that something is moving is an already constituted domain, a space that is there waiting for the arrival of something. In our case, the internalization of language would mean that language is simply "transferred" from an external domain to an internal domain, and that this internal domain is an already constituted space waiting for the arrival of language. This view is referred to by Wertsch (1985) as the "transfer model of internalization".

In contrast, the dialectical approach understands the process of internalization as a phenomenon which is both transformational and

constitutive. The internalization transforms the process itself (Vygotsky, 1981) and creates the internal domain through this very transformation (Leontiev, 1981). Taking this dialectical approach, we see the internalization of language as a process in which: 1) language does not merely "move" from one domain to another but suffers a deep transformation (a metamorphosis), 2) the very process of internalization creates or constitutes the internal domain in which language, so transformed, starts to work. When we say that mind emerges as an internalization of language, what we mean is that mind emerges as a transformation of language, and that the private (inner) character of the mental is constituted by this very process of internalization. From this viewpoint, language does not join an already constituted mind that it is waiting to become a "linguistic mind". Language, in being internalized, *generates* the mental space as a new experiential domain in the organism. The mental does not pre-exist the linguistic; it is rather its transformation or metamorphosis.

When something suffers a transformation, the result is something that conserves certain features from its previous state but that bears new characteristics too. In short, after a process of transformation, that something is something different. How different? Well, that depends on the particulars of the case. A transformation may be more or less superficial, more or less profound. The image of the caterpillar turning into a chrysalis, and the chrysalis turning into a butterfly, may be useful for visualizing our hypothesis about the relation between language and mind.

Which and how many features can you recognize in the butterfly as coming from the caterpillar? If you see a caterpillar walking on a leaf and a butterfly flying above you, could you deduce, by the mere observation of their bodies and behaviors, that one of them is nothing more than the mature version of the other? That is unlikely. What you need is to observe the *process* of metamorphosis (the development, the history, the genesis) that leads the caterpillar toward the butterfly. Once you have done that, you can see that despite the deep transformation suffered by the caterpillar, there are still some essential features that are conserved in the butterfly.

We see the relation between mind and language in a similar way. According to our view, mind emerges as a metamorphosis of language. Yet

we know that metamorphosis is a process in which, although many things change in a dramatic way, some essential features are conserved. Which features of language are lost and which ones are conserved in the constitution of mind? Our Vygotskyan hypothesis contends that, when internalized, language dramatically changes its syntactic structure and its forms of semantic articulation, conserving its representational (referential, intentional) power.

First, if the transformative process of language is really profound, as we think it is, then it should not be easy to appreciate, at least at first sight, the genetic continuity between language and mind. I think that is precisely the case when we examine our mental dynamic. Unless we are performing inner speech or thinking verbally in a propositional way, most of the times our mental experience does not reveal its linguistic credentials. As with the butterfly and the caterpillar, mind does not show its linguistic origins in an easy way. Why? Basically because the internalized language suffers a severe process of structural abbreviation (simplification) whose result is an almost unrecognizable syntax (almost absent), wherein what predominates is rather the free flow of oversaturated semantic contents. Let us see these ideas briefly.

Vygotsky (1962) saw this process of abbreviation basically as a predicative tendency–the tendency of propositions to lose their subject and conserve only the predicate. For instance, if I ask you "What time is it?" you might simply respond "Five" rather than "The time is five". The subject "the time" has been suppressed and only the predicate "five" remains. Though this predicative tendency is effective, I tend to think that the syntactic transformation of language goes deeper than this. Our mental experience, in processes like perception and action, does not show any linguistic character at all; we do not feel an internal 'voice' commenting on our perceptions and actions in terms of concepts or predicates. Yet our perceptions and actions appear as meaningful phenomena. This meaningful aspect, according to our hypothesis, is what remains from language once internalized. Language loses its syntactic envelope but conserves its semantic essence, and that continuous flow of semanticity so internalized constitutes what we call our mental experience; the experience of instantiating representations and

intentional states in a private or intrapersonal field.

