History, Philosophy and Theory of the Life Sciences

Alvaro Moreno
Matteo Mossio

Biological Autonomy

A Philosophical and Theoretical Enquiry

Springer

History, Philosophy and Theory of the Life Sciences

Volume 12

More information about this series at http://www.springer.com/series/8916

Alvaro Moreno • Matteo Mossio

Biological Autonomy

A Philosophical and Theoretical Enquiry

Springer

Alvaro Moreno
IAS-Research Centre for Life,
Mind and Society
Departamento de Lógica y Filosofía
Universidad del País Vasco
Donostia – San Sebastian
Guipúzcoa, Spain

Matteo Mossio
Institut d'Histoire et de Philosophie des
Sciences et des Techniques (IHPST)
CNRS/Université Paris I/ENS
Paris, France

ISSN 2211-1948 ISSN 2211-1956 (electronic)
History, Philosophy and Theory of the Life Sciences
ISBN 978-94-017-9836-5 ISBN 978-94-017-9837-2 (eBook)
DOI 10.1007/978-94-017-9837-2

Library of Congress Control Number: 2015933996

Springer Dordrecht Heidelberg New York London

Printed on acid-free paper

Springer Science+Business Media B.V. Dordrecht is part of Springer Science+Business Media (www.springer.com)

Foreword

I am sitting with my grandchild in the park on a fading Australian summer afternoon. The sulphur-crested cockatoos screech as they squabble over the last of the sunlit eucalypt branches and she notices them as they fly over her, their powerful wings beating. "What is flying," she asks, "and why can't we do it?" For a moment I am tempted to respond to the latter question with "We don't have the genes for flight". But not only would this not help, indeed could not help, it also tears that question away from the initial one. While evolution as changes in gene frequencies can track the requisite gene changes involved, the actual features of organisms that make flight possible are left out. Genes can tell us about the appearance, spread and evolution of the *fact* of flight but they cannot in themselves tell us what flight actually *is*, namely the production of suitably spatially distributed and temporally coordinated thrust and lift. To understand that involves understanding, for example, how musculature must be recruited and organised to work wings that provide both lift and thrust, how skeletons must be both organised to effect tail-wing coordination, and be light enough to lift yet strong enough to brace the musculature in flight, to land on moving branches without fracturing legs, etc. and much more.

In short, it is to understand the internal organisation of birds. Without that we are blind to the internal consequences of genetic variation; and without that and ecological organisation, blind also to its external ecological consequences via the new sources of food and nests that become available, the spread of seeds via bird guts and the spread of plants that compete for bird feeding, and so on. Without such understanding the survival-of-the-fittest engine is left spinning its wheels, its simple idea of stochastic selection on populational variety left to sort rocks in a river and straws in the wind as well as gazelle on a savannah but without purchase on the nature and potential of evolving life.

There are three good reasons to read this book about how life is constituted. *First*, its organisational approach to organism is deeply informative, radically different from current orthodoxy and makes a crucial contribution at an important historical juncture in science. *Second*, it provides a detailed, powerful and ultimately elegant

model of the mutual development of scientific and philosophical understanding. *Third*, the pellucid, penetrating and parsimonious character of the writing makes a text dense in precisely characterised ideas quite accessible, including to a non-expert audience.

The last of these features is uncommon, the second is decidedly unusual and the first quite unique. They are all discussed a little further below. If you have an interest in understanding how our world works, or a specific interest in the foundations of biological science and/or philosophy of biology, or in organised complex systems more widely (robotics, cybernetics, intelligent agents, etc.) then this book is for you.

The book expounds and explores the claim that a distinctive organisation is the hallmark of life and that organisation ultimately provides a framework for understanding the evolution of life forms, of agency and of intelligence/intentionality. A quick review of chapter content can be found towards the close of the Introduction. As you might expect, it starts with the basics, closure and self-maintenance, then a complex form of closure called autonomy, the foundational organisation of all life, and then explores the still more complex topics just noted. Moreno especially has pursued the organisational approach consistently over decades and, with Mossio as collegial co-writer, this book is the summative outcome. I have helped to make the odd contribution to this position myself, partly on its systems foundations (organisation), but largely concerned with the adaptive roots of cognition (see references, this book), and in my view this book is unique in offering the first high quality conceptually integrated, empirically grounded, in depth exposition of this approach. It shows just how far the organisational perspective can take us in understanding the nature and evolution of life (answer: very far) and its exposition bids fair to remain the standard for some time to come.

The Organisational Approach

Listing genes and gene-trait associations tells you little about how the creatures that carry the genes are put together. The common presumption is that those latter answers come after the genetic work is done and will be found by studying the biochemical detail. Then whatever organisation there is will drop out as a consequence. But there is another, reverse possibility, one that has been largely neglected, namely that there are irreducible structures of nested correlated interactions, that is, organisations, that are key to understanding why the biochemical details are as they are, genomes included, and that such organisational design is as fundamental to understanding as is the biochemistry. That is the approach taken here.

Organisation (think car engines) happens when many different parts (cylinders, cam shafts, fuel injection ...) interact in specific, coordinated ways (cylinder rod rotates on cam shaft, fuel injected into cylinder, ...) so as to collectively support some global functioning (convert chemical potential energy into torque). It is roughly measured by the numbers of nested layers of different correlations

among different parts of a system. It is their functional contributions to the overall organisation of car engines that require the parts to have the shapes, sizes and material compositions they have. You can study these parts separately but unless you relate them to their organisation you will not understand their particular features. Organisation is not the same as order; pure crystals are highly ordered, so uniform they cannot show any organisation. Neither can gases, because they are too random to be organised. Organisation lies between the crystal and gas extremes but we don't have a good theory that tells us exactly where and why. Some may worry that talk of organisational constraints is too "airy fairy" and "metaphysical". But it is just the opposite, a matter of real dynamics found everywhere, from car engines to cellular "engines", for instance, the Krebs Cycle.

In this book the chief exemplar of an organised system is the living cell. The metabolism of a cell has to completely re-build the cell over time (that is its grand cycle). This is because, being material, a cell is a thermodynamic engine whose internal interactions degrade its innards which must then be replaced. But you don't get systematic self-replacement without being highly organised to do it: the particular materials and energy needed for each repair must be available at just the right location at just the right time, otherwise the cell will malfunction. In a cell more than 3,000 biochemical reactions are so organised that with each kind distinctively distributed throughout the cell their joint products re-make the cell, including themselves (and remove the thermodynamically unavoidable wastes), in the process also re-making the cell's capacity to extract from its environment the resources it needs. Thus at the heart of every cell is, and must be, a massive self-maintenance organisation cycle, operating under just the right constraints. This kind of organisation is called autonomy, with its core sense of self-governance applying all the way "up" from self-restriction by constraints to the more familiar socio-political notion.

Moreno and Mossio show that such organisation is central to cellular function, essentially defining all life. They also show that it is the necessary precursor to a well-defined evolutionary process, rather than the other way around. This is because the internal organisation of organisms secures the reproducibility of functionality which permits the inheritable traits, including those for mutant genomes, on which evolutionary selection operates. The interaction between evolutionary and developmental dynamics, in the context of epigenetic organisation, once mostly ignored but now richly studied, throws into stark relief the role of organism organisation in framing evolutionary process. All this is a relatively new perspective for evolutionary theorists, whose pure population statistics in themselves discourage awareness of organismal, communal and ecological organisation (cf. flight, above; albeit the theory has itself evolved significantly over the past 50 years). Moreno and Mossio lay out the issues with meticulous care.

Incidentally, it was the twin successes of the explorations of population genetics and molecular genetics that led to a century-long relative repression of biological organisation as an object of study, a repression that only really receded this century when molecular biology had exhausted simple gene sequencing and medicine simple gene-trait associations and both admitted the study of biosynthetic pathway

organisation as the next major challenge. Thus this book arrives on the scene at this epochal moment, just in time to provide a penetrating framework for understanding what is actually involved in such research.

On that score, note that the science of spatio-temporal organisation of interactions so as to generate global self-maintenance is itself in its infancy; we know relatively little about it, but just enough about the incredibly complex ways reactants are spatially arranged in cells to suppose it is going to be a large, complex and very difficult domain to understand. But it must come if ever we are to develop a thorough cellular biology and much else up to truly life-like robotics beyond the one-dimensional computer-in-a-box toys we focus on at present. (See also Hooker ed. *Philosophy of Complex Systems*, North Holland 2011 for further discussion.)

Multicellular Organisation

The emergence of a biochemical organisation capable of regenerative closure, the cell, is the first decisive step in the evolution of life. A subsequent giant step is the organisation of groups of cells to form multicellular organisms. These must organise their multicellular processes so that cellular metabolism is supported throughout, hence the presence of a cardiovascular system to deliver oxygen and nutrients where needed and remove wastes, the presence of renal and lymph systems to manage toxins and so on. In short, multicellularity requires a set of "higher" organisational layers on top of cellular ones to obtain a functional organism. (Again, we do not as yet understand a lot about such organisational constraints, e.g. respiration, that reach from individual cells across organs and other intermediate organisations, to the whole organism.) But there is a pay-off for all this overhead.

The distinctive twin advantages of multicellularity lie in its increased capacities for more complex behaviours and for more interactively open organisation, each feeding the other, even while closure must still be satisfied for their component cells. Once cellular communication develops to allow cell specialisation compatibly with cellular organisational coherence (as above), the way is thrown open to great increases in both behavioural complexity and interactive openness. The case of expanded behavioural repertoires is obvious enough. No single cell can fly, for the good reason that, whether or not it can muster thrust, it cannot control its surface shape so as to provide lift. But a collection of cells suitably specialised and interconnected can provide the musculature, cardiovascular support, surface controllability and so on to fly, powerfully and elegantly.

The case of greater interaction openness is perhaps less obvious but of even greater significance. Multicellularity has made possible increases in interaction-led adaption of both inner metabolism and outer environment. In the case of inner metabolism, multicellular organisms are able to suspend or adapt aspects of metabolic activity, from speeding up some processes (e.g. removing wastes before conflict) to slowing down and modifying others (e.g. hibernation in bears),

sometimes drastically (e.g. consuming internal organs for energy when fat stores are exhausted in stressful circumstances). Indeed, it is possible for existing organ systems to be entirely transformed in response to circumstances, as the metamorphosis of pupae into butterflies so beautifully illustrates. All of this requires over-arching organisational capacities. In the case of the outer environment, sensory cellular specialisation permits new ways of inward-bound interaction with the environment, leading to increased motor metabolic adaptiveness, from movement (e.g. sitting to running) to fasting, and to new ways of outward-bound interaction with the environment, like fight/flight, but also altering the environment to ease selection pressures (mouse holes for mice, etc.). Humans do not even internally manufacture all of their essential amino acids, relying on these open interaction systems to obtain them from their environment. (Which means that any constraint closures required for organism autonomy must be understood relatively to what can be regulated through external interaction and not only internal metabolic activity.) Just as with flight, all this also transforms ecological organisation.

In sum, if I might exploit a flight metaphor, when it comes to the expansion of life on the planet, it may be evolutionary selection that provides the thrust, but it is organisation that provides the lift. It is, as Howard Pattee taught us, the coordination of organisational constraints that makes possible the accumulating diversity and complexity of life. If organisation without evolution is impotent, evolution without organisation is blind.

Integration of Science and Philosophy

The dominant tradition in (meta-)philosophy is that philosophy and science are not to be integrated because philosophy provides an a priori normative framework for the analysis, conduct and evaluation of science whereas science constructs a posteriori empirical knowledge of the world by applying that framework. But in practice the development of understanding has rarely (really: never) happened like this. Philosophers have always borrowed ideas, theories and methods from science, and vice versa, each fertilising the other, unregarding of the proprieties of doing so. This has been a GOOD THING for both parties, each informing the other and keeping it on its toes. A minority naturalist (meta-)philosophy position would also applaud this intercourse as entirely appropriate. And that is what our authors consciously practice. Here is what they say (see Introduction): "... the approach developed in this book lies in between philosophy and theoretical biology. It deals with philosophical questions, like the nature of autonomy, agency and cognition, as well as their relations with concepts such as function, norms, teleology and many others; yet, it addresses these questions in close connection to, or even deeply entangled with, current scientific research." What emerges from this rich process is a coherent, if unfinished, majestic view of life as a subtly mutually entangled, organised whole from molecules to macro-ecology.

On Chasing Hares

Like all really interesting books, this book is profoundly incomplete: it starts new hares (new lines of thought) running on almost every page. This leaves the curious and/or thoughtful reader to enjoy the pleasure of identifying them and deciding which ones to follow up. A fine example already occurs in Chap. 1, in the nature of the closure found in self-regeneration and its relation to dynamical constraints. This issue is central, for according to the book's story there is no function or organisation, properly so-called, without closure ("an organization is by definition closed and functional", Chap. 3) and hence no autonomy either. I have previously mentioned constraints five times, including in characterising autonomy itself, and closure thrice, as if both notions were well understood. Did you notice any hares leap?

Closure has been an issue in thinking about autonomous systems from the beginning (see the summary in their Chap. 1). Founders like Varela emphasised closure as the distinctive feature of biological organisation and made its discovery at multicellular levels the key requirement for understanding them, even though closure was hard to uncover (it was thought to characterise the immune and nervous systems) and seemed to pull against the increasing interactive and organisational openness that marks multicellularity (see above). Many (myself included) adopted a process model: processes are sequences of dynamical states and process closure occurs where these states cycle through a closed loop of states, returning each time to an initial state, e.g. the normal or "resting" metabolism state. The cellular Krebs cycle is again a useful example. The thermodynamic flow, another process, drives the cycling, thus reconciling openness (flow) with closure (cycling). But Moreno and Mossio find this unsatisfying (for reasons I leave to the reader to pursue) and have developed their own distinctive account on which it is constraints that are closed and not processes, which are open on account of the thermodynamic flow. By constraint closure is meant, roughly, that the constraints so interrelate as to reconstitute one another. (So there is still a process cycle, but it is among constraint conditions, leaving thermodynamic processes to remain open.) To make the distinction between constraints and processes really sharp, they require that constraints do not interact, in the sense of exchange energy/materials, with the thermodynamic flow, only shape its direction. Think of a river flowing between frictionless banks. For this reason, they characterise constraints as not being thermodynamic entities and in Chap. 2 they support that by arguing that they are emergent entities with respect to the thermodynamic flow.

What are constraints, these non-thermodynamic entities that somehow shape the flow while not being of it? In standard mathematical dynamics constraints appear in the application of dynamical models where, although not directly represented in the system dynamics flow equations, they apply forces that constrain the dynamical possibilities of the flow. When they do not interact with the flow (ironically for Moreno/Mossio) those forces can be calculated and, like all modelled forces, are grounded in physical configurations of matter and/or fields of the same sort as make

up the system being modelled, just located externally to it. But in autonomous systems all the matter/fields that give rise to those constraint forces have themselves to be assembled within the autonomous system itself in consequence of its constrained flow. Precisely that is the trick of autonomous self-regeneration, and a problem for understanding constraints.

For this means that constraints repeatedly degrade and have to be physically reconstructed, waste molecules literally replaced with new ones, etc. That is, the system itself must do work on its own constraints, or anyway on the matter/fields that give rise to them. Think of a real river that erodes and reconstructs its own banks as it flows. But that raises a first important issue: we have no workable methods for formulating the dynamics of systems that do work on their own constraints, the standard techniques of Lagrangian dynamics break down in this case. (See my "On the import of constraints in complex dynamical systems", *Foundations of Physics*, 2013, and earlier in Hooker (2011) above.) So how exactly are we to understand these systems and their self-reconstituting constraints? (Hare 1) This issue applies much more widely than biology, of course, since the self-formation and transformation of internal constraints is a major feature of complex dynamics anywhere (Hare 2). And, as noted above, but not in the book, apparently multicellular closed constraints have to be understood relative to an organism's interactional (agency) capacities, which itself depends on its functional, so closure, organisation (Hare 3).

And the manner in which Moreno/Mossio move to avoid facing the problem for autonomous dynamics (by requiring that constraints do no work and have none done on them) raises a second important issue: since constraints have to be reconstituted there are presumably periods of time when work is being done on at least some of them (on their supports): what kind of dynamics then applies to them and the flow? (Hare 4) These concerns are reinforced by a vivid picture in Chap. 3 of self-maintenance extended over time, for both intra-organism and inter-generational autonomous organisations, reinforced by the argument in Chap. 6 that developmental processes are necessary to multicellular constitution. (There is another group of hares loitering around these ideas.) But perhaps it also offers a way out in its conception of transmission of causal organisation over time that does not seem to require continuous satisfaction of closure (Hare 5). Even then, the hare 4 issue would remain to be addressed. And a further issue arises: considering time periods during which various proportions of constraints do not exist as such because they are doing work on some part of the system (including regenerating other constraints) and/or having work done on them (being regenerated), how large can those time spans be before system autonomy is considered disrupted and no longer explanatory of that system, and why? (Hare 6)

No doubt the authors will have anticipated such issues and been thinking about responses. (Their remarks on river banks and in a few other places reflect my earlier probings.) Irrespective, these questions should not be considered criticism of the book; to the contrary, they represent questions that could not be asked until the refined treatment of constraint closure Moreno and Mossio propose was available. And while there are lots of hares to startle, as there must inevitably be given our

ignorance and a penetrating book, the present book succeeds in blunting many of the criticisms (including mine) made of the organisational approach. For instance, theirs is a position that takes the nature and role of biological organisation far beyond simple self-organisation of the kind beloved of the complex-behaviour-from-simple-rules-among-many-components tradition in the physics of complex systems. Indeed, that latter kind of process includes forming crystal lattices and like, so in fact it has no direct relationship at all with the kind of nested-complementary-correlations-and-regulations-among-disparate-components that this book is concerned with. The former could in principle be extended to encompass biology via bringing all organic chemistry under atomic modelling, but even then "organisation" in "self-organisation" remains a misnomer. (Two more hares.) Again, the book's position takes external interaction (individual and evolutionary) as seriously as internal organisation, whereas there are other traditions (discussed in the book) that are more closed-off to its importance, e.g. as illustrated above for understanding multicellular capacities. Nonetheless, we may still wonder whether the full extent of the interactive openness has been appreciated: what would their account of consuming internal organs under stress or adapting closure to environmental extraction of amino acids look like? (Another hare.) Finally, here the organisational approach is used to illuminate a thoroughly embodied approach to mind, for example with a deep connection developed to body plan, that counters the concern with "lifting off" an abstracted organisational pattern that has only nebulous connection to nervous system dynamics, organisation and functioning. However, there is still room to wonder about how neural phenomena characteristic of neural networks, whether distributed representations or waves, fit with organisation. (Another hare.) In these and like ways, this book represents a marked step forward in developing the organisational approach.

Meanwhile, there is the serious fun of chasing down such interesting and epistemically rewarding hares.

Conclusion

The authors describe my review of the draft of this book as, among other things, "relentlessly critical" (see closing remarks, Introduction). This is a compliment to both parties. A decade or more earlier I had entertained the prospect of a book on autonomy and discussed the idea with Moreno – on one occasion after an ocean swim near my Australian home and over a little local sauvignon blanc with freshly shucked Sydney Rock oysters, which he commented were "the best oysters I have ever tasted". (The preceding year at his coastal village I ate the best turbot I had ever tasted.) I hopefully suggested that the book could begin by understanding life through a series of ever tightening dynamical and thermodynamical constraints culminating with a notion of autonomy as the unique allowed evolvable organisation, just as the Krebs Cycle is a solution to capturing free energy for the cell. "Go ahead!" he said, "Be quick! I shall eagerly await your analysis." Of course, he knew better

from years of trying just how hard that scientific task would be, still impossibly hard today where, for example, simple chemical cell models are still under development. I should have paid more attention to the quiet twinkle in his eye.

But we can all pay attention to what has been achieved. This book has thrust and lift. It is a masterly account of the organisational foundations of life, a splendid flight in the firmament of conception and understanding.

Professor Emeritus of Philosophy Cliff Hooker
Fellow of the Australian Academy of Humanities
PhD (Physics, Sydney University)
PhD (Philosophy, York University, Canada)

Contents

Introduction

Life as Autonomy

If we were to point out in a few words what characterises the phenomenon of life, we would probably mention the amazing plasticity and robustness of living systems, the innumerable ways they adapt, and their capacity to recover from adverse conditions. All these capacities have been on the surface of our planet since the origins of life, and for this reason we have become accustomed to seeing life as something almost "normal". And yet, looking at it from a more global perspective, life is quite an extraordinary phenomenon. In a short period of time (compared to the history of the universe), in a very tiny portion of the cosmos, a set of entities has managed to attain extremely improbable configurations, to keep them in far-from-equilibrium conditions, and to thrive under these conditions: self-organising, proliferating, diversifying, and even increasing their complexity. Furthermore, this persistently organised system (or, rather, this global system formed by millions of local, individualised systems, which combine decay and reproduction) has been able to deploy a set of selective forces, modifying its environment so as to enhance its own maintenance. In a word, life seems to be at the same time an extraordinarily precarious (and improbable) phenomenon and a powerful, robust, and easily expansive one.

Actually, this astonishing capacity to maintain highly organised systems seems to be the easiest way to recognise universally living matter beyond the specificities of terrestrial life. Present-day theories estimate that the universe came into being 13.7 billion years ago, while our planet was formed approximately 9 billion years later. In this period of time, or perhaps later, forms of organisation similar to early living systems on our planet possibly appeared in other parts of the universe. Indeed, if life appeared on our planet when certain physicochemical conditions were met, other planets with similar conditions could also have once supported forms of life. This

raises the question of how we could recognise these hypothetical extra-terrestrial living systems, and what would be the essential features of *any* form of (possible) life. In the last decades, this question has been widely discussed.

For some (Cleland and Chyba 2007), it is impossible to say how such "essential features of life" should be conceived, because we only know life as it manifests itself on Earth. Yet, if what we mean by "life" is any material organisation that has evolved from non-living physicochemical systems (therefore obeying the universal laws of physics and chemistry) and has attained at least a degree of complexity capable of generating the properties we associate with the simplest forms of terrestrial life, we should be capable of recognising it anywhere in our universe, regardless of how differently these systems may be constituted (Ruiz-Mirazo et al. 2004). At the same time, the huge variety of life forms that have appeared during the very long history of life on our planet (Ward and Brownlee 2004) might downplay the argument that we have had access only to a unique example of life among a hypothetically huge set of extra-terrestrial biological systems. Be that as it may, when facing the question of the nature of life, we could not do otherwise than formulate theories based on – and tested against – life as we know it.

It is because of its capacity to achieve and maintain higher degrees of complexity that physical sciences find it very difficult to explain how life has originated. For this reason, the question of the origin of life is deeply entangled with the question of its very nature. Is there some law or principle in the physical world that allows explaining the emergence of life as a necessity or, as Monod (1970) thought, is the origin of life so unlikely that it is almost a miracle? How could inert matter originate something that seems to be so deeply different in its properties?

From the perspective of the physical sciences, explaining life is a highly challenging task because the more complex a system is, the less probable it becomes both in its appearance and its persistence. At first approximation, it might be easy to understand how simple building blocks may spontaneously generate composite stable structures (atoms, molecules, macromolecules . . .) due to different levels of forces (Simon 1969): as a result of these interactions, increasingly complex stable structures appear (endowed, in many cases, with new interactive properties, not present in their separate parts, such as superconductivity, chemical affinity . . .). As the complexity of the structures increases, however, its maintenance becomes a problem: thermal noise increases fragility and, moreover, the coincidence or coordination of many highly specific processes becomes increasingly unlikely.

It is true that recent advances in thermodynamics explain the formation of composite aggregates (called "dissipative structures"), whose parts are tied together without intrinsic forces, ensuring their cohesion in far-from-equilibrium conditions. However, as we will discuss at length in this book, these systems appear spontaneously and persist only when specific external boundary conditions are met and, more importantly, they lack internal complexity and functionality. In contrast, biology deals with highly complex systems, so that something more than initial conditions and fundamental laws seems to be required to explain a world of complex biological systems.

Assuming that nature does not make leaps and that, therefore, there is a continuum between non-living matter and life,[1] there should be explanatory principles of the transition from non-living to living matter. As Fry (2000) has pointed out, the fundamental problem of the origin of life lies in the tension between the principle of continuity and the difficulty of explaining the obvious differences between non-living and living matter. If the origin of life is a legitimate scientific question (and we think it is), one should look for a theory that bridges the gap between physics and biology. In particular, since living beings are made of the same constituents as non-living entities, what is the nature of the organisation that enables them to achieve, maintain, and propagate such a high degree of complexity? And what are the consequences of this extraordinary capacity?

On our planet, life has developed for a long period of time and has colonised the most diverse environments – from the deep oceans or even several kilometres under the Earth's crust to the upper levels of the atmosphere; from the hottest environments (over 100°C) to extremely acid or radioactive ones. And if we consider life from an historical perspective, it is even more impressive how it has managed to adapt to the successive catastrophic events that have occurred on our planet during the last 3.5 billion years. Admittedly, only the simplest forms of life are capable of such extreme robustness and versatility; at the same time, these forms of life have also been able to innovate and evolve towards increasingly higher levels of complexity. Life, as it has developed on our planet, has gradually integrated more and more levels of organisation (from unicellular life to colonies, multicellular organisms and societies).[2]

How can we explain all this diversity and complexity? Ever since Dobzhansky's (1973) famous dictum that "nothing in Biology makes sense except in the light of evolution", mainstream thinking in biology has seen evolution by natural selection as the source of diversity at every level of biological organisation. Indeed, the unfolding of an evolutionary process by natural selection, based on heritable genetic mechanisms, allows life to explore many possible combinations and solutions in order to survive. And the evolution-centred view of life has been so dominant that the idea of organism (which played a key role in nineteenth century biology) has become almost dispensable (Morange 2003). However, in a very fundamental sense, we shall argue at length that the reality is rather the opposite: evolutionary mechanisms operate because they are embodied in the complex organisation of organisms. Thus, if we look for the roots of the impressive capacity of life to proliferate, to

[1] Philosophically, this assumption amounts to adopting a monistic stance. Chapter. 2 is devoted to a detailed analysis of the position of the autonomous perspective developed in this book in the debate on emergence, reduction, and related issues.

[2] Nowadays we know that this process of diversification and complexification is not a contingent fact, but rather something "inscribed" in the evolutionary nature of life. As Gould (1994) has argued, evolution is not aimed towards an increase in complexity; in fact, life originates in the simplest form and many organisms have remained successfully as such. However, a few organisms occasionally introduced innovations, "thus extending the right tail in the distribution of complexity. Many always move to the left, but they are absorbed within space already occupied".

create an enormous variety of forms, to adapt to completely different environments, and particularly, to increase its complexity, we shall focus on individual living entities, namely on organisms, because evolution[3] as an explanatory mechanism actually presupposes the existence of organisms. As Varela (1979) pointed out,

> evolutionary thought, through its emphasis on diversity, reproduction, and the species in order to explain the dynamics of change, has obscured the necessity of looking at the autonomous nature of living units for the understanding of biological phenomenology. Also I think that the maintenance of identity and the invariance of defining relations in the living unities are at the base of all possible ontogenetic and evolutionary transformation in biological systems (p. 5).

As Rosen also emphasised, the crucial question for understanding life lies in the nature of its organisation.[4] It is true that any known living being cannot have appeared except as a result of a long history of reproductive events, since such a complex organisation can only be originated through an accumulative historical process and, furthermore, that its long-term sustainability also requires inter-generational entailments. This is clearly reflected in the fact that, in order to be operational, genetic components (which contribute to specify the metabolic machinery and organisation of single biological entities) must be shaped through a process that involves a large number of individual systems and many consecutive generations, or reproductive steps. Yet, this does not mean that the organisation of organisms should be neglected; on the contrary, a theory of living organisation is fundamental for understanding how these evolutionary mechanisms could have appeared and how they could work.

A theory of the living based on the concept of organism aims to review the concept of evolution and its role in a new way, attempting to overcome the dichotomy – and often opposition – between what since Mayr's (1961) work is called the biology of proximate causes and that of ultimate causes. Our vindication of the central role played by the notion of organism in biology should be placed within this wider perspective, in which the explanatory emphasis is placed on organisation. As Hooker and Christensen (1999) have highlighted, in order to

[3]The term 'evolution' could be understood in a very broad sense, just as an historical process of causal entailments. However, since Darwin, the term evolution has acquired a more restrictive sense, referring to specific mechanisms of inheritance and several other conditions (see for example, Godfrey-Smith (2009)). We will discuss the relation between autonomy and evolution in Chap. 5; here, we use the more restrictive sense of the term.

[4]"We cannot answer the question (. . .) 'Why is a machine alive?' with the answer 'Because its ancestors were alive'. Pedigrees, lineages, genealogies, and the like, are quite irrelevant to the basic question. Ever more insistently over the past century, and never more so than today, we hear the argument that biology *is* evolution; that living systems instantiate evolutionary processes rather than life; and ironically, that these processes are devoid of entailment, immune to natural law, and hence outside of science completely. To me it is easy to conceive of life, and hence biology, without evolution" (Rosen 1991: 254–55).

properly understand the evolution of biological systems, traditional approaches need to be embedded within a more general dynamical-organisational theory.[5]

Therefore, it is at the level of organisms, understood as cohesive and spatially bounded entities, that the biological domain's organised complexity is fundamentally expressed. Seen from the perspective of their relations with their environment, individual organisms are systems capable of acting for their own benefit, of constituting an identity that distinguishes them from their environment (at the same time as they continue interacting with it as open, far-from-equilibrium systems). This capacity of living beings to act for their own benefit follows from their peculiar form of organisation.

Living beings are systems continuously producing their own chemical components, and with these components they build their organs and functional parts. In a word, their organisation is maintaining itself. This is why living systems cannot stop their activity: they intrinsically tend to work or they disintegrate. Actually, this inherent tendency of living entities to promote their own existence – to act on their own behalf – could be related to the idea of the *conatus*, to which Spinoza (1677/2002) refers to designate the innate inclination of any entity to continue to exist and enhance itself.[6]

The root of this drive to persist lies in the principles of biological organisation. As Jonas (1966/2001) pointed out, the organisation of living systems is characterised by the inseparability between what they are – their "being" – and what they do – their "doing". This feature is reflected in their metabolism, which consists of a set of processes that allow them to build and replace their structures, grow and reproduce, and respond to their environments. Metabolism is the ongoing activity by which living beings continuously self-produce (and eventually, re-produce), self-repair, and maintain themselves. Unlike the Cartesian argument (which has had so much influence during modernity[7]) that living beings are like man-made machines, Kant was the first author who defended the view that organisms are

[5] As a matter of fact, an organisational perspective seems to be taking shape in the new evolutionary developmental biology, which studies how the dynamics of development determine the phenotypic variation arising from genetic variation and how this affects phenotypic evolution (Laubichler and Maienschein 2007).

[6] As Spinoza (1677/2002) writes, "Each thing, insofar as it is in itself, endeavors to persist in its own being" (Ethics, part 3, prop. 6). This is understood as an intrinsic tendency or force to continue to exist. Striving to persevere is not merely something that a thing does in addition to other activities it might happen to undertake. Rather, striving is "nothing but the actual essence of the thing itself" (Ethics, part 3, prop. 7). See Duchesneau (1974) for an in-depth analysis of Spinoza's account of living systems, and a comparison with the Cartesian one.

[7] Actually, the Cartesian distinction between *res extensa* and *res cogitans*, which subsumed the biological domain within a global mechanistic vision of nature, facilitated a scientific research programme for studying living systems. It should be underscored that, while Descartes' metaphysical dualism is widely recognised and is a prominent feature of his *Meditations*, scholars in the past generation have also focused on the complexity of his natural philosophy, including his work in physiology, medicine but also on the passions, as displaying something very different: a more 'integrated' view of bodily function. See notably the essays collected in Gaukroger et al. (2000).

deeply different from machines because their parts and activities are non-separable, and the functions of these parts are not externally imposed, but rather intrinsically determined. According to Kant (1790/1987), since the activity performed by the parts of the organism is carried out for their own maintenance, organisms are intrinsically teleological. As he writes in the *Critique of Judgement*:

> In such a product of nature each part, at the same time as it exists throughout all the others, is thought as existing with respect to the other parts and the whole, namely as instrument (organ). That is nevertheless not enough (because it could be merely an instrument of art, and represented as possible only as a purpose in general); the part is thought of as an organ producing the other parts (and consequently each part as producing the others reciprocally). Namely, the part cannot be any instrument of art, but only an instrument of nature, which provides the matter to all instruments (and even to those of art). It is then – and for this sole reason – that such a product, as organized and organizing itself, can be called a natural purpose (CJ, § 65).

This view allows him to open up a gap in the physical world, since organisms cannot be brought under the rules that apply to all other physical entities. Thus, Kant asks himself:

> How purposes that are not ours, and that we also cannot attribute to nature (since we do not assume nature to be an intelligent being) yet are to constitute, or could constitute, a special kind of causality, or at least a quite distinct lawfulness of nature (CJ, § 61).

This "special" kind of causality is circular, namely, effects derive from the causes but, at the same time, generate them. The very organisation of living beings, in which the parts generate the whole, and, conversely, the whole produces and maintains the parts, shows a kind of intrinsic purpose. Kant grounds the idea of purposiveness (and teleology) in the holistic and circular organisation of biological organisms and, more precisely, in the fact that they are able to organise by themselves, to *self*-organise.[8] Unlike artefacts, organisms are "natural purposes": they are not produced or maintained by an external cause, but instead have the self-(re)producing and self-maintaining character that is revealed in the kinds of vital properties they display (reciprocal dependence of parts, capacity for self-repair and self-(re)production).

Today, some aspects of the Kantian perspective are undergoing resurgence. For example, the recent blossoming of systems biology (Kitano 2002; Science, *special issue* 2002; Bogeerd et al. 2007), focused on the complexity of biomolecular interaction networks, is much closer to a holistic or integrative conception of living systems than the reductionist views predominant in molecular biology. Thanks to the development of new scientific tools, these more holistic theories place the question of the organisation at the centre of biological research. This recent trend contrasts with the preceding history of biology, during which the Kantian view has often be seen as marginal (even through this view has been corrected by

[8] Actually, Kant has been one of first authors to use the term "self-organisation". In Chap. 1, we will briefly mention how the meaning of this concept has progressively shifted during the 20th century.

the recent historiography, see for instance Huneman 2007; Richards 2000; Sloan 2002), essentially because it was thought to be at odds with the model of causality predominant in Newtonian science.

And yet, the Kantian perspective had continuity in the (mostly Continental) Biology of the nineteenth century, especially in the work of Goethe and Cuvier (Huneman 2006). In the first part of the twentieth century, many biologists were still convinced that the nature of living organisation – understood, following Kant's inspiration, as the form in which the parts interact with each other to bring forth the properties of the whole – was one of the main issues of biology. This view was commonly labelled *organicism* (Wolfe 2010; see also Gilbert and Sarkar 2000). Organicism considers the observable structures of life, its overall organisation, and the properties and characteristics of its parts to be the result of the reciprocal interplay among all its components. The organicist tradition was influential in early twentieth century biology. During the twenties and thirties, a group of researchers, including Woodger, Needham, Waddington, and Wrinch, created the "Theoretical Biology Club", whose objective was precisely to promote the organicist approach to biology. This movement – in which we can include other authors, like Bernal and Bertalanffy – was characterised by a predominant anti-reductionist and holistic inspiration (Etxeberria and Umerez 2006). Among these researchers, the name of Waddington is worth stressing because his work, after the Second World War, permitted the connection between the organicist movement of the thirties and the new tendencies of the sixties and seventies.

To understand the roots of the current blossoming of the "Kantian-inspired organicist ideas" in biology during the twentieth century, let us mention some other scientific trends, falling outside the frontiers of biology.

First, during the thirties and forties, a number of physicists associated with the development of quantum theory, interested in the nature of biological organisation, turned their attention to biology. Among these scientists, it is worth emphasising the name of Schrödinger, who gave his famous lectures "What Is Life?" in 1943 (Schrödinger 1944). Following this work, other quantum physicists addressed the problem of what characterises the specificity of living systems with regard to physical ones. Among these we can include researchers like von Neumann and Pauli. Interestingly, the advances in physics inspired new attempts to challenge reductionist assumptions. For example, Rashevsky, according to his disciple Rosen, defended

> a principle that governs the way in which physical phenomena are organized, a principle that governs the organization of phenomena, rather than the phenomena themselves. Indeed, organization is precisely what relational biology is about (Rosen 1991: 113).

During the seventies, Rosen himself and Pattee (Umerez 2001) also developed an anti-reductionist view of the specific organisation of living systems, based on his analyses of the specific causation associated with emergent constraints that living systems generate (see further below).

Second, special emphasis should be put on the cybernetic movement. The cyberneticists were influenced by the work of the American physiologist Cannon

(1929) who, in the early 1930s, developed the concept of "homeostasis" (whose origins date back to the work of the French biologist Claude Bernard[9]) as a key feature of the organisation of living beings. According to Cannon, the idea of homeostasis expresses the tendency of living systems to actively maintain their identity, despite external perturbations or differences within their environment. During the 1970s, a new generation of cyberneticists, notably Von Förster, Ashby, and Maturana, created the so-called second-order cybernetics. This movement was especially interested in the study and mathematical modelling of biological systems, based on the ideas of recursivity and closure (Cahiers du CREA 1985). Second-order cybernetics is of special relevance for our purposes, since it constituted the scientific environment in which the theory of *autopoiesis* was elaborated (see below).

Third, after the work of Prigogine (1962), the idea of self-organisation in far-from-equilibrium conditions began to enter into scientific discourse in physics, which also helped the Kantian view to gain influence in biology. Yet, as we will discuss at length in Chap. 1, there is an important conceptual difference between the Prigoginian concept of (physical) self-organisation and the Kantian notion of (biological) organisation. As Fox Keller (2007) has pointed out, the kind of complexity of organisms resulting from an iterative processes of organisation that occur over time is completely different from the one-shot, order-for-free kind of self-organisation associated with some kind of non-linear dynamical systems. In particular, the former is constituted by functional parts, whereas the latter lacks functionality. The logic of the metabolism, for example, shows a functionally diversified organisation, clearly different in this sense from any physicochemical dissipative structure. In this sense, as we will see, what we need is a view of biological systems that goes beyond a generic vindication of an organisational-centred biology. What matters is the understanding of the *specificity* of the organisation of biological systems, which are not just self-organised systems.

In the second post-war period, both the New Synthesis in evolutionary biology and the revolution of Molecular Biology created a scientific atmosphere that was quite unprepared to accept organicist and Kantian views (Moreno et al. 2008). Accordingly, this tradition remained, until very recently, marginal in biology. In this context, however, Waddington was the main driver of a movement that advocated an organisational approach in biology, by reviving the "first" Theoretical Biology of the twenties and thirties (Etxeberria and Umerez 2006). This "second" Theoretical Biology was initially developed by several pioneering authors like Waddington himself (1968–1972), Rosen (1971, 1972, 1973, 1991), Piaget (1967), Maturana and Varela (1980), Pattee (1972, 1973), and Ganti (1973/2003, 1975). Many of these authors put strong emphasis on the idea that the constitutive organisation of biological systems realises a distinctive regime of causation, able not only of producing and maintaining the parts that contribute to the functioning of the system as an integrated, operational, and topologically distinct whole but also able

[9] See Bernard (1865) and (1878).

to promote the conditions of its own existence through its interaction with the environment. This is essentially what we call in this book *biological autonomy*.

To give a preliminary idea of what autonomy is about, let us mention one of its first and well-known accounts, the theory of *autopoiesis* proposed by the Chilean biologists Maturana and Varela in the early 1970s (Maturana and Varela 1973; Varela et al. 1974). In the theory of autopoiesis, although the concept of autonomy is applied to different specific biological domains (immune, neural . . . see Varela 1979), it characterises the fundamental feature of the living, namely, the autopoietic organisation. Autopoiesis refers to the capacity of self-production of biological metabolism, by emphasising (in a simplified and abstract way) its causal circularity – which Maturana and Varela called "operational closure". In particular, their model describes the production of a physical boundary, which is conceived as a condition of possibility of the internal chemical network (because it ensures suitable concentrations for the maintenance of the component production network); in turn, the network maintains the physical boundary (because it is the component production network which produces the special self-assembling components that build the membrane). In their own terms (in which the cybernetic flavour is manifest):

> An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components which: (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in space in which they (the components) exist by specifying the topological domain of its realization as such a network (Maturana and Varela 1980: 78).

Thus, autopoiesis consists in a recursive process of component production that builds up its own physical border. The global network of component relations establishes self-maintaining dynamics, which bring about the constitution of the system as an operational unit. In short, physical border and metabolic processes are entwined in a cyclic, recursive production network and they together constitute the identity of the system. From this perspective, phenomena like tornadoes, whirlpools, and candle flames, which are to a certain degree self-organising and self-maintaining systems, are not autonomous, because they lack an internally produced physical boundary, and are not concrete topological units. In that sense, what distinguishes self-organisation from autonomy is that the former lacks an internal organisation complex enough to be recruited for deploying selective actions capable of actively ensuring the system's maintenance.

For the purposes of this book, it is worth mentioning two lines of criticism that have been addressed to the theory of autopoiesis. On the one hand, autopoiesis conceives autonomy as a fundamental internal determination, defined by the operational closure between the production network and the physical border. In this model, interactions with the environment do not enter into the definition-constitution of the autonomous system; rather, the interactions with the environment – that Maturana and Varela called "structural couplings" – follow on from the specific internal identity of each autopoietic system. On the other hand, Maturana and Varela

define autonomy in rather abstract and functionalist terms: material and energetic aspects are considered as purely contingent to its realisation.