We have said that when language is transformed into an individual phenomenon, it undergoes an important structural transformation that deeply changes its syntactic configuration. Why? What is the difference between public and private language? Basically, the transformation of the original dialogical structure of language has to do with the fact that, the more recurrent and familiar the relation between the linguistic participants, the more recursive and codified their interactions become. The structure of the language, the economy of its signs, changes radically as the familiarity between the communicative agents is maximal.

Originally, language is built through the interaction between different organisms. Insofar as this process takes place among different organisms whose different structures entail different experiential domains, language arises effectively as coordination of different experiential domains. Now, the more different the experiential domains, or, equivalently, the less shared the experiences, the more explicit (specified) the linguistic coordination needs to be. On the other hand, the more shared a background of experiences, the less explicit or specified the linguistic coordination needs to be. Think of the way in which you communicate with a person that you meet for the first time, and compare it to the way in which you communicate with someone whom you have known for a long time (e.g., an old friend, a long term partner). With an old friend you may need a couple of words, sometimes just a couple of gestures, to achieve a perfect and complex communication. Sometimes, people who have lived together for years need only one look to coordinate a complex array of actions. On the contrary, with an unknown person you usually need to articulate full sentences using an explicit grammar in order to communicate something. What is the difference? The difference is that with familiar people you have a rich history of recurrent communicative interactions such that you have co-created further recursive levels of coordination. That is, you have created some communicative codes (behavioral linguistic coordination that refers to other linguistic coordination) that remain intelligible only for those who have shared that peculiar history of interactions. This is what allows you to abbreviate the behavioral coordination to simple gestures or looks.

The epistemically private character of certain linguistic interactions refers to a communicative space that is restricted only to those who know the recursive rule behind such interactions. This can happen with people who deliberately agree a code so that other people cannot access the meaning of a communication (like spies and secret agents), or may be the spontaneous result of a history of shared experiences that generates in the participants a sort of secret code (like you and your old friend). These communicative spaces are epistemically opaque for everyone else; only the internal participants have access to the semantic rules.

The peculiarity of our mental experience is that in this case the familiarity of the communicative participants is maximal, since it is one and the same person who establishes the communicative phenomenon within an experiential flow that is always continuous (without epistemic gaps). The abbreviation is maximal, the code ultra-condensed, and the semantic space completely private in the individual sense. This transformation is profound, like a metamorphosis. Consequently, the subjective experience shows almost no trace of language, almost no trace of some inner communicative dialogue. What appears is a meaningful experiential flow without trace of words or sentences.

The syntactic envelope of language has almost completely disappeared to leave room only for its recursive core; such is the semantic, representational and intentional power of mind.

Conclusion

In this thesis we have tried to answer two fundamental questions:

1) What kinds of natural systems are living beings such that they behave in ways that we observers qualify as intelligent, cognitive, purposeful or intentional?
2) How do mental phenomena arise in the life of certain living beings?

We have answered the first question by saying that living beings are:

1) Adaptive dynamic systems
2) Deterministic machines of closed transitions
3) Multistable dissipative systems, and
4) Organizationally closed systems with respect to their sensorimotor activity and their autopoietic dynamic

We have answered the second question by saying that mental phenomena, basically understood as representational and intentional phenomena, arise with language, and that language, in turn, arises as a recursive phenomenon in the communicative domain of third-order biological systems. Additionally, and in a more speculative vein, we have tried to explain the epistemically private character of mental phenomena by appealing to a dialectical process of internalization of language.

The distinctive mark of these answers, as I see it, is that they have been built on the basis of a Strict Naturalistic framework. In this framework living beings are conceived of as metaphysically ordinary natural systems, and studied by applying the same ontological assumptions and explanatory principles applied in the study of any other natural system. This fundamental metaphysical assumption has been condensed in the Maturanian notion of

'structural drift'.