On both these issues, the framework that we will develop in this book takes a different path. The autonomous perspective, we hold, should take into account the "situatedness" of biological systems in their environment, as well as their "grounding" in thermodynamics. As a matter of fact, these issues have been at the centre of the most recent studies on biological autonomy, by authors like Hooker, Collier, Christensen, Bickhard, Kauffman, Juarrero, and the IAS Research Group,[10] who have stressed that the interactive dimensions of autonomous systems in fact *derive* from the fact that they are thermodynamically open systems, in far-from-equilibrium conditions. As these authors have explained, since the constitutive organisation of biological systems exists only in far-from-equilibrium conditions, they must preserve an adequate interchange of matter and energy with their environment or they would disintegrate. For example, in Kauffman's approach, the main condition required for considering a system autonomous is that it should be capable of performing what he calls "work-constraint cycles" (Kauffman 2000). As Maturana and Varela, Kauffman's account envisages how autonomy can come out of the causal circularity of the system; yet, in his view, this circularity is understood not just in terms of abstract relations of component production but in explicit connection with the thermodynamic requirements that the system must meet to maintain itself.

In accordance with this literature, we will make in this book a conceptual distinction between two interrelated, and yet conceptually distinct, dimensions of biological autonomy: the *constitutive* one, which largely determines the identity of the system; and the *interactive* one, which, far from being a mere side effect of the constitutive dimension, deals with the inherent functional interactions that the organisms must maintain with the environment (Moreno et al. 2008). These two dimensions are intimately related and equally necessary. It might be illuminating to think of the example of the active transport of ions across the cell membrane, required to prevent osmotic crises. The cell can be maintained as long as ion transport is performed, but this interaction can only be carried out because there is a constitutive chemical organisation providing the membranous machinery that does the work. In particular, the emphasis on the interactive dimension implies, as we will stress repeatedly, that autonomy should not be confused with independence: an autonomous system must interact with its environment in order to maintain its organisation[11] (Ruiz-Mirazo and Moreno 2004). As we will discuss in Chap. 4, this is what grounds the agential dimension of autonomy.

[10]The IAS Research Group – to which the authors of this book belong – has been working since the last 25 years on autonomous perspective in biology, while extending it to other fields as cognition, society and bioethics. See also the end of this Introduction and footnote 6 in Chap. 1.

[11]Hooker has recently defined autonomy as "the coordination of the internal metabolic interaction cycle and the external environmental interaction cycle so as the latter delivers energy and material components to the organism in a usable form and at the times and locations the former requires to complete its regeneration cycles, including regeneration of the autonomy capacity" Hooker (2013).

Again, there is a reciprocal dependence between what defines the conditions of existence of the system and the actions derived from its existence: from the autonomous perspective, in Jonas's terms, the system's *doing* and its *being* are two sides of the same coin (see also Moreno et al. 2008). In this view, the environment becomes a world full of significance: facts that from the outside may appear just as purely physical or chemical develop into positive, negative, or neutral influences on the system, depending on whether they contribute to, hinder or have no effect on the maintenance of its dynamic identity. Even the simplest living organism creates a set of preferential partitions of the world, converting interactions with their surrounding media into elementary values, as we will explain extensively in Chaps. 4 and 7. Von Uexküll (1982/1940) called this subjective meaningful world of each organism *Umwelt*. The interactive dimension of autonomy is where the nature of living systems as inventors of worlds with meaning becomes manifest (see also Hoffmeyer 1996). Indeed, this aspect was recognised by Weber and Varela (2002) who argued, following Jonas, that autonomy implies a meaningful relation with the environment.

The autonomous perspective that we develop here endeavours then to grasp the complexity of biological phenomena, by adequately accounting for their various dimensions, specificities, and relations with the physical and chemical domains. As we will discuss throughout the book, our framework differs in many ways from preceding related models, mainly because we aim at – simultaneously – enriching and specifying their central tenets, in close contact with current scientific theories. In the remainder of this introduction, let us give a synthetic overview of the ideas that we will be advocating.

First, the self-determination of the constitutive organisation remains the conceptual core of autonomy. We share with existing accounts of autonomy the idea that biological systems are constituted by a network of causal interactions that continuously re-establish their identity (see also Bechtel 2007). The aim of Chap. 1 will be to provide an explicit conceptual and (preliminarily) formal account of self-determination in terms of what we will label "closure of constraints".

Biological systems determine (at least in part) themselves, we will contend, by constraining themselves: they generate and maintain a set of structures acting as constraints which, by harnessing and channelling the processes and reactions occurring in the system, contribute to sustain each other, and then the system itself. The core of biological organisation *is* the closure of constraints. We will discuss how the concept of closure allows specifying what kind of "circularity" is at work in the biological domain, and how it fundamentally differs from other "process loops" and self-organising phenomena in Physics and Chemistry. In particular, we will emphasise that biological closure requires taking into account, at the same time, the conceptual distinction, and yet inherent interdependence, between two causal regimes: the constraints themselves and the thermodynamic flow on which they act. In the autonomous perspective, closure (of constraints) and (thermodynamic) openness go hand in hand. Self-determination as closure constitutes the pivotal idea on which we will build our account of autonomy. A first step is made in the last section of Chap. 1, where we will claim that biological organisation, to

be such, requires regulation. The long-term preservation of biological organisms supposes the capacity to self-maintain not only in stable conditions but also, and crucially, before potentially deleterious internal or external perturbations. In such circumstances, regulatory capacities govern the transition towards a new viable situation, be it by countering the perturbations or by establishing a new constitutive organisation. In all cases, we will account for regulation in terms of a specific set of constraints, which contribute to the maintenance of the organisation *only* when its closure is being disrupted: accordingly, we will argue that regulatory constraints should be understood as being subject to *second-order* closure.

Does the autonomous perspective require appealing to some form of emergentism? In previous years, some authors have argued that accounts dealing with concepts like self-organisation, closure, constraints, autonomy, and related ones are indeed committed to the idea that biological organisation is an emergent determination. In Chap. 2, we deal with this issue, advocate a monistic stance, and provide a twofold argument. First we argue, against exclusion arguments, that closure can be consistently (with respect to our monistic assumption) understood as an emergent regime of causation, in the specific sense that the relatedness among its constituents provides it with distinctive and irreducible properties and causal powers. Second, although the closure of the constitutive organisation makes sense of the claim that "the very existence of the parts depends on their being involved in the whole", we hold that closure does *not* imply inter-level causation, in the restrictive sense of a causal relation between the whole and its own parts (what we label "nested" causation). Yet, we leave room for appealing to nested causation in the biological domain, if relevant cases were identified and the adequate conceptual justification were provided.

With these clarifications in hand, Chap. 3 addresses the question of the distinctive emergent features of organisms by arguing, in particular, that the closure of constraints provides an adequate and naturalised grounding for the teleology, normativity, and functionality of biological organisation. When closure is realised, the existence of the organisation depends, as we have already emphasised, on the effects of its own activity: accordingly, biological systems are teleologically organised, in a specific and scientifically legitimate sense. Because of teleology, moreover, the activity of the organism has an "intrinsic relevance" which, we submit, generates the norms that the system is supposed to follow: the system must behave in a specific way, otherwise it would cease to exist. Hence biological organisation, because of closure, is inherently normative. And then, by grounding teleology and normativity, closure grounds also functionality in biological organisation: the causal effects produced by constraints subject to closure define biological functions. The general upshot of the analysis, at the end of Chap. 3, will be the deep theoretical binding between "closure", "organisation", and "functionality": it will be our contention that, from the autonomous perspective, they are reciprocally defining concepts, which refer to the very same causal regime.

The constitutive dimension of closure, however, is not autonomy. As mentioned in the preceding pages, autonomy also includes an interactive dimension, dealing with the relations between the organism and its environment. We deal with the

interactive dimensions in Chap. 4, and refer to it as *agency*, characterised as a set of constraints subject to closure, exerting their causal effects on the boundary conditions of the whole system. At the end of the chapter, we argue that a system whose organisation realises closure, regulation, and agency, as defined in the first part of the book, is an autonomous system, and therefore a biological organism. More precisely, Chap. 4 elaborates on a definition of minimal autonomy that captures the essential features of biological organisation in its (relatively) most simple manifestations, typically in unicellular organisms.

What has the autonomous perspective to say about more complex organisations and specifically about multicellular organisms? One of the main weaknesses of the organisational tradition in biology is arguably the fact that it has never explicitly addressed the issue of higher *levels* of autonomy: How many levels of autonomy can be identified in the biological realm, and what are their mutual relations? In Chap. 6, we make a first step in this direction: we try to frame the issue of higher-level autonomy in precise terms and submit some explicit hypotheses on its features. The central idea will be that what matters for higher-level autonomy is *development*. More specifically, multicellular systems are relevant candidates as organisms when their organisation exerts a functional control over the development of unicellular components, so to induce their differentiation which, in turn, makes them apt to live only in the very specific environment constituted by the multicellular system: in a word, the control over development produces the relevant degree of *functional integration* that distinguishes multicellular organisms (as autonomous systems) from other kinds of multicellular systems. What about the relations between levels of autonomy? In spite of their differentiation (and then of the loss of some of their capacities), we will argue that unicellular constituents of higher-level organisms still meet the requirements of autonomy. In fact, the very possibility of higher-level autonomy seems to require that lower-level entities preserve an adequate degree of complexity: multicellular autonomy requires unicellular autonomy. One of the objectives of Chap. 6 (and partly of Chap. 4, last section) will be, by relying on an explicit definition of autonomy, to provide relevant criteria for examining different kinds of higher-level associations and organisations of autonomous systems and to compare them on theoretical grounds. In particular, our framework could allow locating them in a *continuum* of organised systems going from, at one extreme, those cases fulfilling only the requirements for closure (as ecosystems) to systems being progressively more integrated (as the cyanobacterium *Nostoc punctiforme*), up to genuine multicellular organisms (higher-level autonomous systems) at the other extreme.

The transition to multicellular autonomy paves the way towards cognition, which is possibly the most amazing innovation during the evolution of life. Cognition, as discussed in Chap. 7, is much more than a complex form of agency. It is better conceived as a radically new kind of autonomy whose specific features and dynamics go, qualitatively, far beyond multicellular autonomy, opening the way towards our own origins as human beings. In this sense, the analysis of cognition is related to the nuclear problem of the gap between the "biological" (broadly understood) and the "human" domains. Yet, the autonomous perspective strives

to understand and explain cognitive capacities in close connection to a bodily organisation, which is in turn the product of a long evolutionary process, through which new phenomena such as emotions or consciousness – and a world of meaning and values – have been generated. The appearance of cognition is the result of the evolution towards increasingly higher degrees of both constitutive and interactive complexity: in this sense, with all its specificities, cognition is still a "biogenic" (Lyon 2006) phenomenon. By framing the issue of cognition in these terms, we think that it can be better handled in naturalised terms, without underestimating the formidable difficulties that any satisfactory account of cognition has to face to understand its complex nature and phenomenological novelty. Accordingly, Chap. 7 is possibly the most ambitious and yet incomplete, since it sketches in a preliminary way many problems for which much more work will be required.

Autonomy, as conceived in this book, lies at the intersection between different dimensions, and specifically the constitutive and interactive ones, on which we put strong emphasis in the previous pages. Yet, this is not the whole story: autonomy also has a *historical* dimension. As we will discuss in Chap. 5, no adequate understanding of the emergence of autonomous systems (and specifically highly complex autonomous systems, as present biological organisms) can be obtained without taking into account the evolutionary process that brought them about. Autonomous systems are too complex to be spontaneous and cannot self-organise (in the sense of generate themselves) as dissipative systems do: their complexity requires an evolutionary process of accumulation and preservation. Yet, in addition to acknowledging the fundamental place of history in the autonomous perspective, we will submit two related ideas. First, the historical dimension does not have the same theoretical status as the constitutive and interactive ones: while the latter two *define* autonomy, the former does not. The reason is that we do not need history to understand what biological systems are, but rather to understand where they come from: these two questions are of course related, but conceptually distinct. Second, we will restate the relations between selection and organisation, by advocating the general picture according to which the evolution of biological systems stems from the mutual interplay between organisation and selection: this is because, as we will argue, organisation is a condition, and not only an outcome, of evolutionary processes.

Having outlined the central ideas of the book, let us point out that it is, of course, not our intention to develop an exhaustive account of biological autonomy, which would deal with all aspects and implications of the philosophical and theoretical framework. Rather, our ambition is to offer a coherent and integrated picture of the autonomous perspective, by focusing on what we think are some of its central tenets. Much more could (and hopefully will) be written on biological autonomy, but we hope that the ideas of this book can be a useful ground on which future investigations will rely.

This book is the result of a collaboration that goes far beyond that between the two authors. After having promoted (together with Julio Fernandez, Arantza Exteberria, and Jon Umerez), more than 20 years ago, the creation of the *IAS Research Centre for Life, Mind and Society*, at the University of the Basque Country,

in Donostia – San Sebastian, Alvaro Moreno has had since then the chance to work in this highly stimulating intellectual environment. In this respect, a special thought goes to Francisco Varela, who has been a fundamental source of inspiration for the creation of the *IAS Research* group and, for many years afterwards, a close collaborator and a friend.

Matteo Mossio joined the group in 2008 as a postdoctoral fellow and, after having moved back to Paris in 2011, maintains close collaborations with many of its members. Since the constitution of the *IAS Research* group its members have collectively developed the autonomous perspective in the biological, cognitive, biomedical, and ethical domain. The ideas developed in this book, then, are deeply grounded into the substantive and extensive philosophical and theoretical work undertaken by our colleagues and friends.[12]

It then goes without saying that we are intellectually indebted with many people. Let us thank first those who co-authored previous publications with (at least one of) us and allowed us to rework and use in this book some of the ideas advocated there: Argyris Arnellos, Xabier Barandiaran, Leonardo Bich, Maël Montévil, Kepa Ruiz-Mirazo, and Cristian Saborido. At the beginning of each chapter, we inserted a note in which we give the references of the specific publications from which some ideas and text portions have been taken and adapted.

We are sincerely grateful to the other members of the IAS Research Centre for continuous interactions, over the years, on a variety of topics related to this book: Antonio Casado da Rocha, Jesús Ibañez, Hanne de Jaegher, Asier Lasa, Ezequiel di Paolo, and Agustin Vicente. Also, we thank many other researchers with whom one of us (AM) has worked for a long time: Francisco Montero, Federico Morán, Juli Peretó, and more recently, Nei Nunes and Charbel El-Hani.

In Paris, the whole *Complexité et Information Morphologique Team* (CIM), at the Ecole Normale Supérieure, deserves a special mention. Some years ago, Giuseppe Longo created a small but very active interdisciplinary team, nourished by the talent of several young fellows: among them, let us thank with special emphasis Nicole Perret and Paul Villoutreix. We would also like to express our deepest gratitude to Giuseppe, a remarkably brilliant and profound scientist, for his wise guidance on – and unfailing support to – Matteo's academic and scientific trajectory. Recently, Matteo has been invited, together with some of the CIM members, to join the new *Theory of Organisms* research group at the Ecole Normale Supérieure, supervised by Ana Soto. We warmly thank her, as well as Carlos Sonnenschein and Paul-Antoine Miquel, for this unique opportunity to engage in stimulating and quality discussions and exchanges.

Since Matteo's appointment, the *Institut d'Histoire et Philosophie des Sciences et des Techniques* (IHPST) has constituted a privileged scientific environment and

[12]Over the years, the activites of the *IAS Research* group have received funding by both Basque and Spanish public institutions. This book, in particular, was supported by grant IT 590–13 of the Gobierno Vasco, and grant FFI2011-25665 of the Ministerio de Industria e Innovación.

provided him with ideal conditions of work. For that, we want to thank its director Jean Gayon, as well as all the members and colleagues who, in many cases, have become good friends.

Lastly, we are greatly indebted to those colleagues and friends who took the time to read and critically comment on early versions of the manuscript: Philippe Huneman, Johannes Martens, Arnaud Pocheville, and Charles Wolfe. In most cases, their observations and criticisms were decisive to highlight some of the weaknesses of the arguments and force us to improve their clarity and accuracy. In this respect, we owe a lot to Alicia Juarrero, who has not only made a number of precise and lucid comments on various ideas developed in the book but also crucially contributed to bringing the initial unstable language closer to correct English. We also want to warmly thank Juli Peretó for his help in the elaboration of many figures.

The final acknowledgements go to Cliff Hooker: his meticulous, lucid, and uncompromisingly critical reading of the entire manuscript has induced substantial changes (and, we hope, improvements!) in the formulation of the ideas, regarding both the form and the content. Last but not least, he has kindly written the best foreword we could expect.

Donostia – San Sebastian/Paris
October 7th 2014

1
Constraints and Organisational Closure

The first and most fundamental tenet of the autonomous perspective is the idea that the constitutive dimension of biological systems is inherently related to self-determination. As we recalled in the introduction, what constitutes biological systems is the fact that the effects of their activity and behaviour play a role in determining the system itself. As autonomous systems, biological systems "are (at least in part) what they do".

To a first approximation, all accounts of biological autonomy developed during recent decades share this idea, which most of them refer to using the technical term "closure". Despite the differences in existing formulations, the concept of closure aims to ground the intuition about self-determination in a biologically relevant and treatable way. In very general terms, it designates a feature of biological systems by virtue of which their constitutive components and operations depend on each other for their production and maintenance and, moreover, collectively contribute to determining the conditions under which the system itself can exist (Mossio 2013).

The term was first used in the biological domain by Varela in his *Principles of Biological Autonomy* (Varela 1979), and was later adopted by several other authors, including Howard Pattee, Robert Rosen, and Stuart Kauffman, in a similar or complementary sense. Varela's account constitutes here a relevant starting point, since it explicitly establishes a theoretical connection between closure and autonomy through the so-called "Closure Thesis", according to which "every autonomous system is operationally closed" (Varela 1979: 58). Although the thesis does not enunciate an equivalence – which means that, for Varela, closure does not *define* biological organisation (a point on which other authors, such as Rosen, would disagree) – it does put closure at the core of biological organisation, viewing it as a necessary and constitutive feature of autonomy.

Some of the ideas exposed in this chapter, as well as some parts of the text, are taken from (Montévil and Mossio 2015).

A. Moreno, M. Mossio, *Biological Autonomy*, History, Philosophy and Theory of the Life Sciences 12, DOI 10.1007/978-94-017-9837-2_1

Since its formulation, the Closure Thesis has indeed remained a common assumption in the philosophical and scientific tradition centred around biological autonomy, and the concept of closure has being increasingly developed in recent theoretical, computational, and experimental studies (Chandler and Van De Vijver 2000).

Yet, in spite of the current interest in closure as a key notion for understanding biological organisation, it should be noted that no consensus has yet been reached regarding a precise definition. Of course, definitions are not a goal in themselves, and the degree of accuracy which is required may depend on the role played in the general framework. In the case of closure, in our view, the lack of precision does indeed constitute an obstacle for the further development of the autonomous perspective, since closure is a fundamental pillar of the whole account, on which many (or even most) other aspects rely either directly or indirectly, as we will discuss at length in the following chapters.

In particular, the very status of closure as a causal regime with distinctive properties remains somehow controversial since, to date, no explicit account of the relations between closure and other kinds of causal regimes at work in physics and chemistry has been offered. This is a crucial issue since it might be possible that all accounts of biological organisation referring to closure could be reformulated in terms of other physicochemical causal regimes without any relevant information being lost. If this were the case, the concept itself, as well as all other notions relying on it, would have some heuristic value for biological research, but no explanatory role. Consider for instance the central feature of closure, i.e. the mutual dependence[1] between constituents and their collective capacity to self-determine. At first glance, indeed, mutual dependence seems to be by no means a distinctive feature of the biological domain. Let us mention an example that is frequently referred to in this kind of debate, namely, the Earth's hydrologic cycle.[2] Here, a set of water structures (e.g. clouds, rain, springs, rivers, seas, etc.) generate a cycle of causal relations in which each contributes to the maintenance of the whole, and is in turn maintained by the whole. Clouds generate rain, which (contributes to) generates a spring, which gives rise to a river, which (contributes to) generates a lake, which regenerates clouds, and so on. Is this a case of closure?

Arguably, a large number of physical and chemical systems could be described as generating some form of mutual dependence of this kind between their constitutive entities and processes. As a consequence, a coherent account of closure has to choose between two alternative options: either closure is to be conceived as a specific variant of other kinds of causal regimes encountered in physics and

[1] Strictly speaking, "mutual dependence" and "closure" are not synonymous. While the former is realised by any (sub)set of entities which depend on each other, the latter is realised by the set of *all* entities which are mutually dependent in a system. So for instance, the heart and the lungs realise mutual dependence among them, but only the whole set of organs of the organism realises (by hypothesis) closure.

[2] Another, more complex, example is the atmospheric reaction networks, which realise a closed loop of chemical reactions (Centler and Dittrich 2007).

chemistry, in which case the difference between physicochemical and biological systems, in this respect, would possibly be quantitative, but not qualitative; or, alternatively, it might be that closure is qualitatively irreducible to most kind of physical and chemical regimes and dependencies, and so specific to the biological domain.

The aim of this first chapter is to propose a theoretical and formal framework that characterises closure as a causal regime specifically at work in biological organisation. In particular, it will be our contention that biological systems can be shown to involve two distinct, although closely interdependent, regimes of causation: an *open* regime of thermodynamic processes and reactions, and a *closed* regime of dependence between components working as constraints.

1.1 Biological Determination as Self-Constraint

In the introduction to *Toward a Practice of Autonomous Systems* (Bourgine and Varela 1992), restate the Closure Thesis and clarify that they build on an algebraic notion, according to which

> a domain K has closure if all operations defined in it remain within the same domain. The operation of a system has therefore closure, if the results of its action remain within the system (Bourgine and Varela 1992: xii).

Applied to biological systems, closure is realised as what Varela labels *organisational* (or *operational*) closure,[3] which designates an organisation of processes such that:

> (1) the processes are related as a network, so that they recursively depend on each other in the generation and realisation of the processes themselves, and (2) they constitute the system as a unity recognisable in the space (domain) in which the processes exist (Varela 1979: 55).

Varela's account is perhaps the best-known and most influential one within the autonomous perspective. It has several qualities, particularly that of providing a general and abstract characterisation, which can be realised in nature by different kinds of biological systems and sub-systems.[4] In each case, the nature and kind of components and processes subject to closure are different, as is the kind of unity that they generate. In the specific case of the cell, as mentioned in the Introduction, closure takes the exemplary form of autopoiesis (Varela et al. 1974), which is

[3]It should be noted that, over the years, Varela himself has proposed slightly different definitions of operational closure. Also, more recent contributions have introduced a theoretical distinction between organisational and operational closure: whereas "organisational" closure indicates the abstract network of relations that defines the system as a unity, "operational" closure refers to the recurrent dynamics and processes of such a system (see Thompson 2007).

[4]According to Varela, three realisations of closure have been described: the cell, the immune system, and the nervous system (see Varela 1981: 18).

realised at the chemical and molecular level, and involves relations of material production among its constituents. In all cases, organisational closure constitutes the *fundamental invariant* of biological phenomena, in spite of the variability of its concrete realisations.

Despite its qualities, however, we would underscore what we take to be the fundamental weakness of Varela's account of closure. The characterisation described above refers to the processes as the relevant constituents of the system that, when organised in a network, must realise mutual dependence and closure. It seems only fair to point out that, for Varela, closure is understood as *closure of processes*. And here, in our view, is where the problem lies. Formulated in these terms, closure can in principle be used to describe not only the constitutive organisation of biological systems – which are by hypothesis the prototypical example of autonomous systems – but also a number of physical and chemical systems such as, for instance, the famous hydrologic cycle.

To this objection, one may reply that the definition emphasises the "spatial localisation" of the closed unity: Varela and colleagues have in mind the fact that biological systems are clearly recognisable as spatial units, distinguishable from their surroundings. Yet the criterion of being spatially localisable appears to be open to interpretation, and one could easily argue that the hydrologic cycle is a "unity recognisable in the space in which the processes exist".[5] Another relevant response would be to claim that, of course, closure is a necessary but not sufficient condition for autonomy; accordingly, those physical and chemical systems that could possibly be shown to realise closure would not be autonomous. The point is well taken but it reveals, we maintain, that we are dealing with an unsatisfactory characterisation of closure precisely because it applies to biologically irrelevant systems. As we mentioned above, and discuss in much more detail below, closure is a pivotal determination of autonomous systems, and grounds many of their distinctive properties such as, for instance, their individuation, normativity and functionality. Hence, although we would not be compelled to conclude that physical cycles are autonomous, Varela's account would indeed force us to ascribe to them many properties that we would like to apply to autonomous systems, and therefore to biological systems.

Our diagnosis concerning Varela's account of closure is that, although it points in the right direction by emphasising the fact that the organisation of autonomous systems somehow involves a mutual dependence between its components, it *fails to locate closure at the relevant level of causation*.

In our view, closure, as it is realised by autonomous systems, does not involve processes and/or reactions, as is the case for physical and chemical cycles. Instead,

[5] As a matter of fact, some authors have recently argued that the requirement for a *physical* boundary should be replaced by one for a *functional* boundary (Bourgine and Stewart 2004; Zaretzky and Letelier 2002). We agree entirely with this suggestion (see also Sect. 1.6 below), but it should be noted that functional boundaries, given that they are more general, might expose even more closure to the danger of applying to irrelevant systems. The appeal to functional boundaries should then go with a more rigorous definition of closure.

we claim that closure consists of a specific kind of mutual dependence between a set of entities having the status of *constraints* within a system.[6]

What are constraints? In contrast to physical fundamental equations, constraints are local and contingent causes, exerted by specific structures or processes, which reduce the degrees of freedom of the system on which they act (Pattee 1972). As additional causes, they simplify (or change) the description of the system, contributing to providing an adequate explanation of its behaviour, which might otherwise be under-determined or wrongly determined. In describing physical and chemical systems, two main features of the explanatory role of constraints should be emphasised. Firstly, constraints are usually introduced as external determinations (boundary conditions, parameters, restrictions on the configuration space, etc.), which means that they contribute to determining the behaviour and dynamics of a system, even though their existence does not depend on the dynamics upon which they act (Umerez 1994; Juarrero 1999; Umerez and Mossio 2013). To take a simple example, an inclined plane acts as a constraint on the dynamics of a ball resting on it, whereas the constrained dynamics do not exert a causal role in the production and existence of the plane itself. Secondly, in those cases in which some constraints are produced within the system being described, the causal relations between these constraints are usually oriented, in the sense that each constraint may possibly play a role in generating another constraint in the system, although no mutual dependence is realised.

In turn, a distinctive feature of autonomous systems is the fact that, in contrast to most physical and chemical systems, the causal relations between (at least one subset of) the constraints acting in the system generate closure. The general idea behind this account of closure is that the specificity of autonomous systems lies in their capacity for self-determination, in the form of *self-constraint*. But what does this actually mean?

Biological systems, like many other physical and chemical systems, are dissipative systems, which means, in a word, that they are traversed by a flow of energy and matter, taking the form of processes and reactions occurring out of thermodynamic equilibrium. In this respect, organisms do not differ qualitatively from other natural dissipative systems. However, what specifically characterises biological systems is the fact that the thermodynamic flow is channelled and harnessed by a set of constraints in such a way as to realise mutual dependence between these constraints. Accordingly, the organisation of the constraints can be said to achieve self-determination as self-constraint, since the conditions of existence of the constitutive constraints are, because of closure, mutually determined within the organisation itself.[7]

[6]The connection between closure and constraints has been already put forward in the work of authors like Bickhard, Christensen, Hooker, and Kauffman, mentioned in the Introduction. Similarly, substantial theoretical work has been done on this issue by various members of the IAS Research Group over the last two decades.

[7]A terminological clarification: For reasons that will become clearer later on (in particular in Chap. 5), we hold that biological self-determination (as self-constraint) implies specifically "self-maintenance" and not "self-generation". Biological systems maintain themselves but do not

As autonomous systems, biological systems do not realise some sort of "process loop" determined by a set of externally determined boundary conditions; rather, they act on the thermodynamic flow to maintain the network of constraints, which are organised as a mutually dependent network. Hence, the *organisation* that realises closure is the organisation of the constraints, and not that of the processes and reactions. What is lacking in Varela's account of closure is, we hold, the (explicit) theoretical distinction between processes and constraints, and the related ascription of closure to the organisation of constraints.

It is worth noting again that, as Varela himself has repeatedly clarified, closure (and autonomy) is by no means meant to signify the "independence" of the system vis-à-vis the external environment. On the contrary, as (Bourgine and Varela 1992) themselves explain, closure goes hand in hand with *interactive openness*, i.e. the fact that the system is structurally coupled with the environment, with which it exchanges matter, energy, and information. In our account, we ground this crucial point through the distinction between constraints and processes: while biological systems are (by hypothesis) closed at the level of constraints, they are undoubtedly open at the level of the processes, which occur in the thermodynamic flow. Autonomous systems are then, in this view, *organisationally closed* and *thermodynamically open.*[8]

Before characterising in more formal terms the fundamental distinction between processes and constraints, we shall discuss this "thermodynamic grounding" of autonomy in more detail.

1.2 The Thermodynamic Grounding of Autonomy[9]

Autonomy, as we characterise it in this book, is essentially grounded in thermodynamics. Autonomous systems, as mentioned above, are dissipative systems dealing in a constitutive way with a thermodynamic flow that traverses them.

To better understand the relevance of this statement, it is worth recalling that scientific tradition in the field of "general systems theory" has, over the last 50 years, made a clear-cut distinction between informational-organisational aspects and energy-material ones. The distinction is at the core of disciplines like Cybernetics, Artificial Intelligence, Computer and Systems Sciences, and the most recent one,

generate themselves spontaneously (as *wholes*, although of course they do generate some their functional components). In this book, we will then use "self-maintenance" to refer to the specific mode of biological self-determination.

[8] In our knowledge, (Piaget 1967) was the first author who has explicitly expressed the conceptual distinction between organisational closure and thermodynamic openness. The treatment of the distinction developed in this chapter is consistent, we think, with his own conception.

[9] Most of the ideas exposed of this section, as well as some parts of the text come from (Moreno and Ruiz Mirazo 1999).

Artificial Life, which all share the idea that considerations about the material or energy realisation[10] of a system do not affect its "organisational essence". Accordingly, although one has to include "a good deal of ancillary machinery for the real implementation of any material system" (Morán et al. 1997; see also Moreno and Ruiz Mirazo 1999), this would not be significant for modelling the organisation as such. Of course, the physical realisation of living organisation requires a certain layout of material components, interactions, and flows. Yet, the common assumption of all these approaches is that the formal and computational models can legitimately abstract from those aspects without losing their relevancy or explanatory power.

A number of abstract computational models of biological organisation have been proposed over the years, some of them at the very heart of the autonomous perspective. Just to mention a few relevant examples, Varela himself and his collaborators have developed computational expressions of autopoiesis (Varela et al. 1974; McMullin 1997; McMullin and Varela 1997), while Kauffman introduced the notion of "autocatalytic sets" (Farmer et al. 1986; Kauffman 1986). In a different computational context, Fontana's "algorithmic chemistry" generated systems (or "grammatical structures") with self-maintaining properties, expressed in the syntactical framework of the lambda-calculus (Fontana 1992; Fontana et al. 1994). All of these are models of "component production systems", sharing the basic property of self-maintenance, and their goal is to determine what is the minimal architecture of interrelations able to generate that property. In these approaches, the aspects related to energy and matter (dissipation, irreversibility, couplings, currencies, etc.) are assumed to be negligible in order to understand the principles of biological organisation.

Undoubtedly, such an abstract approach has proved to be productive for scientific research. Yet, as several authors have argued (Pattee 1977; Emmeche 1992; Moreno et al. 1994; Moreno and Ruiz Mirazo 1999), the watertight separation between "matter" (i.e. the material basis, including the energy-related aspects) and "form" (the "abstract" organisation) can be misleading when studying living systems. Rather, an adequate understanding of biological organisation should reconcile form and matter, insofar as many fundamental features of biological organisation make sense, in this view, only in relation to the conditions of their realisation in nature.

In this sense, a number of authors (Bickhard 2000; Christensen and Hooker 2000) have emphasised that what matters in this respect is precisely the fact that biological systems are also dissipative systems, so the autonomous perspective must understand biological organisation, first and foremost, in light of its "thermodynamic grounding". In general terms, this is not a new idea. Authors like (Maynard Smith 1986) and (Morowitz 1992) have already observed that the maintenance of the living

[10]The terms implementation and realisation are often used to denote a very similar meaning. However, in a strict sense, an implementation is interpreted as a kind of physical realisation of a given formal organisation which is not unique (i.e., where there are multiple possibilities of realisation of said organisation: e.g., a computer programme can be completely specified in an abstract way and then implemented in various kinds of hardware). Consequently, in this book, when we want to avoid such an interpretation, we shall use the term (physical) realisation.

state requires a constant flow of energy through the system. Either by means of an input of suitable chemical components or sunlight, energy must be supplied to the system and then (part of it) given back to the environment, typically as heat. In particular, the continuous flow of energy and matter through the system is somehow controlled by the system itself, and the way in which this control occurs is the object of disciplines such as biophysics, biochemistry, and physiology.

In accordance with these ideas, it is our contention that the autonomous perspective should integrate such dimensions into its conceptual framework. In general terms, acknowledging the thermodynamic grounding means assigning a key role to the physical magnitude *energy,* and specifically to two basic types of energy transformations (within the system and between the system and its environment): *work* and *heat*. Work is typically related to those transformations of energy that maintain or increase its "quality" (i.e. the energy gets ordered[11] – usually because it is localised[12] – as the result of the transformation and is not completely reusable a posteriori, see Atkins 1984), while heat is connected with those which "degrade" it (the energy gets dispersed, and is no longer recoverable).[13] In more technical terms, work is generated as a result of endergonic-exergonic couplings, which are not spontaneous and absorb and store energy, whereas heat is related to *exergonic* transformations, which are spontaneous and release energy.

What is to be explained, then, is how biological systems manage energy flow so as to maintain themselves, and how the very nature of their organisation is shaped to achieve this goal. If one adopts Atkins' view (Atkins 1984), then the issue can be restated as that of how biological systems succeed in constraining flows of (incoming or internally degraded) energy in order to generate work, which in turn contributes to their maintenance.[14]

The details of how living beings actually carry out this efficacious management of energy can be very complex, but, essentially, there is reasonable agreement regarding the fact that it involves the realisation of *couplings* between endergonic

[11] In a cohesive or coherent movement of constituents in the system, for instance, the classic idea of mechanical work.

[12] Typically, in a molecular bond, related to chemical work.

[13] 'Heat' refers to energy that is disordered relative to the initial state of the current exothermic transition. Only in that situation does the fact of "not being recoverable any more" have a clear meaning. As we see, the concepts of work and heat are defined in terms of possibilities of energy use. But "use" here refers to a functionality which is clearly external to the system; hence, it lacks all significance without the presence of an outside observer. Insofar as living organisms are autonomous systems, we shall have to restate these concepts in such a way that they acquire meaning within the operational framework of the actual system (in Chap. 3, we shall discuss this question in detail).

[14] Actually, (Atkins 1984) does not speak about self-maintenance or biology. Rather, his book is about a general interpretation of thermodynamics, and the asymmetry between heat and work (work as a "constrained release of energy").

and exergonic processes which, in turn, requires that at least two important conditions be met (Moreno and Ruiz Mirazo 1999)[15]:

1. First, the presence of "energy intermediaries" (currencies), which enable the establishment of the couplings, so that the exergonic drive of certain reactions can be exploited or later invested in processes of an endergonic nature (typically, self-construction and repair).
2. Second, the presence of components able to modify the rates of reaction in such a way as to ensure the suitable synchronisation of the couplings. The reason is that processes, in addition to being thermodynamically feasible,[16] also follow specific kinetics. For instance, although the combustion of a glucose molecule is exergonic, its rate of reaction at physiological temperature is such that it would stay stable for ages. Accordingly, one has to take into account not only the amount of time that a reaction – or some other process – requires in order to be carried out, but also (and most especially) the time it needs in relation to other reactions with which it could become coupled. In other words, metabolism necessarily requires the *synchronisation* of a whole set of biophysicochemical processes.

Thermodynamically speaking then, biological systems are self-maintaining organisations in far-from-equilibrium conditions, which means, among other things, that their constitutive structures and relations tend to decay and cannot exist except in the presence of the continuous regeneration of the whole organisation. Self-maintenance then, occurs in spite of the continuous replacement of the material components. Indeed, biological systems may also undergo major structural and morphological changes during their lifetime, due to adaptations, accidental events (injuries, etc.) and, especially, because of development, but they keep their organisational identity whilst undergoing constant change.

1.2.1 Kauffman's Work-Constraint Cycle

(Kauffman 2000) has proposed an account that explicitly states the consequences of thermodynamic grounding on the interpretation of autonomy. His central argument consists of retrieving the classic idea of "work cycle" (as in an ideal thermal Carnot machine),[17] and applying it to the context of biochemical,

[15]On the one hand, the system must couple with some external source of energy (sunlight or chemical energy in the autotrophic case; extraneous organic matter in the heterotrophic one). On the other hand, it is also fundamental that *internal* energetic couplings take place, because this allows certain processes (of synthesis, typically) to occur at the expense of others (degradation), when in principle the former ones alone would not be spontaneously viable.

[16]Feasible in the sense that, when coupled, a *global* decrease of free energy should take place.

[17]Originally, a work cycle is a set of externally controlled processes that takes a thermodynamic system back to its initial state, giving as a result an overall production or consumption of work.

self-maintaining reactions. Specifically, he interprets these work cycles as couplings between endergonic and exergonic reactions, so characteristic in living organisms.

Based on Atkins' ideas about work, conceived, as mentioned, as a *constrained* release of energy (Atkins 1984), Kauffman argues that a mutual relationship between work and constraints must be established in a system in order to achieve autonomy in the form of a "work-constraint (W-C) cycle". The basic idea is simple yet deep: constraints are required to harness the flow of energy (in Carnot's machine, for instance, one needs the walls of the cylinder, the piston, etc.), so that the system can generate work and not merely heat (due to the dispersion of energy). In the case of systems able to determine themselves, these constraints are not pre-given, but rather produced and maintained by the system itself. Hence, the system needs to use the work generated by the constraints in order to generate those very constraints, by establishing a mutual relationship between the constraints and the work.

Accordingly, the work-constraint cycle explicitly constitutes a thermodynamically grounded self-determination, through which a system is able to self-constrain by exploiting *part* of the flow of energy and matter to generate work. Of course, the system is still thermodynamically open, and is by no means independent: it dissipates energy and matter, and has "to take in from a source" and "give away into a sink" in order to stay away from equilibrium (Morowitz 1968).

At least two implications of Kauffman's account should be emphasised.

First, the characterisation of work and constraint depends essentially on the fact that they realise the cycle. Energy is work insofar as it contributes to generating and maintaining constraints that facilitate the suitable endo-exergonic couplings. Constraints are constitutive when they are, at the same time, the *condition* (or one of the conditions) for the renewal of work in the system and the *product* of such work. As we will discuss in detail in Chap. 3, this points to the fact that, from the autonomous perspective, the meaning and value of biologically relevant determinations (such as work, constraints and the related concept of "usefulness") are grounded in the self-determining nature of biological organisation.

Second, the W-C cycle makes it clear that the autonomy of the system inherently involves the contribution of constraints. The cycle is maintained precisely thanks to the action exerted by constraints on the thermodynamic flow, which in turn regenerates the constraints. Kauffman, in our view, was one of the first authors to see not only that any understanding of biological autonomy must acknowledge its thermodynamic grounding, but also, and perhaps more crucially, that such grounding brings into focus the role of constraints exerted at the thermodynamic level.

The typical thermodynamic system would be a gas enclosed in a thermal machine (with walls that can be adiabatic or kept at constant temperature if required) undergoing successive expansions or compressions until it is brought back to its original thermodynamic state. The Carnot cycle in particular is completed through two isothermal and two adiabatic processes, producing *ideally* an amount of work that equals (or is exactly proportional to) the area limited by the lines representing those processes in a pressure-volume diagram.

Yet, although Kauffman's account is a highly relevant step towards an adequate account of autonomy, it suffers from a central weakness, namely that organisational closure implies not only the constraining action exerted on the thermodynamic flow, but also a specific *organisation* among the constitutive constraints. And the work-constraint cycle does not elaborate on the nature of this organisation.

Before explicitly addressing the characterisation of closure, let us first, in the following section, focus in precise theoretical and formal terms on the theoretical distinction between constraints and processes, as well as on two corresponding regimes of causation.

1.3 Constraints and Processes

The claim according to which biological closure is realised by the organisation of constitutive constraints acting on the thermodynamic flow requires a theoretical and formal account of the relations between the two causal regimes involved, and specifically between thermodynamic openness and organisational closure.

In this section, we provide an account of the distinction between processes and constraints (exerted on these processes). Processes refer to the whole set of physicochemical changes (including reactions) occurring in biological systems, which involve the alteration, consumption and/or production of relevant entities. Constraints, in turn, refer to entities that, while acting upon these processes, can be said to remain unaffected by them, at least under certain conditions or from a certain point of view.

We propose to ground the theoretical and formal distinction between processes and constraints in the concept of symmetry. In very general terms, symmetries refer to transformations that do not change the relevant aspects of an object: these aspects are said to be conserved under the transformations. In mathematical approaches to natural phenomena, symmetries are at the core of the constitution of the scientific objects themselves, to the extent that they ground their stability and justify the objectivity of the theories formulated to describe them. In this section, we suggest defining constraints as entities that exhibit a symmetry with respect to a process (or a set of processes) that they help stabilise. Specifically, given a process $A{=}{>}{>}B$ (getting B from A), C is a constraint on $A{=}{>}{>}B$, at a time scale τ, if and only if two conditions are fulfilled. Let us discuss each of them by explaining their meanings and referring to two concrete examples, i.e. the action of the vascular system on the flow of oxygen, and that of an enzyme on a chemical reaction.

I/ The situations $A{=}{>}{>}B$ and $A_C{=}{>}{>}B_C$ (i.e. $A{=}{>}{>}B$ under the influence of C) are not symmetrical by permutation at time scale τ.