In assuming that living beings are systems in structural drift, we have rejected a series of assumptions and explanatory notions that, though relatively popular in cognitive sciences, overlook or violate this metaphysical condition. Conspicuously, we have rejected the idea that living beings, or subsystems such as the nervous system operate:

1. As cognitive agents that relate with the environment in epistemic terms
2. With the aim of fulfilling tasks, solving problems or meeting goals
3. As control systems that regulate the behavior according to the requirements of 2
4. On the basis of plans, models, maps or any form of internal representation
5. With a certain margin or degree of freedom to act
6. Intentionally open or oriented to (having in view, concerned about) the external world

In this thesis we have not directly confronted this view with any cognitive theory in particular, but it should be apparent that, if we did it, our Strict Naturalistic stance would easily find a considerable number of opponents. This is because what we object to, to a large extent, is a series of fundamental assumptions which are more or less shared by a wide spectrum of cognitive theories.

For example, points 1 and 6 refer to a fundamental assumption in cognitive science; the assumption that living beings make 'cognitive' contact, or establish an epistemic and intentional relationship, with the environment.

Most cognitive theories understand this epistemic and intentional relationship in terms of information. The basic idea is that organisms are informed 'about' the environment through their sensory 'windows'; i.e., organisms are *informavore* systems (Pylyshyn, 1984). Different approaches may disagree about, for example, whether this information is simply collected and directly used to produce some action (Gibson, 1966, 1979; Chemero, 2009; McDowell, 1994), or processed to build some form of internal representation (Millikan, 1993; Menary, 2007). Others, I would say

a minority, may even disagree with the idea that the epistemic and intentional relationship that the organism establishes with its environment is informational in character. Enactivist and some Heideggerian authors, for example, prefer to see this relationship in phenomenological terms; i.e., as meaning creation or sense-making (Varela, Thompson and Rosch, 1991; Gallagher, 2008; Thompson, 2007; Hutto, 2011; Dreyfus, 2008).

What our thesis questions, nonetheless, is the intentional and epistemic relationship itself, regardless of the way it is understood.

Another example is points 2 and 3. The idea that the organism, or its nervous system, is able to control the behavior according to certain adaptive goals is shared by both representational (Clark, 1997; Wheeler 2005; Grush 2004; Clark and Grush, 1999) and non-representational (enactive) cognitive theories (Di Paolo, 2005; Froese and Ziemke, 2009). Both of them assume that the organism or its nervous system anticipates, monitors, and regulates the actions in order to meet some adaptive goal.

What our thesis questions is this very assumption, regardless of whether it is, or is not, spelled out in representational terms (Maturana, 2008).

The reader might find, I am sure, several other examples of potential theoretical confrontations. If this thesis has motivated the reader to search for such confrontations, or to think (Why not?) of possible philosophical allies, then the thesis has met one of its goals.

Future directions of research

There are many aspects that were not addressed in this thesis, or that were too superficially addressed, and that might be explored in future works. Among them, perhaps the most important is human mental activity and experience.

For example, when analyzing the process of perception in animals we argued that the environment, as an object of perception, exists only for the observer, not for the animals. Well, what happens when the one who perceives is not an animal but the observer? What is the case with *our* perceptual experience as observers? If our brain is a structurally determined system, and is therefore subject to all the restrictions we have examined in

this thesis, what do we perceive when we perceive? Can we humans assert the existence of an external world, an independent reality populated, among other things, by animals and environments?

When applied to our own perceptual experience, PSD seems to lead us, perhaps unavoidably, toward a kind of skepticism. Is this an unavoidable result?

We have argued that living systems, including the nervous system, do not operate on the basis of intentional and representational elements. On the other hand, we have said that intentional and representational phenomena arise with language, and that language emerges in humans as a recursive communicative phenomenon. Thus, in this thesis we have not denied the existence of intentional and representational phenomena, rather we have placed them outside the internal structural dynamic of living systems; more concretely, outside the brain. Where? It is not clear at all.

Strict Naturalism, recall, distinguishes between two different Natural domains: the domain of natural phenomena and systems (rivers, stars, volcanoes), and the domain of human or cultural phenomena and systems (social institutions, symbols, mental phenomena). It asks us not to conflate these domains, and avoid any form of over-naturalization, anthropomorphism and observational projection. Rivers and stars are natural entities. They do not have minds; therefore, they do not have interests, purposes, temporal experience, epistemic states, etc. We humans are observers and have mental life; therefore, we have epistemic states, interests, purposes, temporal categories, etc. But where does all this mental activity take place? We cannot answer "In our brains", for our brains are just natural systems like rivers and stars.