We note $C_{A=>>B}$, those aspects of C relevant for $A{=}{>}{>}B$ which, when transformed, alter $A_C{=}{>}{>}B_C$.

This condition requires that a constraint play a causal role in the target process. In formal terms, we express this by saying that the situations with and without C are not

symmetrical, which simply means that they are different, even without considering the constraint itself, but just its effects on the process.[18]

Consider the vascular system. There is an asymmetry associated with the flow of oxygen when considered under the influence of the vascular system (A_C=>>B_C) or not (A=>>B), since, for instance, A_C=>>B_C occurs as a transport (canalised) to the neighbourhood of each cell, whereas A=>>B has a diffusive form. Consequently, the situation fits condition I, which means that the vascular system plays a causal role in the flow of oxygen.

Similarly, there is an asymmetry associated with a chemical reaction when considered under the influence of an enzyme (A_C=>>B_C), or not (A=>>B), since, typically, (A_C=>>B_C) occurs faster than (A=>>B).

II/ A temporal symmetry is associated with $C_{A=>>B}$ in relation to the process A_C=>>B_C, at time scale τ.

A constraint, while it changes the way in which a process behaves, is not changed by (conserved through) that same process. The second condition captures this property by stating that C or, more precisely, those aspects $C_{A=>>B}$ by virtue of which the constraint exerts the causal action, exhibit a symmetry with respect to the process ACB.[19]

Again, consider the examples. A temporal symmetry is associated with the vascular system C with respect to the transformation A_C=>>B_C, since, among other things, the spatial structure of the vascular system remains unaltered at the time scale required, for instance, to accomplish the transport of a set of molecules of oxygen from the lungs to the cells. Hence, the situation fits condition II, which means that the relevant aspects $C_{A=>>B}$ are conserved during the process.

Similarly, a temporal symmetry is associated with the configuration of an enzyme, which is preserved during the reaction.

Since they meet the two conditions, both the vascular system and enzymes can be taken as constraints within the organism.[20] All situations which fulfil conditions I and II will be expressed as C(A=>>B)τ or, in an expanded form (Fig. 1.1):

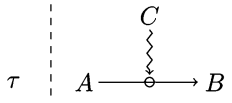

Fig. 1.1 Constraint (*Credits: Maël Montévil*)

[18]The latter precision is important because it would otherwise be trivially true that a situation AB and a situation ACB are different, because of the new object (C) that has been added. Yet, the presence of C does not necessarily change something for the objects present only in the first situation (A and B), since this depends on whether they interact with C in a relevant way.

[19]It is crucial to stress that the conservation concerns *only* these relevant aspects, while other aspects of the entity that exerts the constraint might undergo alteration, even at τ.

[20]The definition of constraint provided above is reminiscent of (and, we think, consistent with) Pattee's account of this concept (see for example, Pattee 1972, 1973). This author defines a constraint as an "alternative description" of the dynamical behaviour of a system, in which a

It is of fundamental importance to emphasise that each condition is met only *at the relevant time scales* and, in particular, that the time scale τ at which conditions I and II must be fulfilled is the same. This means that a constraint, to be such, must conserve its relevant aspects at the same time scale at which its causal action is exerted, even though it may undergo changes and alterations at shorter and/or longer time scales. Consider our two examples. The structure of the organism's vasculature does not change at those time scales at which it channels the flow of oxygen; yet, the structure of the system *does* change at greater time scales due to the effects, for example, of neovascularisation. The same holds for enzymes, which are conserved at the time scale of catalysis, while decaying and randomly disintegrating at larger scales. Moreover, enzymes are also altered at shorter time scales (since they bind with the substrate and lose or gain hydrogen, electrons or protons, etc . . .) and then restored when catalysis is achieved. In spite of the changes at longer and shorter time scales then, constraints are conserved and exhibit a symmetry at that time scale (τ) at which their causal action is exerted.[21]

In most biological cases, a constraint alters the behaviour of a system but does not lead to new behaviours. More technically, the space of possible dynamics of A_C=>>B_C is smaller or equal to the space of possible dynamics of A=>>B, each space being described at the relevant scale (the relevant scales at which the space of possible dynamics of A=>>B and A_C=>>B_C can be described may be very different from τ, and usually they are much longer, possibly infinite). In the case of the vascular system, the flow of oxygen could reach each cell at an adequate rate even in the form A=>>B, i.e. in the absence of the vascular system, from the point of view of statistical mechanics. Hence, the vascular system does not extend the space of possible dynamics of the process A=>>B. In other words, the vascular system is not required, at least in principle, for oxygen to reach the cells at an adequate rate (although the probability of the unconstrained situation occurring is extremely low, at least at biologically relevant time scales, see below). Similarly, an enzyme does not make an otherwise impossible reaction possible, but it does lead to a (possibly far) greater speed of reaction.

---

macroscopic material structure selectively limits the degrees of freedom of a local microscopic system. For an extensive discussion of Pattee's account, see also (Umerez 1994, 1995).

[21]Note that the conservation supposes that a specific time scale τ, at which the target process occurs, *is* to be specified which, in turn, requires determining when the process begins and ends. As a consequence, in those cases in which the process is continuously occurring, discretisation might be necessary to describe the constraints. Let's take the physical example of a river continuously eroding its banks. At first sight, the banks could not be taken, according to our definition, as constraints on the dynamics of the river, precisely because they are transformed by the river. But in fact this description of the system is inadequate, because it fails in specifying the relevant time scale. Although the banks are of course not conserved at the very long time scale at which the entire existence of the river can be described, their *relevant aspects* by virtue of which the river (i.e. a specifiable set of water molecules) moves from a specific point upstream to a specific point downstream in given period of time are presumably conserved during that period. Accordingly, the banks, at that time scale, fit our definition.

It is worth emphasising that the interplay between different time scales allows accounting for an apparent divergence between the idea that constraints are, in many biological cases, theoretically unnecessary, and related analyses of the role of this concept in explaining biological organisation. In particular, as (Juarrero 1999) has pointed out, constraints at work in biological systems are *enabling*, in the sense of being able to generate behaviours and outcomes that would otherwise be impossible. Now, in all those cases in which they do *not* generate new dynamics or behaviours, constraints are *limiting*: they just canalise (condition I) the constrained processes toward a specific outcome among a set of possible ones. Is there a theoretical disagreement? In fact, we think that the distinction between limiting and enabling constraints corresponds to a difference with respect to the time scale at which their causal effects are described. We maintain that, in principle, the constrained dynamics or outcomes could in most biological cases occur in an unconstrained way at the relevant (very long, or infinite) time scale; yet, at *biological* (shorter) time scales, constraints are indeed required for actually getting these specific dynamics and outcomes, because they contribute to the production of otherwise improbable (or *virtually* impossible) effects. In particular, as we will discuss in Sect. 1.5 below, each constitutive constraint of biological organisms enables the maintenance of other constrains and, because of closure, of the whole system. So, although constraints are mostly limiting at longer time scales, they can always be pertinently conceived as enabling at biological shorter time scales: in this sense, it is perfectly consistent with our account to claim that biological organisation could not exist without the causal action of constraints.[22]

Before moving on, let us discuss in some detail the theoretical and epistemological implications stemming from the distinction between constraints and processes. The central point consists in obtaining a description in which biologically relevant entities (the constraints) can be *extracted* from the thermodynamic flow to which biological systems are subject.

Condition II stipulates that the relevant aspects $C_{A=>>B}$ of the constraint are conserved, at τ, as the constrained process continues. In particular, this implies that no relevant flow of matter or (free) energy (or any conserved quantity) occurs between $C_{A=>>B}$ and A=>>B.

Consequently, we submit that constraints can be treated, at τ, as if they were *not* thermodynamic objects because, by definition, they are conserved with respect to the thermodynamic flow, on which they exert a causal action. A description of constraints in thermodynamic terms would be possible in principle, but irrelevant to

[22] At biologically relevant time scales, then, the distinction between constraints and processes roughly maps onto Rosen's distinction between *efficient* and *material* causes (Rosen 1991): constraints might indeed be said to "efficiently" produce an effect by acting, for instance, on the underlying "material" input of a reaction. In spite of this (approximate) correspondence, however, we do not adopt Rosen's terminology, which can be confusing in some respect (see also Pattee 2007 on this point), and will maintain in this book the distinction between constraints and processes. Actually, it might be argued that constraints should rather be intended as "formal" causes (see for example Emmeche et al. 2000; we also briefly discuss this question in Chap. 2).

understanding their causal role, since such a description would show that the flow between $C_{A=>>B}$ and A=>>B is at equilibrium, i.e. no alteration, consumption and/or production would be observed with respect to the constraint. A description of the causal role of constraints in terms of thermodynamic exchanges may possibly be relevant to understanding the intermediate steps leading to the effect (such as, for instance, the sequence of alterations of an enzyme during catalysis), but would be dispensable for understanding the overall effect, which does not involve a flow between the constraint and the constrained process or reaction.

Yet, according to condition I, constraints do play, at τ, a causal role in the process. How is such a role to be conceived in this framework? How can constraints be conserved and yet, at the same time, play a causal role? In our view, constraints do not produce their effects by transmitting energy and/or matter to the process or reaction, but rather by channelling and harnessing a thermodynamic flow, without being subject to that flow. Accordingly, the vasculature channels the blood flow, and the enzyme the reaction (the latter by lowering the activation energy). Even in those cases in which the constraints appear, at first sight, to transmit energy (such as, for instance, the heart which "pumps" blood), the constraint can be pertinently described as a structure which channels a source of energy (in the case of the heart, the free energy available in the cardiac cells) in order to modulate the blood flow. Again, the constraint is conserved; it exploits energy and matter to act on processes and reactions.

The central outcome of the theoretical distinction between constraints and processes consists of the claim that it corresponds to a distinction between two regimes of causation. For a given effect of a process or reaction, one can theoretically distinguish, at the relevant time scale, between two causes: the inputs or reactants (in Rosen's terms, the "material" causes) that are altered and consumed through the reaction, and the constraints (the "efficient" causes, at τ), which are conserved through that very reaction. Constraints are irreducible to the thermodynamic flow, and constitute for this reason a distinct regime of causation.

As mentioned in the previous sections, the distinction and relation between these two causal regimes is a central pillar of any adequate description of biological organisation, specifically as regards its capacity for self-determination. In the following section, we will take a preliminary step towards showing how constraints can realise self-determination in the physical domain.

1.4 From Self-Organisation to Biological Organisation

Self-determination exists in the physical and chemical domain, in the well-known form of self-*organisation*.

A classic example of self-organisation are dissipative structures (Glansdorff and Prigogine 1971; Nicolis and Prigogine 1977), in which a huge number of microscopic elements adopt a global, macroscopic ordered configuration (a "structure") in the presence of a specific flow of energy and matter in far-from-thermodynamic

equilibrium conditions. In turn, the macroscopic configuration exerts a constraint on the microscopic interactions among the surrounding molecules, which contributes to the maintenance of the required flow of energy and matter, and therefore, to the maintenance of the very macroscopic configuration (Ruiz-Mirazo 2001).

A number of physical and chemical systems, such as Bénard cells, flames, hurricanes, and oscillatory chemical reactions, can be pertinently described as self-organising dissipative systems. Let us take the example of "Bénard cells", i.e. macroscopic structures that appear spontaneously in a liquid when heat is applied from below (Chandresekhar 1961). In the initial situation, in which there is no difference in temperature between the upper and lower layers, the liquid appears uniform in terms of the statistical distribution of the molecules' kinetic energy. When heat is applied, and the temperature in the lower layer is increased up to a specific threshold, the liquid's dynamics change dramatically: the random movements of the microscopic molecules spontaneously become ordered, creating a macroscopic pattern (convection cells). In each cell, billions of microscopic molecules rotate in a coherent manner along a hexagonal path, either clockwise or anticlockwise, and always in the opposite direction to that of their immediate neighbours on a horizontal plane.

Bénard cells appear when some specific boundary conditions (e.g. the heat applied from below), which exert *external* constraints on the dynamics of a given set of molecules, are imposed. Yet, once they have appeared, the maintenance of Bénard cells depends not only on these external boundary conditions, but also on the constraint exerted by the configuration itself on its surroundings. For instance, the cells *capture* surrounding water molecules in their dynamics, turning them into constituents. It is through this action that Bénard cells contribute to maintaining the flow of energy and matter traversing them.

Self-organisation, as it occurs in physics and chemistry, constitutes then a case of self-determination, described by appealing to the action of an emergent constraint on the thermodynamic flow. One needs to appeal to the constraining action of the Bénard cell itself in order to find an explanation for its own maintenance, which would otherwise be impossible on the basis of the sole properties of the boundary conditions. The macroscopic constraint determines some of the conditions required for its own existence, and then contributes to its own determination.

Is self-organisation a specific case of closure?

As we have recently emphasised (Mossio and Moreno 2010), dissipative systems realise a *minimal* form of self-determination, in the sense that they generate a *single* macroscopic structure acting as a constraint on its surrounding microscopic dynamics that, in this way, become a part of the system. Accordingly, dissipative systems make a single contribution to their own maintenance, since they contribute to maintaining the single constraint involved in the self-maintaining loop between the structure and the surroundings.

In relation to biological systems, the situation is more complex. In contrast to minimal self-organising systems, biological systems are able to exert a high number of constraints, each of them making a different contribution to the maintenance of the whole.

In doing so, they generate a network of structures, exerting mutual constraining actions on their boundary conditions, so that the whole organisation of constraints realises *collective* self-maintenance. In biological systems, constraints are not able to achieve self-maintenance individually or locally: each of them exists insofar as it contributes to maintaining the whole organisation of constraints that, in turn, maintains (at least some of) its own boundary conditions. This makes a clear-cut categorical distinction between minimal self-organisation and biological closure: while in the first case a single constraint is able to determine itself, in the second case self-determination can only be *collective*, i.e. by contributing to the maintenance of one or several other constraints, each constraint contributes indirectly to its own maintenance, because of mutual dependence (Ruiz-Mirazo and Moreno 2004).

In the following section, we will provide an explicit formal characterisation of organisational closure. Here, we would like to emphasise what is behind the distinction between self-organisation and closure in terms of the underlying complexity and interaction with the environment.

The distinction between self-organisation and closure basically involves the *takeover of (some of) the boundary conditions* required for the maintenance of the system. On the path to autonomy, closed organisations help control several environmental factors, something that requires a degree of internal complexity that simple self-organising systems do not possess.[23]

Of course, autonomous systems do depend on their coupling with the environment; as we stated earlier, autonomy is not independence. Yet, in comparison with self-organising systems, the interaction is different, the degree of dependence is lower and the system is less menaced by external changes. In cases like Bénard convection cells, for instance, small variations in the external conditions, such as the temperature gradient or the inward flow of some substrate, can provoke dramatic changes in the pattern displayed, and may even result in the complete disappearance of the structure. Generally, however, this is not what happens with autonomous systems.

Consider, for instance, the membrane. While self-organising systems are not delimited by a physical border, all biological cells possess a membrane, which not only helps to maintain an adequate internal concentration of materials and nutrients but also, because of its selective permeability, helps control their inward and outward flow. Membranes help distinguish the system from the environment, while at the same time enabling it to act on relevant factors.

The same holds in relation to those internal constraints in charge of the synchronisation of process kinetics and the establishment of global endo-exergonic couplings (see also Morán et al. 1997). The action of catalysers enables autonomous systems to take over the synchronisation of kinetics, which would otherwise depend

[23] As we will discuss at length in Chap. 5, due to the degree of complexity required by autonomous systems, these can only be historical systems (i.e. systems whose complexity has emerged through a cumulative phylogenetic process), and can by no means appear spontaneously (as dissipative structures do).

on very specific (and very unlikely) boundary conditions. Similarly, biological systems, in contrast to self-organising structures, are able to store energy so that, again, they can take over the energy supply for relatively long periods of time and be less affected by a lack of external resources.

In a word, the higher degree of complexity inherent to autonomous systems in comparison with self-organising ones corresponds to a higher degree of self-determination, because of the takeover of boundary conditions over which dissipative structures have no influence or control. The *qualitative* change from minimal (self-organisation) to collective (closure) self-determination goes hand in hand, then, with a *quantitative* increase of the underlying complexity.

One last point should be emphasised here. As we will discuss extensively in Chap. 5, the distinction between self-organisation and closure lies not just in the fact that the latter requires a higher degree of actual complexity than the former, but also in that closure allows for the *potential increase* of functional complexity. Self-organising systems, despite realising a minimal form of self-determination (we will come back to the implications of this point in Chaps. 2 and 3), are not relevant for understanding autonomy, not only because they are "too simple" and categorically different from closed systems, but also because they cannot be taken as a "starting point" for the emergence of closure and autonomy. Closure (and autonomy) is not self-organisation, and neither does it straightforwardly emerge from self-organisation.

We cannot, therefore, understand much about autonomy by looking only at self-organisation as it occurs in physics and chemistry.

1.5 Dependence

Organisational closure occurs in the specific case of mutual dependence between (at least some of) the constraints acting on a biological system. Before discussing closure as such, let us first focus on the relationship of dependence between constraints.

In the previous section, constraints were defined as entities that, among other things, are conserved (symmetrical) with respect to the thermodynamic flow. As specified above, constraints are such only at specific time scales, which means that, at other times scales, they are subject to the thermodynamic flow. In particular, at longer time scales, constraints are subject to degradation and must be replaced or repaired.[24] When the replacement or repair of a constraint depends (also) on

[24]In the case of repair the entity is maintained, while in the case of replacement it is destroyed and reconstructed. Note that the same situation can be interpreted as a case of replacement or repair following the scale at which the constraint is described: individual enzymes are replaced, while the population is repaired. This holds for all those cases (mainly at the molecular level) in which both individual and populations exert the same constraint. See the discussion about scale invariant constraint in Sect. 1.6 below.

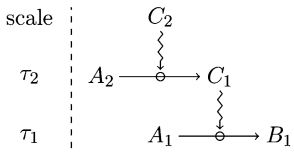

Fig. 1.2 Dependence between constraints (*Credits: Maël Montévil*)

the action of another constraint, a relationship of dependence between the two constraints is established.

As we said, a constraint C_1 is associated with a time symmetry at the scale at which it acts on the process (τ_1 below) and with respect to the relevant aspects for this process, but not necessarily at other scales (τ_2). At the same time C_1, and more precisely the relevant aspects of C_1 (as defined above), can themselves be the product of a process that, in turn, may be constrained by another constraint. This situation leads to the diagram of minimal causal dependence between constraints (Fig. 1.2).

Let us now consider a constrained process C_1 $(A_1 =>> B_1)\tau_1$. Because of condition II, there is a time symmetry at scale τ_1 associated with C_1, which concerns those aspects relevant to the constrained process. At the same time, C_1 is the product of another constrained process $C_2(A_2 =>> C_1)\tau_2$, at a different time scale. At τ_2, C_2 plays the role of constraint, whereas C_1 does not, being the product of the process $C_2(A_2 =>> C_1)$. This situation generates dependence between constraints, where C_1 (the *dependent* constraint) depends on C_2 (the *enabling* constraint, see Sect. 1.3 above). In more general terms, we define a relationship of dependence between constraints as a situation in which, given two time scales, τ_1 and τ_2 considered jointly, we have:

1. C_1 as a constraint at scale τ_1;
2. An object C_2, which is a constraint at scale τ_2 on a process producing aspects of C_1 relevant for its role as constraint at scale τ_1 (which would not appear without this process).

As a simple example, consider the case of an enzyme acting on the reaction that it catalyses at some time scale τ. At longer scales, enzymes are subject to degradation and are replaced by the cell via the translation process, on which ribosomes and mRNA (the DNA sequence being, in turn, a constraint on mRNA) act as constraints. Hence, dependence between constraints holds between enzymes on the one side, and ribosomes and mRNA on the other.

Several important clarifications are required here.

First, the relationship of dependence that is relevant for biological closure must be a *direct* one. This specification is necessary because the definition given above would otherwise apply to a wide range of relationships between constraints, including those in which the enabling and dependent constraints are linked through very long chain of processes. Consequently, dependence would cover many biologically irrelevant situations. Hence, we restrict the relevant meaning of dependence to direct dependence, i.e. a situation in which, considering the different processes that occur

at τ_2 and contribute to maintaining a relevant aspect of C_1 that depends on C_2, there is no process starting after the one constrained by C_2. For example, if we consider an enzyme formation, the maturation of the protein can be successively constrained by a chain of structures. In fact, the catalytic capacity depends directly only on the constraint acting on last process involved, which determines the conformation of the protein or, more precisely, its ability to react with other chemicals. Accordingly, the population of mRNA, as discussed above, is a constraint on the production of protein, but contributes indirectly to their conformation.

Second, dependence between constraints is logically different from dependence between processes. Indeed, at τ_2, where C_2 plays the role of constraint, the conservation of C_2 implies that no thermodynamic exchange occurs between the constraint and the constrained process, and therefore between C_2 and C_1 (see Sect. 1.3 above). In contrast, at scales other than τ_2, the relationship between constraints may involve a thermodynamic exchange, but these exchanges do not interfere with the causal dependence described at the relevant scale. At scales shorter than τ_2, exchanges are possible but irrelevant, since these exchanges would *in fine* be compensated at τ_2, at which C_2 is conserved. At scales longer than τ_2, the interaction between C_2 and $A_2 => C_1$ might contribute to the degradation of C_2; but in the case of biological systems that degradation would be also irrelevant, since C_2 is replaced or repaired by the organisation, because of closure.

Third, in most biological cases, $A_2 => C_1$ does not require C_2 in order to occur, at least at the very large scale of the possible evolutions of $A_2 => C_1$. In contrast, C_2 is required, at the specific scale τ_2, to actually observe the production of C_1. The appeal to different time scales therefore allows us to circumvent the apparent contradiction between the claim that a constraint is conserved through and unaffected by the thermodynamic flow, and the fact that it depends on another constraint.

With the concept of dependence in hand, we can now turn to closure.

1.6 Closure

Closure is a specific mode of dependence between a set of constraints. In very general terms, it refers to all those cases in which, instead of having a linear chain of dependence relationships between constraints, the chain folds up and establishes *mutual* dependence.[25]

In formal terms, a set of constraints **C** realises closure if, for each constraint C_i belonging to **C**:

1. C_i depends directly on at least one other constraint of **C** (C_i is dependent);
2. There is at least one other constraint C_j belonging to **C** which depends on C_i (C_i is enabling).

[25] See also note 1 above on the conceptual relations between "closure" and "mutual dependence".

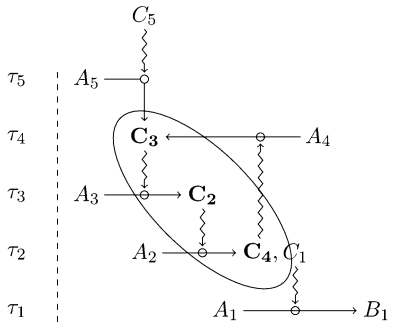

Fig. 1.3 Closure of constraints (*Credits: Maël Montévil*)

Closure refers then to an organisation in which each constraint is involved in at least two different dependence relationships in which it plays the role of enabling and dependent constraint, respectively. The network of all constraints, which fit the two requirements, is – we hold – collectively able to self-determine (or, more specifically, self-maintain; see note 5) through self-constraint.[26]

As a very general abstract illustration, consider the network of dependent constraints shown in Fig. 1.3.

In Fig. 1.3, C_1, C_2, C_3, C_4 and C_5 satisfy, by hypothesis, the definition of constraint at τ_1, τ_2, τ_3, τ_4 and τ_5 respectively. Furthermore, C_1, C_2, C_3 and C_4 play the role of dependent constraints, while C_2, C_3, C_4 and C_5 are enabling constraints. The subset that includes those constraints that are both enabling and dependent is then (C_2, C_3, C_4). The organisation constituted by C_2, C_3 and C_4 realises closure.

This definition of closure is, of course, very general and, as we will discuss in the following section, too schematic to capture the complexity of its actual realisations in biological systems. Yet, it is precise enough to illustrate some of its implications.

[26]The relations brought about by constraints responsible for closure in living systems have received two characterisations by Howard Pattee, in different stages of his work: *statistical* closure (1973) and *semantic* closure (Pattee 1982). By "statistical closure" he (1973: 94–97) means a collection of elements that may combine or interact with each other individually in many ways, but that nevertheless persists as the same collection largely because of the rates of their combination. This in turn implies a population dynamics for the elements and therefore a real-time dependence. Furthermore, the rates of specific combinations of elements must be controlled by collections of the elements of the closed set. The adjective *statistical* refers to the "selective loss of detail" of a statistical classification presents in relation to the underlying dynamics. It explains, according to Pattee, the nature and function of control constraints within a hierarchical system.

In turn, Pattee defines "semantic closure" as follows: "We can say that the molecular strings of the genes only become symbolic representations if the physical symbol tokens are, at some stage of string processing, directly recognized by translation molecules (tRNA's and synthetases) which thereupon execute specific but arbitrary functions (protein synthesis). The semantic closure arises from the necessity that the translation molecules are themselves referents of gene strings." (Pattee 1982: 333). Semantic closure is then based on the idea of symbolic records that preserve those constraints, and of how they are interpreted within the living system as a whole (Umerez 1995; Etxeberria and Moreno 2001).

Firstly, it is worth noting that, in this definition, constraints subject to closure can be interpreted both as *individual* entities or *classes* of entities. In biological systems, constraints are exerted by entities that can be described at different spatial, temporal, or organisational scales. In general, an entity exerting a constraint at a given scale also contributes to the constraints exerted, at different scales, by the larger entities of which it may be a part. For instance, an enzyme working as a catalyst in a cell could also contribute to the function of pumping blood (a different constraint) if the cell belongs to a cardiac tissue. Hence, constraints are not usually scale invariant, insofar as entities described at different scales do not exert the same constraint. Yet in some cases, a constraint might indeed be scale invariant. For instance, an individual enzyme and a group of enzymes exert the same catalytic constraint at different scales, and the same function can be ascribed both to each individual and to the group as a whole. In this case, since both the individuals and the group fit the same characterisation of "constraint" and are subject to closure, it is legitimate to claim that individual entities may be (at least to some extent) redundant,[27] since the network of causal dependencies between constraints would not be affected or altered in the event of breakdown or suppression. The constraint would still be performed by the group at a higher scale. Accordingly, the definition of closure given above covers the case in which some constraint is exerted, in a given system, by an individual entity (say: the heart) as well as those cases in which it may be exerted by an individual *or* a collection of entities.

Secondly, closure of constraints is irreducible to the underlying open regime of thermodynamic processes and changes. As discussed in Sect. 1.3, individual constraints are irreducible to the thermodynamic flow, each constraint being conserved at the relevant time scale. Hence, a reductive description of closure in terms of the causal regime of thermodynamic changes would be inadequate, since it would be unable to include constraints as such and their contribution as causal factors.[28] In particular, a description of biological organisation which does not appeal to the causal power of constraints and their closure would amount to a system constituted by a cluster of *unconnected* processes and reactions, whose coordinated occurrence would be theoretically possible at very long time scales (see discussion in Sect. 1.3), but extremely unlikely (virtually impossible) at biologically relevant time scales.[29]

[27] Scale-invariant constraints may be realised in the form of both *redundancy* or *degeneracy* of functional parts. As (Tononi et al. 1999) have pointed out, redundancy refers to the situation in which structurally similar elements produce the same effects, whereas degeneracy occurs when structurally different elements perform the same function.

[28] It is, of course, conceivable that a description of constraints might possibly be given in terms of thermodynamics, specifically as entities *that are not affected* by the thermodynamic flow. However, in this case, constraints (and hence closure) would not be reduced to a different causal regime, but simply re-described in different terms.

[29] This implication allows us to distinguish between a closure of constraints and a cycle of processes or reactions such as, for instance, the hydrologic cycle mentioned in the introduction to this chapter. In this case, the entities involved (e.g. clouds, rain, springs, rivers, seas, clouds, etc.) are connected to each other in such a way that they generate a cycle of transformations and changes between

Thirdly, as mentioned in Sect. 1.1 above, as a dimension of autonomy, closure should be carefully distinguished from independence, since a system that realises closure is a thermodynamically open one, inherently coupled to the environment. Among other things (discussed at length in Chap. 4), this implies that closure is a *context-dependent* determination, to the extent that it is always realised with respect to a set of specific boundary conditions, which include several external (and independent) constraints acting on the system. Consequently, closure does not, and cannot, include all constraints with which the system may have a causal interaction, but only the *subset* of all those that fit the definition above.

Fourthly, we understand – in accordance with Maturana and Varela – closure as a general invariant of biological organisation. Whatever its specific architecture may be, the organisation of a biological system realises closure between a subset of the constraints acting on it. Constraints subject to closure *constitute* the biological organisation and, accordingly, make an essential contribution to determining the identity of the system. Biological individuality, we think, has much to do with organisational closure, to the extent that one may conjecture that closure in fact defines biological individuality. Although this claim would require a full-fledged argument (that we leave for a future work) we do hold that, by relying on closure, the autonomous perspective clearly favours (as other authors has pointed out, see note 4) functional criteria over physical ones to define the boundaries of biological organisms. In Chaps. 4 and 6, we will make some preliminary steps to apply this view to both unicellular and multicellular organisms.

Fifthly, closure is the fundamental *principle of order* of biological phenomena, which underlies the stability of each biological system and controls the transitions and modifications that the said system undergoes over time. Many different sources of the various kinds of ordered biological patterns can of course be described; yet, what generates the distinctive order of biological organisation as a whole are – fundamentally – the principles governing the integration and coordination of its constitutive constraints in the form of closure. Accordingly, the autonomous perspective can be said to make a significant departure from molecular biology, in the sense that it advocates a shift of focus from genetic information to organisation itself as the central source of order of biological phenomena. Although, as we will discuss in more details in Chap. 5, the autonomous perspective obviously acknowledges that genetic mechanisms do play a crucial role in generating and maintaining biological organisation, it also takes an explicit *holistic* stance in claiming that the role of these mechanisms is to be understood in the light of their contribution to the whole system, the latter being governed by organisational principles.

---

them. In turn, these entities do *not* act as constraints on each other (among other reasons, precisely because they are transformed when they produce another water structure), and the system can be adequately described by appealing to a set of external boundary conditions (soil, sun, etc.) acting on a single causal regime of thermodynamic changes (see also Mossio and Moreno 2010).

Lastly, it is crucial to point out that the invariance of closure by no means implies that biological systems are not subject to variability (specifically functional variability), or that variability is not a central aspect for understanding biological systems. In our framework, closure is described by considering a temporal interval that is wide enough to encompass all constraints and their dependencies. In this sense, the organisation of mutual dependencies is described by abstracting from the physical time in which they occur. In this formal framework, the claim according to which closure constitutes an invariant of biological organisation means that a description of closure is possible for any temporal interval that is wide enough to encompass all constraints and dependencies. In other words, given a minimal interval in the thermodynamic time, closure is realised for whatever interval chosen within the system's lifetime.[30]

At the same time, biological systems may (and do) undergo changes in their organisation throughout their life. Of course, the kind of changes that are relevant here are functional changes, i.e. (as we will discuss in Chap. 3) changes involving one or more constitutive constraints. Let us emphasise that functional variations are not only a contingent fact of biological organisation but also, in many cases, a crucial requirement for adaptivity, the increase in complexity and, in the end, the long-term sustainability of life (see Longo and Montevil 2014, for an original analysis). In Chap. 5 we will discuss these issues at length. Here, it is our contention that, as biological systems undergo continuous and even inherent functional variations, *their organisation maintains closure, albeit realised in different variants*, by adding or suppressing specific constraints or sets of constraints.

Closure is a sort of organisational general invariant: it is the common property of each specific organisation that an individual system may instantiate.

1.7 A Word About Related Models

The definition of closure that we have proposed in the previous section is closely related to a number of models and proposals regarding biological organisation that have been developed during recent decades, mainly in the field of Theoretical Biology. In previous sections, we have discussed two of them in some detail, namely the theory of autopoiesis, which is also the best-known one, and the work-constraint cycle. We have also emphasised the intellectual debt that we have to Howard Pattee, specifically in relation to his work on the concept of constraint and its role in biology. In this section, we would like to say a word about some other accounts, that, of course if dealt with properly, would deserve a full-fledged analysis.

[30]See also (Montévil and Mossio 2015) for more details in this issue. In Chap. 3, Sect. 3.3.1 below we will explore the issue of the temporal boundaries of closure, when these go beyond the lifespan of an individual organism.

A first line of research has been developed by those authors who have suggested formulations of organisational closure in more chemically "realistic" terms. One example is "reflexive catalysis", defined as a gang of molecules each exerting some catalytic function so that, as a net result, the incorporation of all members of the gang is ensured by the gang itself (Szathmary 2006). A very similar concept was proposed by Stuart Kauffman in the 1980s (Kauffman 1986), under the term "catalytic closure", which refers to the mutual dependence between a set of catalysers, each of which constrains a chemical reaction that, in turn, produces at least another one of the set. Although a number of computational models and simulations of catalytic closure have been developed over the years, it should be emphasised that, to date, no chemical realisations have been obtained, which shows to what extent even minimal instantiations of closure require a non-trivial degree of complexity.

In this sense, a very relevant contribution has been made by the Hungarian biologist Tibor Gánti through his model of the *chemoton* (Gánti 1975/2003). The chemoton consists of three functionally dependent autocatalytic subsystems: the metabolic chemical network, the template polymerisation and the membrane subsystem enclosing them all. The correct functioning of the chemoton relies on the precise stoichiometric coupling of the three subunits. The most important of these cycles is the first one because it transforms the chemical energy of nutrients into useful work and constitutes the material support for the other two subsystems. The compartment isolates the autocatalytic subsystem, ensuring an adequate concentration of components and making a certain selection in the transport of matter between the environment and the system. As in the case of autopoiesis, the chemoton creates its own membrane: the metabolic cycle generates not only more intermediates of the cycle but also components of the membrane. Moreover, the chemoton includes the capacity for self-reproduction, since the dynamics of both the compartment and the inner components evolve, doubling their initial value and leading to the subsequent division into two identical chemotons. Lastly, Gánti added a third subsystem – the "template cycle" – to ensure a kind of "control" or "regulation" of the other two dynamically coupled subsystems. It is the length of the polymer that matters in the regulation of the other two subsystems, since it affects the replication rate. The role of the template has nothing to do with any informational control of present-day cells; rather, it is more like a kind of "buffering" system, acting as a "sink" soaking up the waste products of the metabolism, and so affecting the metabolic rate (we shall discuss this point further in the next section). In sum, the combination of the three subsystems gives rise to what Gánti characterises as a supersystem, displaying biological features. In this way, Gánti defines a threshold of minimal tasks and avoids trivial forms of self-maintenance (Fig. 1.4).

Although Gánti does not explicitly refer to constraints closure, the chemoton is undoubtedly an example of closure of constraints because (at least) the membrane that contains the reaction network and the catalysts driving these reactions operate as constraints, and they depend on each other. Moreover, since the length of the template also acts by affecting the rates of other basic processes, it fits our characterisation of a constraint as well. And in turn, this constraint is also dependent on the other constraints. Hence, the chemoton fulfils the criteria of an

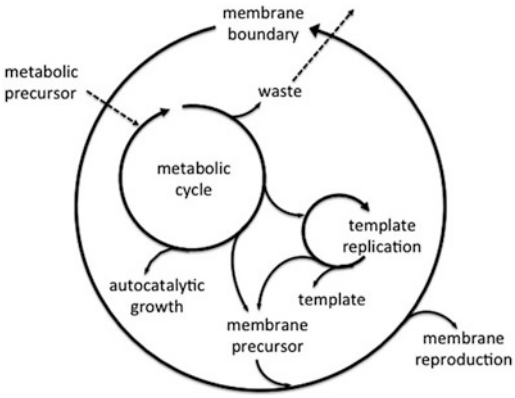

Fig. 1.4 Scheme of Gánti's Chemoton with the three coupled cycles *(credits: Juli Peretó)*

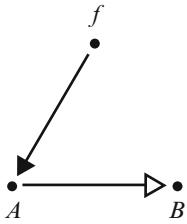

Fig. 1.5 Rosen's distinction between efficient and material causes

organisationally closed system, and provides very relevant insights into the degree of chemical complexity that even its minimal realisations must attain in order to show relevant biological features.

The second line of research is that established by Robert Rosen, and currently being developed by several authors (see for instance Letelier et al. 2003, 2006; Cárdenas et al. 2010; Piedrafita et al. 2010). Rosen's account is complex and profound, and aims to provide a conceptual, theoretical, and formal characterisation of the general principles of biological organisation (as well as of the modelling relationship itself, although we will not discuss this aspect of his work here). Although he was probably not the first author to have used the term "closure" to refer to a distinctive property of biological systems, he was certainly the first one to have explicitly seen and claimed that a sound understanding of closure in biological organisation should make the distinction between two causal regimes at work in biological systems and should locate closure within the relevant regime. In this sense, we acknowledge the intellectual debt we owe to Rosen's work, and see our work, in many ways, as an attempt to further develop his ideas and insights.

As Rosen's account has been developed over 40 years, we refer here only to his latest contributions, and in particular to his book: *Life Itself* (Rosen 1991). As described in that volume, his account is based on a rehabilitation and reinterpretation of the Aristotelian categories of causality and, in particular, on the distinction between efficient and material cause. Let us consider an abstract mapping f between the sets A and B, so that $f: A=>>B$. Represented in a relational diagram, we have Fig. 1.5.

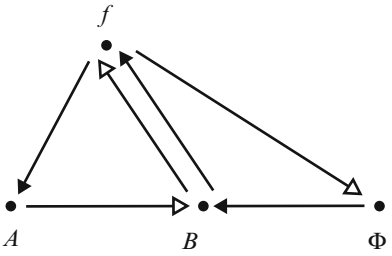

Fig. 1.6 Rosen's closure to efficient causation

When applied to model natural systems, Rosen claims that the hollow-headed arrow represents material causation, a flow from *A* to *B*, whereas the solid-headed arrow represents efficient causation exerted by *f* on this flow.

Rosen's central thesis is that "a material system is an organism [a living system] if, and only if, it is closed to efficient causation" (Rosen 1991: 244). In turn, a natural system is closed to efficient causation if, and only if, its relational diagram has a closed path that contains all the solid-headed arrows. According to Rosen, the central feature of a biological system is the fact that all components having the status of efficient causes are materially produced by and within the system itself. At the most general level, closure is realised in biological systems between three *classes* of efficient causes corresponding to three broad classes of biological functions, which Rosen denotes as *metabolism* (f: A=>>B), *repair* (Φ: B=>>f) and *replication* (B: f=>>Φ) (Fig. 1.6).

By providing a clear-cut theoretical and formal distinction between material and efficient causation, Rosen, as mentioned above, explicitly distinguishes between two coexisting causal regimes: closure to efficient causation, which grounds its unity and distinctiveness, and openness to material causation, which allows material, energy, and informational interactions with the environment. Clearly, the distinction between constraints and processes maps onto the distinction between efficient and material causes. As a matter of fact, as Pattee discusses in a recent paper (Pattee 2007), in his previous work Rosen himself used to use a terminology that was closer to the one adopted in this book.

An analysis of Rosen's account in all its richness would far exceed the scope and limits of this book.[31] Here, we would simply like to mention a specific point about which we believe Rosen has proved particularly insightful. As mentioned earlier, closure to efficient causation is realised by and between very general classes of functions, not just between individual constraints. Accordingly, Rosen's closure occurs at a *higher level of description* with respect to ours, and the relevant question is why Rosen chose to define closure at that level, and what are the implications of that choice. To the best of our knowledge, no clear answers have

[31]We made a contribution in Mossio et al. (2009a), in which we analysed one of Rosen's claims, according to which closure to efficient causation has non-computable models. (Cárdenas et al. 2010) offers a detailed reply to our analysis.

yet been provided to these questions, although in recent times, some studies have taken important steps in this direction. In particular, Letelier and co-authors have published an analysis of Rosen's account in which they propose an interpretation of its central features, and focus in particular on the biological meaning of the classes of function subject to closure (Letelier et al. 2006). In their view, Rosen's labels are somehow misleading, and they suggest using "metabolism", "replacement", and "organisational invariance". They discuss at length the last class of functions, and claim that Rosen's central result was the mathematical demonstration that a system endowed with metabolism and replacement functions can also be inherently organisationally invariant.

Without entering into any mathematical detail, we would simply like to emphasise that, in our view, Rosen's account touches on a crucial issue related to the principles of biological organisation, namely the fact that we should be able to understand its invariance through time, and that said understanding requires an appeal to higher-order closure and organisation. The connection between the invariance (or at least stability) of organisation and the hierarchical nature of constraints and closure is, we believe, a key topic for future research in the field of theoretical biology, especially from the autonomous perspective. In the following section, we make some preliminary headway in this direction.

1.8 Regulation

The characterisation of closure offered in the previous section is extremely general, and aimed at covering all its possible concrete realisations in nature. Any system realising closure, we submit, has to fulfil the above characterisation. Yet concrete realisations of closure require a minimal degree of complexity in order to be not only possible, but also biologically relevant. Indeed, not all closed systems belong to the biological domain, since some viable closed networks may not possess biologically relevant properties or features.

In this section, we focus specifically on this issue, and try to clarify how such "biologically relevant" properties and features should be understood. Furthermore, it will be our contention that those realisations, which are complex enough to be biologically relevant, provide the theoretical groundwork for understanding *metabolism*, interpreted within the framework of the autonomous perspective.

Under what conditions, then, can actual instantiations of closure be taken as "metabolic"? Usually, the simplest realisations of closure are conceived in the chemical domain, in the form of catalytic closure, as briefly discussed in the previous section. The simplest option, of course, is to claim that any minimal chemical network is *ipso facto* metabolic, providing it realises catalytic closure. As a matter of fact, a number of physicists and chemists, typically interested in the origins of

life, have characterised minimal metabolism in a very simplified way,[32] as a closed network of reactions, typically driven by pre-enzymatic catalysts.[33] Interestingly, these networks are usually referred to as "proto-metabolisms[34]" meaning that, although they are very simple, there is a fundamental organisational continuity between them and fully-fledged metabolisms. What constitutes metabolisms is already there, although *in nuce*, in proto-metabolisms.