Chapter 6 tried to elaborate some vague ideas, mainly through metaphors and analogies, about this point. They are, however, clearly insufficient. Much more philosophical work needs to be done in order to make legitimate room, within a Strict Naturalistic picture of the world, for our mental experience and its intentional and representational phenomena.

This thesis, in a sense, is an invitation to undertake the task.

Bibliography

Achinstein, P. (1983). *The nature of explanation*. New York: Oxford University Press.

Ashby, W. R. (1947). The Nervous System as Physical Machine: With Special Reference to the Origin of Adaptive Behavior. *Mind, New Series, 56, (221),* 44-59.

Ashby, W. R. (1956). *An introduction to cybernetics*. London: Chapman & Hall.

Ashby, W. R. (1960). *Design for a brain*. London: Chapman & Hall. 2nd edition (revised).

Ashby, W. R. (1962). Principles of the self-organizing systems. In H. Von Foerster and G. W. Zopf, Jr. (Eds.), *Principles of Self-Organization: Transactions of the University of Illinois Symposium* (pp. 255-278). London, UK: Pergamon Press.

Austin, J. L. (1962). *How to do things with words*. Oxford: Clarendon Press.

Barbieri, M. (2008). Life is semiosis: The biosemiotic view of nature. *Cosmos and History: The Journal of Natural and Social Philosophy, 4 (1-2),* 29-51.

Barbieri, M. (2012). Codepoiesis – The deep logic of life. *Biosemiotics, 5,* 297–299.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science, 22 (3),* 295-318.

Bermúdez, J. L. (2003). *Thinking without Words.* Oxford: The Oxford University Press.

Bickhard, M. (2007). Mechanism is not enough. *Pragmatics and Cognition, 17 (3),* 573-585.

Bishop, R. (2002). Deterministic and indeterministic descriptions. In H. Atmanspacher and R. Bishop (Eds.), *Between chance and choice: Interdisciplinary perspectives on determinism* (pp. 5-32). UK: Imprint Academic.

Bitbol, M. & Luisi, P.L. (2004). Autopoiesis with or without cognition: defining life at its edge. *Journal of the Royal Society Interface 1*, 99–107.

Bourgine, P. & Stewart, J. (2004). Autopoiesis and cognition. *Artificial Life 10(3),* 327–345.

Brentano, F. (1874/1973). *Psychology from an Empirical Standpoint*. London: Routledge & Kegan Paul.

Canguilhem, G. (1965). *La connassaince de la vie.* France: Vrin.

Cannon, W. B. (1932). *The wisdom of the body*. New York: The Norton Library.

Chemero, A. (2000). Anti representationalism and the dynamical stance. *Philosophy of Science, 67,* 625-647.

Chemero, A. (2009). *Radical embodied cognitive science.* Cambridge, MA: The MIT Press.

Clark, A. (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. Cambridge, MA: MIT Press.

Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, Mass.: MIT Press.

Clark, A. (2001). *Mindware: An introduction to the philosophy of cognitive science*. New York: Oxford University Press.

Clark, A. & Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior, 7,* 5-16.

Clark, A. & Toribio, J. (1994). Doing without representing. *Synthese, 101 (3),* 401-431.

Crane, T. (1992). The Non-conceptual Content of Experience. In T. Crane (ed.), *The Contents of Experience* (pp. 136-157). Cambridge: Cambridge University Press.

Crane, T. (1998). Intentionality as the mark of the mental. In A. O'Hear (ed.), *Current Issues in the Philosophy of Mind* (pp. 229–51). Cambridge: Cambridge University Press.

Crane, T. (2001). *Elements of Mind*. Oxford: Oxford University Press.

Damiano, L. & Luisi, P.L. (2010). Towards an autopoietic redefinition of life. *Origins of life, evolution and biosphere, 40*, 145-149.

Di Paolo, E. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences, 4(4),* 429–452.