Now, present-day metabolisms, however simple, show a rich organisational diversity. Actually, in the prokaryotic world, the diversity of metabolisms is truly astonishing. Therefore, it is only logical to consider that the concept of metabolism should somehow imply the capacity to harbour, at least potentially, an indefinite organisational diversity. In fact, the kind of organisation that constitutes the core of biological systems implies a capacity to potentially enlarge indefinitely the number of constraints and therefore, its complexity; otherwise the system would not have the capacity to evolve in an open way.

Yet, as we shall explain, minimal conceivable realisation of organisational closure, as we have already described it, does not ensure this capacity. The central point is that systems realising closure do not necessarily possess the capacity to compensate for variations, be they internally or externally generated. Consequently, variations (such as, for instance, changes in component concentrations in one reaction) can affect the output of a specific constitutive constraint, which in turn may affect the structure and activity of other constraints, and so on. Because of this "transmission of variation", due to the closure between constraints, the organisation may progressively "drift" and, most likely, become disrupted after a short time. Moreover, given the "delicate balance" (see below) between the constituents, the more the organisational complexity increases, the more crucial the capacity to compensate for variations becomes. As an example, take the case of Gánti's chemoton, which can be disrupted even by slight variations due to perturbations in the environment. As (Bechtel 2007) pointed out:

[32]It should be mentioned that there is another conception of minimal metabolism, typically put forward by biochemists (see for instance Gil et al. 2004), according to which it is the characterisation of "minimal genomes" through the simplification of existing ones, under the assumption that their associated metabolic networks will drastically reduce the complexity of extant metabolisms. Here, minimal metabolisms are still "genetically-instructed metabolisms", similar (although highly simplified) to those realised by fully-fledged living organisms. As discussed by (Morowitz 1992) and (Morange 2003), one of the problems of this conception seems to be that, since metabolic simplicity depends on the environment, it is highly problematic to elaborate shared criteria to determine what *the* minimal metabolic network actually is.

[33]For example, (Eschenmosser 2007: 311), writes that "another type of reaction loop that can emerge as a consequence of the exploration of a chemical environment's structure and reactivity space is one that, driven by the free energy of starting materials, connects intermediate products (substrates as opposed to catalysts) in a cyclic pathway: such a cycle is referred to as autocatalytic *metabolic cycle*."

[34](De Duve 2007) uses the term "proto-metabolism" to denote those chemical networks driven by catalysts that, whatever their nature, cannot have displayed the exquisite specificity of present-day enzymes and must necessarily have produced some sort of "dirty gemisch".

> Imagine the environment changed so that a new metabolite entered the system which would react with existing metabolites, either breaking down structure or building a new additional structure. This would disrupt the delicate balance between metabolism and membrane generation that Gánti relies on to enable chemotons to reproduce. What this points to is the desirability of independent control of different operations within the system (p. 299).

As we will see in Chap. 5, variations and the transmission of variations play a fundamental role in enabling the generation of novelty in biological systems, and contribute to their long-term evolution. Yet, as we claimed in Sect. 1.6 above, variation cannot play any evolutionary role unless it can be governed by biological organisation, in order to guarantee its stability while at the same time enabling it to integrate novelty. Biological organisation must be able to handle variations, and then conserve closure, otherwise it would be extremely fragile and its realisations in the natural world would hardly move beyond a very low level of organisational complexity. Any perturbation would be more likely to drive the system to disruption than to result in an increase of complexity. What is then required for biological organisation not only to remain stable in the face of perturbations, but also be able to increase its complexity? The answer is, we submit, *regulation*. Biological autonomy requires regulated closure.

When a set of constraints realise closure, the collective maintenance of the organisation lasts for as long as the activity of each constraint stays within admissible ranges and, moreover, adequate external boundary conditions (on which, as we will see in Chap. 4, the system has a partial influence) persist. If a variation occurs,[35] the system drifts and (possibly, see below) collapses unless it possesses some additional capacity enabling it to respond to the variation, and compensate for its effects.[36] How can closed organisations handle deleterious variations? Let us leave aside those local cases that we could label "local robustness", i.e. the capacity of a *single* constraint or structure to compensate for a perturbation, without altering its behaviour or its causal effects. In this case, the perturbation is handled and compensated for locally, and does not produce a variation affecting the relations which exist between the constitutive constraints; in this sense, it is irrelevant from the point of view of the whole organisation, and as such, is not included in our discussion. In contrast, we focus on those variations that do alter the activity of local

[35] We focus here on deleterious variations, i.e. variations that do not lead to new viable organisations and would disrupt the system if not compensated.

[36] It should be noted that, in some cases, variations may be neutral with regard to the self-maintenance of the system: in spite of the variation, the system may drift, but closure is conserved. And it might be the case that the biological system exerts a form of compensation even on this kind of harmless variation, counteracting its effects. In what follows, however, we shall not discuss these forms of compensation because they are negligible with respect to maintaining closure, which is, after all, the main reason for requiring regulation. Regulation will then specifically be characterised in relation to cases of "deleterious" variations that disrupt closure: a gap is generated between the conditions of existence and the activity of the system, which is no longer able to meet those very conditions of existence, and is therefore destined to collapse.

constraints, thus calling for a response by the global organisation (Barandiaran et al. 2009). Regulatory capacities are about these global responses.

In order to understand how these systems deal with perturbations, we shall introduce a distinction between two general ways self-maintaining systems deal with environmental variations and/or, in general, perturbations. Examples of the first way can be grouped under the general label *constitutive stability* against some range of perturbations. The second way is *regulation (or adaptive regulation)*.

1.8.1 Constitutive Stability

Constitutive stability is the capacity of the whole biological organisation to respond to, and compensate for, variations, thanks to the specific structure of the network of constraints, which might for instance instantiate loops of negative feedback. A variation affecting a given constraint can propagate within the system, and produce the variation of one or several other constraints that in turn compensate for the initial one. As a result, the system is stable, homeostatic.

One way of thinking about a primitive form of constitutive stability was proposed by (Deamer 2009), in a discussion concerning the origin of Life:

> No one has yet attempted to develop an experimental system that incorporates all of the above components and controls, so one can only speculate about how control systems might have developed in early forms of life. One obvious point in the network offers a place to start. Small nutrient molecules must get across the membrane boundary, and so the rate at which this happens will clearly control the overall process of growth. I propose that the first control system in the origin of life involved an interaction of internal macromolecules with the membrane boundary. The interaction represents the signal of the feedback loop, and the effector is the mechanism that governs the permeability of the bilayer to small molecules. As internal macromolecules were synthesized during growth, the internal concentration of small monomeric molecules would be used up and growth would slow. However, if the macromolecules disturbed the bilayer in such a way that permeability was increased, this would allow more small molecules to enter and support further growth, representing a positive feedback loop. The opposing negative feedback would occur if the disturbed bilayer could add amphiphilic molecules more rapidly, thereby reducing the rate of inward transport by stabilizing the membrane. This primitive regulatory mechanism is hypothetical, of course; however, it could be a starting point for research on how control systems were established in the first forms of life.

Another example of constitutive stability is the chemoton itself. In this kind of system, disequilibrations are compensated for through the mutual interaction of stoichiometrically-coupled subsystems, that act by affecting the concentrations of the products of the distinct cycles, in such a way that the rates and speed of reactions inside the system are collectively determined. Moreover, the capability of the template subsystem to activate the production of the membrane once a certain threshold of concentration in the product of the metabolic subsystem (determined by the length of the template polymer) is reached, constitutes a mechanism of delay or damping of internal disequilibrations. This mechanism is analogous to a water reservoir that establishes a threshold of concentration of the side product

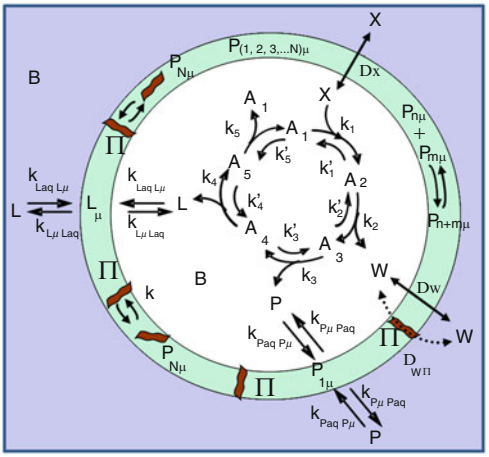

Fig. 1.7 Schematic graph of a minimal self-maintaining compartmentalised organization based on the complementarity between an internal autocatalytic reaction cycle and the self-assembly processes that make up the membrane (from its lipidic and peptidic building blocks). Peptides inserted in the membrane ensure the constitutive stability of the system by opening channels when internal pressure attains a critical threshold (Source: Adapted with permission from (Ruiz-Mirazo and Mavelli 2007). Copyright 2007 Springer-Verlag)

of the metabolism (used in the replication of the template), for the activation of the production of the membrane. This threshold is dependent on the length of the template polymer. Its role would consist of the system responding to an increase in internal pressure by building more membrane and thus avoiding an osmotic burst.

A final and interesting example is the recent model proposed by (Ruiz-Mirazo and Mavelli 2007, 2008). This is a self-reproducing vesicle whose membrane consists of both fatty acids and small peptides, such that the "mechanical" dynamics of the membrane are operationally coupled to the chemical dynamics of the internal autocatalytic network. The system realises control operations so as to maintain a steady state: when the osmotic pressure reaches a certain threshold, peptides in the membrane open channels; and this happens because, due to the elastic tension (a mechanical process), peptides inserted in the membrane adopt the conformation required to become waste-transport channels, thus enabling a faster release of the waste molecules and, consequently, a decrease in osmotic pressure differences (Fig. 1.7).

Constitutive stability requires not only that a set of constraints be mutually dependent but also that their activity can be modulated by specific (internal or external) perturbations in a way that preserves closure. What matters for our discussion here is that the response to the perturbation consists of a chain of changes affecting the constitutive organisation that, because of the architecture of the network of constraints, compensates for the initial variation. This means, in particular, that constitutive stability *does not require that we appeal to a different subset of constraints* specifically in charge of handling deleterious variations; rather, the variations of the constitutive organisation itself, induced by the perturbation, are sufficient.

Constitutive stability is *conservative*, and brings the system back to the same organisation that was in place when the perturbation occurred. The only way in which the system can change is by moving to a different organisation (i.e. a different "regime of self-maintenance") through the establishment of new stable dependencies between constraints, but this would be just a transition between different regimes, determined by the perturbation, and not an increase in overall complexity. Accordingly, although it enables the system to handle certain deleterious variations, constitutive stability is not a relevant starting point for the increase of organisational complexity since it does not enable the system to explore different regimes of closure.

1.8.2 Regulation[37]

The second way of handling perturbations is regulation, which is based on a qualitatively different form of organisation. Regulation requires that the closed organisation possesses a set of constraints exclusively operating when closure is being disrupted by a deleterious variation. The role of these constraints consists of re-establishing closure and bridging the gap between the activity of the system and its conditions of existence, by modulating (and possibly modifying) the constitutive organisation itself and/or its interaction with the environment. By definition, therefore, regulatory constraints are different (and complementary) with respect to constitutive ones: they do *not* contribute to the maintenance of closure in stable conditions (while constitutive ones do) but, when closure is being disrupted, they govern the transition towards its re-establishment (while constitutive ones do not).

As an example, consider the lac operon system, which regulates the metabolism of lactose in bacterium *E. coli*. In normal circumstances, *E. coli* metabolises the glucose taken in the environment. When the level of glucose becomes very low, and lactose is abundant, a mechanism called lac-operon is activated: the detection of lactose disinhibits the expression of a cluster of genes that enable lactose metabolism. In circumstances in which the availability of glucose is constant (that we take here as a set of constraints acting on a sequence of changes of metabolic pathways) these genes do not contribute to the maintenance of the organisation. The cluster of genes remains dormant and would not be included in the characterisation of the current closed organisation of constraints: in those conditions, nothing else in the system depends on the lac operon for its own maintenance. In turn, the lac operon becomes operational when a perturbation (the decrease of glucose levels) occurs and the maintenance of the organisation is menaced: the lac operon re-establishes closure by modifying the constitutive organisation (which shifts from glucose to

[37]The content of this section owes a lot to preliminary discussions with Leonardo Bich and Kepa Ruiz-Mirazo. See Bich et al. (forthcoming) for details.

lactose metabolism), and bridges then the gap between the activity of the system and its conditions of existence. Accordingly, the lac operon mechanism is regulatory.

A major implication is that *regulatory constraints are not subject to constitutive closure*, precisely because in a stable situation in which no deleterious variations occur they are not enabling (see Sect. 1.5. above), i.e. there is no constraint that depends on them. In turn, we claim that regulatory constraints are *second-order* constraints that, unlike constitutive ones, exert their causal actions *on changes of other constitutive constraints* of the organisation. In the case of the lac operon, for instance, the regulatory mechanism governs the transition from glucose to lactose metabolic pathways, which themselves consists of a set of constraints acting on underlying chemical reactions. In particular, in accordance with the definition given in Sect. 1.3 above, regulatory constraints exert a causal action, at time scale τ, on a change related to one (or a set of) other constraint(s), while being conserved through such change at τ. As with any constraint, their causal role at τ is intimately linked to their conservation, since the properties that are conserved are precisely those that provide them with specific causal powers. The lac operon mechanism is conserved at the time scale at which the shift from glucose to lactose metabolic pathways occurs.

The fact that regulatory constraints are not subject to first order closure is then what allows distinguishing them from the first order organisation, i.e. for distinguishing the "regulating system" from the "regulated system" in a principled way. One important implication is that, since they do not participate to constitutive closure, regulatory actions are *triggered* when the relevant (class of) perturbations occur. Therefore, a conceptual distinction can be made between the "constitutive" processes that maintain the regulatory functions (as for instance those which maintain the cluster of dormant genes responsible for the glucose/lactose switch in the lac operon case) and the processes (or changes) that trigger their action (as the increase of lactose and decrease of glucose in the environment). The triggering processes, ultimately due to an external or internal perturbation, may take many specific forms: in particular, it is worth noting that they are in many cases completely *distinct* from the constitutive ones, as for the lac operon. As a consequence, regulatory constraints realise a sort of *decoupling* from the constitutive organisation not only with respect to their effects, but also with respect to their dependence from the constitutive organisation for their triggering. Not only does regulation not contribute to constitutive closure but typically it is not even triggered by (changes of) processes involved in the constitutive closure. Such a decoupling of regulatory constraints vis-à-vis first-order organisation is, we will point out below, what allows them to play a crucial role in the increase of complexity (Fig. 1.8).

The regulatory subsystem (R), when activated (R/P) by a triggering perturbation (P), governs the transition from one constitutive organisation ($C_1 \ldots C_n$) to another one ($C_1' \ldots C_j$). In this specific case, the difference between the two constitutive organisations consists in the replacement of the constraint C_n with C_j.

At this point, a possible objection could be the following: if regulatory constraints are not subject to closure, can we still claim that they are part of the system? Does this characterisation imply that they are *external* to the organisation? We reply by claiming that, if one adopts a broader view, regulatory constraints can

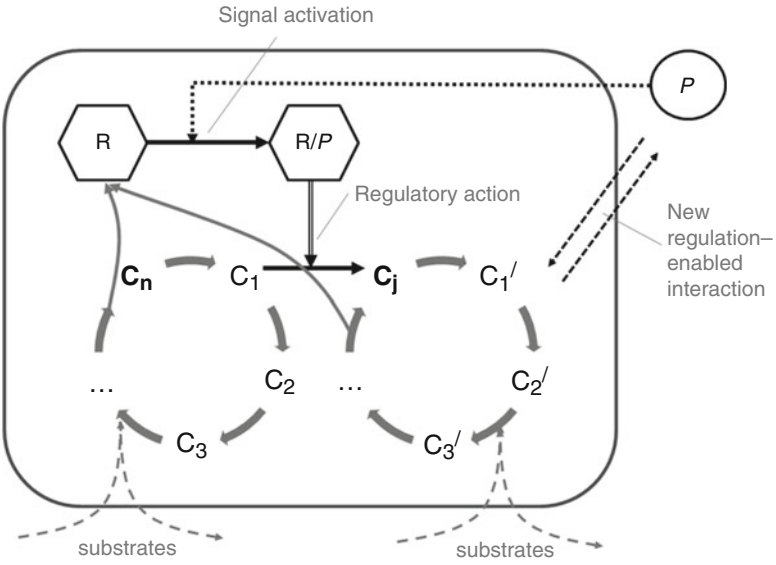

Fig. 1.8 Regulation (*Credits: Leonardo Bich*)

be shown to be both dependent and enabling at the same time, and therefore still subject to closure. In particular, they govern the *transition* between two organisations, the one whose closure is collapsing (the glucose-based one, in the example of the lac operon), and the one that they contribute to establishing (the lactose-based one): regulatory constraints depend on the (constitutive constraints of the) former, and enable the (constitutive constraints of the) latter. We argue that, accordingly, regulatory constraints are subject to a *second-order closure* between both themselves and the whole *set* of organisations among which they govern the transitions. The closure is of second-order because it is realised by a second-order organisation constituted by its set of regulatory constraints, on the one hand, and the set of available instantiations (one of which is enabled/channelled at a given moment) on the other. In other words, this second-order organisation consists of the set of available constitutive regimes of a closed organisation *given* a specific set of regulatory constraints and a set of deleterious variations to which the regulatory constraints are specifically sensitive.

Usually, these changes of regime are reversible and the second-order organisation may instantiate a previously collapsed first-order organisation if a new variation (or an end to the previous variation) were to activate regulatory capacities in this direction (in the case of the lac-operon, this would occur if the availability of lactose decreased, and that of glucose increased again). While the first-order organisation of constraints allows a modulation of the basic physicochemical processes, regulatory second-order constraints modulate the structure of the (first order) organisational closure. To take another example, a typical mechanism of regulation involves

allosteric enzymes, which have two (or more) binding sites and the capacity to switch between different metabolic paths. Accordingly, regulation responds to variations by inducing the (reversible) switch from one first-order instantiation to another: the system changes its organisation to maintain closure, which makes it not only robust, but also *adaptive*.[38]

In Chap. 3, we will argue at length that closure provides a naturalised ground for functionality, and its normative and teleological dimensions. Constraints subject to closure – we will claim – correspond to biological functions, and the conditions of existence of the closed organisation are the norms that they are supposed to satisfy. Here, let us emphasise that the idea according to which regulation is subject to second-order closure implies that it is subject to *second-order norms*, i.e. the norms generated by the conditions of existence of the second-order organisation. Accordingly, regulatory constraints are supposed to contribute to its maintenance by inducing the realisation of one of its possible instantiations (in the example of the lac operon, the possible instantiations being the glucose-based and the lactose-based organisations), according to the specific perturbation that affects the system. To the extent that the effects of regulation involve a shift from one specific constitutive organisation to another, and then from one closure to another, it follows that *regulation modifies first-order norms according to second-order norms*. More subtly: not only can regulatory constraints modulate the inherent norms of the organisation (as constitutive constraints do when the organisation varies for some reason) but, crucially, that modulation is *itself* teleological and normative: regulation is then, in the autonomous perspective, *functional modulation*. And metabolisms are those organisations realising regulated closure.

1.8.3 Regulation and the Increase of Complexity

At this point, we can come back to the initial question: why is regulation, characterised in this specific way, a relevant starting point for explaining not only stability, but also the increase of biological complexity? Or, as we framed the issue, why should we take regulated closure, and not just constitutive stability, as a characterisation of metabolism?

The central point is that regulation allows stability while enabling the increase of complexity, because second-order constraints are decoupled from the constitutive organisation, and therefore less affected by the perturbations impinging on it.

[38] As Di Paolo puts it, adaptivity is "a system's capacity to regulate, according to the circumstances, its states and its relation to the environment with the result that, if the states are sufficiently close to the boundary of viability, (1) tendencies are distinguished and acted upon depending on whether the states will approach or recede from the boundary and, as a consequence, (2) tendencies of the first kind are moved closer to or transformed into tendencies of the second and so future states are prevented from reaching the boundary with an outward velocity" (Di Paolo 2005: 438). For a detailed discussion, see also (Barandiaran et al. 2009); (Barandiaran and Egbert 2013).

What does this imply? In the case of constitutive stability, as (Christensen 2007) has also argued, achieving compensation depends on propagating changes through many local interactions within the organisation: this means that the time taken to achieve it can be long and, crucially, increasingly long as the size of the system increases. Regulation instead allows a decoupled subsystem to induce the appropriate collective pattern in a more rapid and efficient way. The modulation of the system is more efficient if, instead of modifying the very constitutive organisation, it can control the switches between available regimes, through a dedicated mechanism able to cope with specific perturbations. "Freed" from first-order organisation, at least at relatively short time scales, higher-order constraints can be left to "spontaneous" dynamics and allowed to explore higher degrees of organisational complexity, providing that, at longer time scales, it contributes to maintaining second-order closure, as explained above.