Di Paolo, E. (2009). Overcoming Autopoiesis: An Enactive Detour on the Way from Life to Society. In R. Magalhães and R. Sanchez (Eds.), *Autopoiesis in Organization Theory and Practice* (pp. 43–68). UK: Emerald Group Publishing Limited.

Dicks, H. (2011). The self-poetizing Earth: Heidegger, Santiago theory, and Gaia theory. *Environmental Philosophy, 8(1),* 41-61.

Dougall, C. (1999). Autopoiesis and Aristotle: Rethinking organisation as form. *Kybernetes, 28 (6),* 777-791.

Dougall, C. (2000). Reconstructing Maturana - Metaphysics and method. *Kybernetes, 29 (4),* 491-498.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, Mass.: The MIT Press.

Dreyfus, H. L. (2008). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. In P. Husbands, O. Holland and M. Wheeler (Eds.) *The Mechanical Mind in History* (pp. 331–71). Cambridge, Mass.: MIT Press.

Dupuy, J-F. (2009). *On the origins of cognitive science: The mechanization of the mind*. Trans. M. B. DeBevoise. Cambridge, MA: MIT Press.

Flanagan, O. (2006). Varieties of Naturalism. In P. Clayton and Z. Simpson (Eds.), *The Oxford Handbook of Religion and Science* (pp. 430-452). NY: Oxford University Press.

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

Fodor, J. (1981). *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press.

Fodor, J. & Pylyshyn, Z. (1988). Connectionism and the cognitive architecture: A critical analysis. *Cognition, 28,* 3-71.

Froese, T. & Ziemke, T. (2009). Enactive artificial intelligence: investigating the systemic organization of life and mind. *Artificial Intelligence, 173,* 466–500.

Gallagher, S. (2008). Are minimal representations still representations? *International Journal of Philosophical Studies, 16 (3),* 351-369.

Gallagher, S. & Miyahara, K. (2012). Neo-pragmatism and enactive

intentionality. In J. Schulkin (Ed.) *Action, perception and the brain: Adaptation and cephalic expression* (pp. 117-146). New York: Palgrave-Macmillan.

Gardner, B.T. & Gardner, R.A. (1971). Two way communication with an infant chimpanzee. In A.M. Schrier and F. Stollnitz (Eds.), *Behavior of nonhuman primates* (Vol. 4, pp. 117–135). New York: Academic Press.

Gibson, J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Godfrey-Smith, P. (1996). *Complexity and the function of mind in nature*. NY: Cambridge University Press.

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines, 8,* 101-118.

Goyal, P. (2010). What is quantum theory telling us about how nature works? In C. Rangacharyulu and E. Haven (Eds.) *Proceedings of the first interdisciplinary CHESS interactions conference* (pp. 203-215). Singapore: World Scientific.

Goyal, P. (2011). Deciphering quantum theory. In M. Emam (ed.) *Are we there yet? The search for a theory of everything* (pp. 106-115). Sharjah: Bentham Science Publishers.

Grush, R. (1997). Review of "Mind as Motion: explorations in the dynamics of cognition." *Philosophical Psychology, 10 (2),* 233-242.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences, 27,* 377-442.

Hall, N. (2004). Two Concepts of Causation. In J. Collins, N. Hall and L.A. Paul (Eds.) *Causation and Counterfactuals* (pp. 225-276). Cambridge: MIT Press.

Haselager, W.F.G. (2004). O mal estar do representacionismo: sete dores de cabeça da Ciência Cognitiva. In A. Ferreira, M.E.Q. Gonzalez and J.G. Coelho (Eds.), *Encontros com as Ciências Cognitivas, 4* (pp. 105-120). São Paulo: Coleção Estudos Cognitivos.

Heidegger, M. (1977). *The question concerning technology, and other essays*. New York & London: Garland Publishing.

Hutto, D. (2011). Philosophy of mind's new lease on life: Autopoietic enactivism meets teleosemiotics. *Journal of Consciousness Studies, 18(5-6),* 44-64.

Hutto, D. & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content*. Cambridge, MA: MIT Press.

Ilharco, F. (2003). Building Bridges in Phenomenology: Matching Heidegger and Autopoiesis in Interpretive Research. 2nd POT Workshop, *Phenomenology, Organisation and Technology 2nd International Workshop*, September 18th-19th, 2003, Catholic University of Portugal, Lisbon.

Jaeger, G. (2003). *Quantum entanglement, information, and the foundations of quantum mechanics* [electronic resource]. Berlin: Springer.

James, W. (1909). *The Meaning of Truth*. London: Longmans, Green & Co.

Jaynes, E. T. (1989). Clearing up mysteries—the original goal. In J. Skilling (Ed.) *Maximum Entropy and Bayesian Methods* (pp. 1-27). Holland: Kluwer Publishing Co.

Jaynes, E. T. (1990). Probability in quantum theory. In W. H. Zurek (Ed.) *Complexity, Entropy, and the Physics of Information* (pp. 381-404). Redwood City, CA: Addison-Wesley Pub. Co.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. United Kingdom: Cambridge University Press.

Jenann, I. (Fall 2009 Edition).Quantum Mechanics. Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy.*URL=<http://plato.stanford.edu./archives/fall2009/entries/qm/>

Ji, S. (2012). *Molecular theory of the living cell: Concepts, molecular mechanism, and biomedical applications*. New York: Springer.

Johnson, D.K. (1991). Reclaiming reality: A critique of Maturana's ontology of the observer. *Methodologia, 9,* 7-31.

Jonas, H. (1966). *The phenomenon of life*. New York: Harper & Row.

Kelso, J. A. S. (1995). *Dynamic patterns: The Self-Organization of Brain and Behavior*. Cambridge, Mass.; London: MIT Press.

Kennedy, J. (1992). *The new anthropomorphism*. Cambridge, England: Cambridge University Press.

Khalifa, K. (2013). The role of explanation in understanding. *The British Journal for the Philosophy of Science, 64 (1),* 161-187.

Kim, J. (2011). *Philosophy of mind*. CO: Westview Press.

Laloë, F. (2012). *Do we really understand quantum mechanics?* New York: Cambridge University Press.

Leontiev, A.N. (1981). *Problems of the development of mind*. Moscow: Progress.

Lipton, P. (2009). Understanding without Explanation. In H. W. de Regt, S. Leonelli and K. Eigner (Eds.), *Scientific Understanding* (pp. 43–63). Pittsburgh: University of Pittsburgh Press.

Maturana, H. (1975). The organization of the living: A theory of the living organization. *International Journal of Man-Machine studies, 7,* 313-332.

Maturana, H.R. (1978a). Biology of Language: The Epistemology of Reality. In G. Miller and E. Lenneberg (Eds.), *Psychology and Biology of Language and Thought: Essays in Honor of Eric Lenneberg* (pp. 27-63). New York, NY: Academic Press.

Maturana, H. (1978b). Cognition. In P. M. Hejl, W. K. Köck, and G. Roth (Eds.) *Wahrnehmung und Kommunikation* (pp. 29-49). Frankfurt: Peter Lang.

Maturana, H. (1970/1980). Biology of cognition. In H. Maturana and Varela, F., *Autopoiesis and Cognition: The Realization of the Living* (pp. 5-56). Dordrecht, Holland: Kluwer Academic Publishers.

Maturana, H. (1980). Autopoiesis: Reproduction, heredity and evolution. In M. Zeleny (Ed.), *Autopoiesis, dissipative structures, and spontaneous social orders* (pp. 45-79). Colorado: Westview Press.

Maturana, H. (1981). Autopoiesis. In M. Zeleny (Ed.), *Autopoiesis: a theory of living organization* (pp. 21-33). New York; Oxford: North Holland.

Maturana, H. (1987). Everything is said by an observer. In W. I. Thompson (Ed.), *GAIA: A way of knowing* (pp. 65-82). Hudson, N.Y.: Lindisfarne Press.

Maturana, H. (1988). Reality: the search for objectivity or the quest for a compelling argument. *Irish Journal of Psychology, 9 (1),* 25-89.

Maturana, H. (2000). The nature of the laws of nature. *Systems Research and Behavioural Science, 17,* 459-468.

Maturana, H. (2002). Autopoiesis, structural coupling and cognition: A history of these and other notions in the biology of cognition. *Cybernetics and Human Knowing, 9(3–4),* 5–34.

Maturana, H. (2003). The biological foundations of self-consciousness and the physical domain of existence. In N. Luhmann, H. Maturana, M. Namiki, V. Redder, and F. Varela, *Beobachter: Konvergenz der Erkenntnistheorien?* (2nd ed., pp. 47-117). Munich: Wilhelm Fink Verlag.

Maturana, H. (2008). Anticipation and Self-consciousness. Are these functions of the brain? *Constructivist Foundations, 4 (1),* 18-20.

Maturana, H. (2011). Ultrastability... autopoiesis? Reflective response to Tom Froese and John Stewart. *Cybernetics and Human Knowing, 18(1–2),* 143–152.

Maturana, H. & Pörksen, B. (2004). *From being to doing: The origins of the biology of cognition*. Bingley, UK: Emerald Group Publishing Limited.

Maturana, H. & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht, Holland: Kluwer Academic Publisher.

Maturana, H. & Varela, F. (1987). *The Tree of Knowledge*. Boston and London: Shambhala New Science Library.

Menary, R. (2007). *Cognitive Integration: Mind and Cognition Unbounded*. New York: Palgrave Macmillan.

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

Millikan, R.G. (1993). *White queen psychology and other essays for Alice*. Cambridge, MA: MIT Press.

Mingers, J. (1995). *Self-Producing Systems: Implications and Applications of Autopoiesis*. New York: Plenum.

Mitchell, R. W. & Hamm, M. (1996). The interpretation of animal psychology: Anthropomorphism or behavior reading? *Behaviour, 134,* 173–204.

Müller, A. & Müller, K. (Eds.) (2007). *An unfinished revolution*. Vienna, Austria: Edition Echoraum.

Nicolis, G. & Prigogine, I. (1977). *Self-Organization in Nonequilibrium Systems: From Dissipative Structure to Order through Fluctuations.* New York: Wiley.

Peacocke, C. (1992). Scenarios, Concepts, and Perception. In T. Crane (ed.), *The Contents of Experience: Essays on Perception* (pp. 105-135). Cambridge: Cambridge University Press.

Peacocke, C. (1998). Nonconceptual Content Defended. *Philosophy and Phenomenological Research 58*, 381-388.

Port, R. & van Gelder, T. J. (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge MA: MIT Press.

Prigogine, I. (1980). *From Being To Becoming*: *Time and Complexity in the Physical Sciences.* USA: Freeman and Company.

Prigogine, I. & Stengers, I. (1984). *Order out of Chaos: Man's New Dialogue with Nature*. USA; Canada: Bantam Books.

Pylyshyn, Z. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, Mass.: MIT Press.

Rae, A. (2004). *Quantum physics: Illusion or Reality?* 2nd edition. United Kingdom: Cambridge University Press.

Ramsey, W. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Rosenberg, A. (1996). A field guide to recent species of Naturalism. *The British Journal for the Philosophy of Science*, *47 (1)*, 1-29.

Russell, S. & Norvig, P. (2010). *Artificial intelligence: A modern approach.* USA: Prentice Hall.

Ryle, G. (1949). *The concept of mind*. USA: The University of Chicago Press.

Savage-Rumbaugh, E.S. (1986). *Ape language: From conditioned response to symbol*. New York: Columbia University Press.

Searle, J. (1975). A taxonomy of illocutionary acts. In J. Searle, *Experience and meaning: Studies in the theory of speech acts* (pp. 1-29). Cambridge: Cambridge University Press.

Searle, J. (1995). *The construction of social reality*. New York: The Free Press.

Sklar, L. (2009). Determinism. In J. Kim, E. Sosa and G.S. Rosenkrantz

(Eds.), *A Companion to Metaphysics* (pp. 2011-2013). UK: Wiley-Blackwell.

Smolensky, P. (1987). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy, 26* (Supplement), 137–163.

Spencer-Brown, G. (1979). *Laws of form*. New York: E. P. Dutton.

Sperry, R. W. (1943). Effect of 180 degree rotation of the retinal field on visuomotor coordination. *Journal of Experimental Zoology, 92 (3)*, 263-279.

Sperry, R. W. (1945). Restoration of vision after crossing of optic nerves and after contralateral transplantation of eye. *Journal of Neurophysiology, 8,* 15-28.

Stroud, B. (1996). The charm of naturalism. *Proceedings and Addresses of the American Philosophical Association, 70*, 43-55.

Thompson, E. (2007). *Mind in Life: Biology, phenomenology, and the sciences of mind.* Cambridge, MA: Harvard University Press.

Tolman, E. C. (1932). *Purposive behavior in animals and men*. New York: Appleton-Century-Crofts.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55,* 189–208.

Tolman, E. C. & Honzik, C. H. (1930). "Insight" in rats. University of California, Berkeley, *Publications in Psychology, 4*, 215–232.

Tyler, T. (2003). If horses had hands. *Society & Animals, 11*, 267–281.

Ulanowicz, R. E. & Hannon, B. M. (Nov. 23, 1987). Life and the production of entropy. *Proceedings of the Royal Society of London, Series B, Biological Sciences, 232 (1267)*, 181-192.

Valsiner, J. & Van der Veer, R. (1988). On the social nature of human cognition: An analysis of the shared intellectual roots of George Herbert Mead and Lev Vygotsky. *Journal for the Theory of Social Behaviour, 18*, 117-135.

van Fraassen, B. (1980). *The scientific image*. Oxford: Clarendon Press.

van Gelder, T. J. (1995a). The distinction between mind and cognition. In Y.-H. Houng and J.-C. Ho (Eds.), *Mind and Cognition* (pp. 57-82). Taipei: Academia Sinica.

van Gelder, T. J. (1995b). What Might Cognition Be, If Not Computation? *Journal of Philosophy, 92*, 345–381.

van Gelder, T. J. (1998). The Dynamical Hypothesis in Cognitive Science. *Behavioral and Brain Sciences, 21*, 615–65.

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

von Eckardt, B. (1995). *What is cognitive science?* Cambridge, MA: The MIT Press.

von Foerster, H. (1970). Thoughts and notes on cognition. In P. Gavin (ed.), *Cognition: A Multiple View* (pp. 25–48). New York: Spartan Books.

von Foerster, H. (2003). *Understanding understanding: essays on cybernetics and cognition*. NY: Springer Verlag.

Vygotsky, L.S. (1962). *Thought and language*. Cambridge, Mass.: MIT press.

Vygotsky, L.S. (1981). The instrumental method in psychology. In J.V. Wertsh (Ed.), *The concept of activity in Soviet psychology*. NY: Sharpe, M.E.

Weber, A., & Varela, F. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences 1,* 97–125.

Wertsch, J.V. (1985). *Vygotsky and the social formation of mind.* Cambridge, Mass.: Harvard University Press.

Wheeler, M. (2005). *Reconstructing the Cognitive World*. Cambridge, Mass.: MIT Press.

Wheeler, M. (2011). Mind in life or life in mind? Making sense of deep continuity. *Journal of Consciousness Studies, 18(5-6),* 148-168.

Wheeler, M. (Fall 2014 Edition). Martin Heidegger. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Forthcoming ULR <http://plato.stanford.edu/archives/fall2014/entries/heidegger/>.

Wiener, N. (1948). *Cybernetics: or Control and Communication in the Animal and the Machine*. Cambridge, Mass.: MIT Press.

Wynne, C. (2004). The perils of anthropomorphism. *Nature, 428,* 606.

Wynne, C. (2007). What are animals? Why anthropomorphism is still not a scientific approach to behavior. *Comparative Cognition and Behavior Reviews, 2*, 125–135.

Zolo, D. (1991). Autopoiesis: critique of a postmodern paradigm. *Telos, 86*, 61-80.