Let us illustrate how this can happen. Imagine a modified chemoton in which some modular components acquire functional tasks due to the specific sequence of their constitutive modules, as (Griesemer and Szathmáry 2009) have argued. If, instead of just one type of molecule being combined into the sequence of this modular component (say, a short polymer), two or more types constitute the building blocks, then the system will exhibit both a composition of its building blocks in specific concentrations and a sequence. While the concentrations, like other features of the chemoton, will depend on specific stoichiometric relations, the *sequence* is, stoichiometrically speaking, a "free" property. In turn, this stoichiometrically decoupled property can possibly be linked to component operations in the chemoton, so as to control them. In this situation the sequence, which does not participate directly in maintenance, takes over regulatory functions.[39] This allows a free exploration of the reaction space and, furthermore, once a regulatory hierarchical order has been shown to be possible, this in turn opens up the path to the creation of a new higher regulatory orders, and so on, indefinitely. As (Mattick 2004) has pointed out, what really enables the increase in complexity in biological evolution is not so much the capacity to generate a rich variety of elements, but rather the capacity for functional and selective control. The preliminary account of regulation from the autonomous perspective developed in these pages points to the same direction.

[39](Bechtel 2007) has pointed out the same argument, which he states as follows: "if control is to involve more than strict linkage between components, what is required is a property in the system that varies independently of the basic operations. The manipulation of this property by one component can then be coordinated with a response to it by another component so that one component can exert control over the operation of the other component " (Bechtel 2007: 290).

1.8.4 Towards Autonomy

Before concluding this chapter, an important clarification concerning the conceptual connection between regulation and autonomy is needed. In our view, regulation represents a qualitative transition on the path towards autonomy, not only because it enables an increase in functional and structural complexity, but also because the system has internalised constraints which are able to modify norms (those belonging to first-order organisation) in order to preserve its own existence. When closure is regulated, the system not only generates intrinsic norms but, as we emphasised, modulates these norms in order to promote its own maintenance; and this does not happen randomly, but in accordance with second-order norms. This means that self-determination has a stronger sense here: it is not just the generation of intrinsic norms, but also their submission, in accordance to (other) norms, to the maintenance of the system. And only the realisation of a *hierarchy*[40] of (at least) two orders of closure allows this distinction.

Hierarchical (regulated) closure, however, is not autonomy. As we will discuss in the following chapters, and particularly in Chap. 4, autonomy also implies the integration of an interactive dimension, which deals with the relations existing between the biological system and its environment. And we will see that, since regulation may concern both internally and externally-generated perturbations, the system can also exert a causal influence on its environment in order to promote its own maintenance: this is what we call adaptive agency.

[40]It is worth emphasising that, from the autonomous perspective, biological organisations might be hierarchical with respect to both *orders* and *levels* of closure. Chapter. 6 addresses explicitly this conceptual distinction.

2
Biological Emergence and Inter-level Causation

Whether adequate explanations in biology require appealing to an emergent and distinctive causal regime seems to have an obvious positive answer, insofar as biological systems evolve by natural selection (Mayr 2004: 31). Yet, as Wesley Salmon has pointed out (Salmon 1998: 324), one may distinguish between etiological explanations, which tell the story leading up to the occurrence of a phenomenon, and constitutive explanations, which provide a causal analysis of the phenomenon itself. Accordingly, whereas this goes without saying for etiological explanations, there seems to be no obvious answer to the question of whether a constitutive explanation of biological systems would also appeal to a distinctive regime of causation, emergent from and irreducible to that at work in natural physical and chemical systems.

What does the autonomous perspective have to say with respect to this issue? The view according to which closure constitutes a distinctive feature of biological organisation seems to require the adoption of a non-reductivist stance, according to which biological systems realise a regime of causation that is irreducible to (and then distinct from) those at work in other classes of natural (i.e. physical and chemical) systems. Indeed, in Chap. 1, we explicitly claimed that constraints and closure are irreducible to the thermodynamic flow on which the causal action is exerted, because of their conservation at the relevant time scales. Accordingly, the autonomous perspective seems to clearly advocate an emergentist stance with respect to constitutive explanations in biology. Yet, this claim calls for an adequate philosophical justification: which properties do make closure of constraint irreducible and emergent? What is the precise account of emergence invoked here? In this chapter, we aim to develop these issues in some details, by situating the autonomous perspective within the context of the more general discussion on emergence and reduction.

The ideas presented in this chapter, as well as most parts of the text, were originally presented in (Mossio et al. 2013).

A. Moreno, M. Mossio, *Biological Autonomy*, History, Philosophy and Theory of the Life Sciences 12, DOI 10.1007/978-94-017-9837-2_2

Similarly, it is currently unclear whether or not closure involves inter-level causation. At first sight, it seems obvious that closure inherently relies on the causal interplay between entities at different levels of description: the integrated activity of lower-level constituents contributes to generate the higher-level organisation, and the higher-level organisation plays a crucial role in maintaining and regenerating its own constituents, as well as controlling and regulating their behaviour and interactions. As a matter of fact, Moreno and Umerez did argue in a previous contribution that inter-level causation is a fundamental aspect of biological organisation (Moreno and Umerez 2000). However, the appeal to inter-level causation in biological systems may oscillate between two different interpretations of the concept: on the one hand, the causal influence of an entity located at a given level of description on an external entity located at another (upper or lower) level of description; on the other hand, the causal influence of an entity, taken as a whole, on its *own* parts. As we will discuss, while it might seem quite obvious that closure involves inter-level causation in the first sense, a more difficult issue is whether this holds also for the second sense, which requires complying with more restrictive conceptual conditions.

Our argument in this chapter will be twofold. On the one hand, we will argue that closure can be consistently understood as an emergent regime of causation even though the autonomous perspective is interpreted, as we do, as being fundamentally committed to a *monism* (of properties). On the other hand, we will maintain that, although the mutual relations between constraints are such that the very existence of each of them depends on their being involved in the whole organisation, an emergent closed organisation does *not* necessarily imply inter-level causation, be it upward or downward, in the restrictive sense of a causal relation between the whole and its own parts (what we will label *nested* causation). Yet, as we will suggest, the appeal to inter-level causation in this sense (which is the philosophically more interesting and more widely discussed one) may possibly be relevant for organisational closure, if the adequate conceptual justification were provided.

The structure of the chapter being quite complex, let us provide a synthetic overview. In Sect. 2.1 we discuss one of the main philosophical challenges to the idea of emergence – Kim's exclusion argument – by focusing on the fact that it applies to a specific account of emergence formulated in terms of supervenience and irreducibility. In Sect. 2.2, we recall the distinction between two dimensions of the debate about emergence – ontological irreducibility and epistemological non-derivability – and clarify that a pertinent defence against the exclusion argument can be expressed, as we shall demonstrate, exclusively in terms of irreducibility. Section. 2.3 offers a conceptual justification of emergent properties and argues that configurations, because of the relatedness between their constituents, possess ontologically irreducible properties, providing them with distinctive causal powers. In Sect. 2.4, we focus on the specific case in which configurations exert distinctive causal powers as constraints, and argue that closure can be taken as a specific kind of higher-level emergent configuration (an organisation), ontologically irreducible and possessing distinctive causal powers: in particular, we will emphasise self-determination itself, which grounds most other features of autonomous systems.

Section. 2.5 concludes the analysis, and claims that closure can be justifiably taken as an emergent biological causal regime without admitting that it inherently involves inter-level causation in the precise sense of nested causation. Yet, the connection between closure and nested causation remains an open issue requiring further theoretical and scientific research.

2.1 The Philosophical Challenge to Emergence

The very idea of closure as a "distinctively biological" regime of causation cannot be justified unless it can be shown that, in some way, a given system possesses characteristic properties by virtue of which it may exert distinctive causal powers. A conceptual justification of emergence seems then to be a necessary requirement for a coherent account of biological causation.

Philosophical work on emergence began during the late-nineteenth and early-twentieth centuries, with the writings of the so-called "British Emergentists" (Mill 1843; Alexander 1920; Lloyd Morgan 1923; Broad 1925), and has developed considerably over recent decades.[1] As has often been underscored, a central contribution to this debate was made by Jaegwon Kim, who developed one of the most articulated conceptual challenges to the idea of emergence (Kim 1993, 1997, 1998, 2006).

In a recent survey of these issues (Kim 2006), Kim recalls what are, in his view, the two necessary ingredients of the idea of emergence, i.e. supervenience and irreducibility. *Supervenience* is a relationship by virtue of which the emergent property of a whole is determined by the properties of, and relations between, its realisers. As the author himself puts it (Kim 2006: 550):

> *Supervenience*: If property M emerges from properties $N_1, \ldots N_n$, then M supervenes on $N_1, \ldots N_n$. That is to say, systems that are alike in respect of basal conditions, $N_1, \ldots N_n$ must be alike in respect of their emergent properties.

In turn, *irreducibility*, and more precisely, according to Kim, *functional* irreducibility, is expressed as follows:

> *Irreducibility of emergents*: Property M is emergent from a set of properties, $N_1, \ldots N_n$, only if M is not functionally reducible with the set of the N_s as its realizer (Kim 2006: 555).

Given the account of emergent properties in terms of supervenience and irreducibility, the central issue is whether these properties may possess distinctive causal powers. In his work, Kim has developed several lines of criticisms vis-à-vis emergence. The one that is particularly relevant here claims that emergent properties are exposed to the threat of epiphenomenalism. Kim's argument on this matter is known as the "exclusion argument", and has been expounded on several occasions. Very briefly, the idea is the following. If an emergent property M emerges from some

[1] See (Sartenaer 2013) for an informed and insightful analysis of both the history and contemporary structure of the debate about emergence and reduction.

basal conditions P, and M is said to cause some effect, one may ask "why cannot P displace M as a cause of any putative effect of M?" (Kim 2006: 558). If M is nomologically sufficient for whatever effect X, and P is nomologically sufficient for M (because of the supervenience relation), it seems to follow that P is nomologically sufficient for X, and M is "otiose and dispensable as a cause" for X. As a result, invoking the causal power of emergent structures would be useless, since it would be epiphenomenal.

The exclusion argument has crucial implications for the debate about emergence and reduction. If one admits (1) that the relation between M and P is correctly described in terms of supervenience and (2) the validity of what we will call here the *principle of the inclusivity of levels,*[2] i.e. "the idea that higher levels are based on certain complicated subsets from the lower levels and do not violate lower level laws" (Emmeche et al. 2000: 19), then two problematic consequences follow.

First, the explanation is exposed to the danger of causal drainage. Indeed, if the causal powers of an emergent entity can be reduced to the causal powers of its constituents, and if, as may indeed be the case, there is no "rock-bottom" level of reality, then it seems that "causal powers would drain away into a bottomless pit, and there would not be any causation anywhere" (Campbell and Bickhard 2011: 14).[3] Second, if there were some scientifically justifiable rock-bottom level of reality (which is a far from trivial assumption),[4] and causal drainage were blocked, the exclusion argument would force reductive physicalism (see Vicente 2011 for a recent analysis). In this second case, any appeal to distinctively biological causal relations (such as closure itself, and related notions such as "integration", "control" and "regulation" etc.) would, at best, constitute an heuristic tool, unless it could be

[2]We take here the notion of "inclusivity of levels" as being analogous to Kim's "Causal Inheritance Principle" (Kim 1993: 326), according to which if a property M is realised when its physical realisation base P is instantiated, the causal powers of M are identical to the causal powers of P. By opting to use the term "inclusivity of levels", we wish to emphasise the idea that in the natural world, all causes are either physical or the result of the interaction between physical entities: no special causes (vitalist, spiritual, etc., that are not physically instantiated) are introduced at different levels, e.g. at the biological and mental ones. It should be noted that Kim's argument requires also the Causal Closure Principle as a premise, in the sense that the ultimate reduction of an emergent property to its fundamental realisation base is possible only if the basal level is causally closed (Kim 2003). Yet, we maintain that the validity of the inclusivity of levels does not necessarily require an appeal to the causal closure: emergent causal powers can be reduced to basal powers even though the latter are not shown or are supposed to be closed. Consequently, the argument that we develop in this chapter does not depend on the Causal Closure Principle.

[3]In Kim's intentions, the exclusion argument is originally targeted at mental causation and is not supposed to imply causal drainage. As a matter of fact, Kim himself has vehemently tried to avoid causal drainage as the ultimate consequence of the argument in favour of reduction. Moreover, on the basis of a commitment to the Standard Model and its lowest level of fundamental physical particles, he rejects the arguments based on the possibility of the absence of a rock-bottom level of reality. For a detailed discussion of these issues see, for example, Block's criticism of Kim's reduction argument (Block 2003) and Kim's reply (Kim 2003).

[4]The idea of a basic level with self-sufficient basic entities has been deeply questioned in microphysics, the very domain reductionist approaches appeal to as fundamental, where relational and heuristic accounts have been developed (Bitbol 2007).

demonstrated that said relations can be adequately reduced to physical causation or, more generally, to any "more fundamental" regime of causation.

In both cases, the very possibility of biological explanation is menaced. An adequate justification of a distinctive regime of biological causation should be provided, in order to (1) avoid the danger of endless causal drainage and (2) make the biological explanation theoretically independent from the physical and chemical ones, and directly related to the specificity of biological phenomenology, instead of being derived from lower level explanations and dependent on a single physical "theory of everything" (Laughlin and Pines 2000).

2.2 Irreducibility Versus Non-derivability

Before addressing the exclusion argument, let us make a preliminary conceptual distinction between two dimensions of the debate about emergence, i.e. irreducibility and non-derivability.

The exclusion argument challenges the status of emergent properties as causal agents of the world: how can a property be supervenient on something while being, at the same time, irreducible, and then possessing distinctive causal powers? An appropriate reply should then deal with the ontological issue of irreducibility, and justify emergent properties by showing that they are something ontologically "new" with respect to their realisers. Irreducibility, therefore, is inherently linked to *ontological novelty*.

Irreducibility should not be confused with the possible non-derivability of emergent properties from the emergence base, which is an *epistemological* issue. Non-derivability refers to the fact that given a description of the properties of the realisers, it is not possible to predict, explain or deduce the emergent properties of the whole.

As a matter of fact, most of the philosophical debate has tended to merge the two issues,[5] and to take both irreducibility and non-derivability as marks of emergence: emergent properties are not only irreducible but also, and crucially, non-derivable. Consider, for instance, the classic distinction between "resultant" and "emergent" properties, which is based precisely on criteria of non-derivability (or "non-deducibility", in Kim 2006: 552).[6] Resultant properties are aggregative properties, which the whole possesses at *values* that the parts do not (i.e. a kilogram of sand has a mass that none of its constituents has). Emergent properties, in turn, are properties of a *kind* that only the whole possesses, whereas the parts do not (i.e. a system can be alive, whereas none of its parts are alive). Although resultant

[5]The distinction has however been formulated, for instance, by (Silbersten and McGeever 1999), according to whom *epistemological* emergence concerns models or formalisms, while *ontological* emergence involves irreducible causal capacities. Here, we follow this conventional distinction.

[6](Van Gulick 2001) refers to resultant and emergent properties as "specific value emergent" and "modest kind emergent" properties, respectively.

properties can be said to be, in a general sense, irreducible to the properties of their realisers, they are not what British emergentists (and most contemporary authors) had in mind when they talked about emergence. In fact, when appealing to notions like "unpredictability" or "unexplainability" as the mark of emergence, most authors are focusing on epistemological non-derivability.[7]

Yet we maintain that ontological irreducibility and epistemological non-derivability are logically distinct dimensions, and call for independent philosophical examinations. In what follows, we will discuss them separately, as two different issues.

On the one hand, we will develop throughout most of the chapter, a philosophical defence of emergence against the exclusion argument and the danger of epiphenomenalism, by relying exclusively on the irreducibility of emergent properties, without addressing the issue of their non-derivability. Emergent properties, we will argue, can be defined *exclusively in terms of irreducibility* and, crucially, they provide the system with distinctive causal powers *even though* they are derivable from their emergence base.

On the other hand, the issue of the non-derivability of emergent properties may play an important role in the discussion on whether emerging properties enable a system to exert inter-level causation between the whole and the parts. As we will suggest in the last part of the chapter (Sect. 2.5.2), if an emergent property is proven to be also non-derivable from the properties of the constituents, it may be possible to interpret the relationship between the whole and the parts as involving inter-level causation, because of the epistemological gap between them.

2.3 Irreducibility and Emergence

The aim of this section is to offer, in response to the exclusion argument, a conceptual justification of emergent properties provided with irreducible and distinctive causal powers. The core of the argument consists in suggesting that a coherent account of emergent causal powers can be obtained by rejecting the identification between the "supervenience base" and the "emergence base" of a property. As we will propose, a property of a whole can be functionally reducible to the set of properties of its constituents (its supervenience base) while being functionally irreducible to, and hence emergent on, various categories of entities that are distinct from that set. Once the distinction between the supervenience and emergence base is conceded, the resulting account of emergence, we will argue, eludes the exclusion argument and justifies the existence of distinct regimes of causation, even when maintaining the principle of the inclusivity of levels.

The argument will proceed in two steps. First, we will advocate (Sect. 2.3.1) an interpretation of the relation between the whole and the parts in terms of

[7]Crutchfield, for instance, distinguishes between two different definitions and classes of models of emergence, according to two different limitations in our capability "in principle" to describe emergent phenomena: nonpredictability and nondeducibility (Crutchfield 1994).

relational mereological supervenience, according to which a supervenience relation holds between the whole and the *configuration* of its own constituents, and not between the collection of constituents taken separately. We will then put forward a *constitutive* interpretation of relational supervenience, according to which supervenient properties can in principle be reduced to the configurational properties of the supervenience base. The main implication is that a supervenient property M and its basal properties $S_1, \ldots S_n$ have identical causal powers. In the adoption of such a constitutive interpretation of relational supervenience lies, in our view, the *monist* stance of the autonomous perspective.

Second, we suggest (Sect. 2.3.2) that, even under the constitutive interpretation of relational mereological supervenience, a relation of emergence (as irreducibility) holds not between M and configurational properties, but instead between configurational properties and the properties of different categories of entities which do not belong to the configuration. Consequently, configurations can be justifiably said to possess irreducible and emergent properties and hence be able to exert non-epiphenomenal causal powers (in particular, as recalled in Sect. 2.4, as constraints) even under a monist interpretation of the autonomous perspective.

2.3.1 Supervenience and Constitution

The logic of the exclusion argument is based on the way in which the relation between an emergent property M of the whole W and the set of basal properties $N_1, \ldots N_n$ of its constituents P is conceived. Namely, the relation is held to be simultaneously one of (mereological) supervenience and functional irreducibility, while assuming at the same time, as mentioned above, the validity of the principle of inclusivity of levels.

In his *Mind in a Physical World*, Kim paved the way for an answer to the exclusion argument capable of maintaining the inclusivity of levels, by clarifying the terms of the supervenience relation, and particularly specifying how the supervenience base is to be conceived. Kim argues that emergent properties are *micro*-based *macro* properties, i.e. second-order properties emerging out the first-order properties and relations of the basal constituents (Kim 1998: 85–86). The central idea is that the relevant supervenience base is not a set of properties of constituents taken individually or as a collection, but rather the properties of the *configuration* of constituents, i.e. the whole set of inherent *and* relational properties of the constituents. In other words, mereological supervenience should not be interpreted as atomistic but rather relational (see also Thompson 2007: 427–8; Vieira and El-Hani 2008).[8]

[8]The debate between a relational and an atomistic interpretation of the supervenience and emergence bases has a long history that dates back to the first formulations of the notion of Emergence in British Emergentism. In Alexander's framework, for example, space and time, the lower level on

The move to adopt relational mereological supervenience makes configurations of constituents the relevant supervenience base. The basal properties $S_1 \ldots S_n$ that bring about a supervenient property M are not the properties of the collection of constituents taken separately, but rather the configurational properties of the constituents *qua* constituents (including their mutual *relations*, which alter their intrinsic properties as separate elements), which appear only when the configuration is actually realised. If the basal constituents actually and collectively constitute a global pattern or system W, then their properties would now include those generated by their being involved in specific relations and interactions with other elements.

The adoption of relational mereological supervenience has relevant implications for the question concerning the distinctive causal powers of the supervenient property with respect to its supervenience base. Indeed, the idea that emergent properties would be reducible to the properties of the constituents taken in isolation seems to be excessively committed to an atomistic view of nature, which does not take relations into account (Campbell and Bickhard 2011). In turn, the claim that a supervenient property M is in principle reducible to the set of configurational (i.e. including relations) properties $S_1, \ldots S_n$ of its constituents is more convincing (again, by assuming the principle of inclusivity of levels), since configurations are far richer and more complex determinations than the mere collection of intrinsic properties of constituents.

Accordingly, we hold that relational supervenience does not imply functional irreducibility but rather, on the contrary, *constitution*: M supervenes on $S_1, \ldots S_n$ since it consists of $S_1, \ldots S_n$. A supervenient property M of a whole W corresponds to the set of configurational properties $S_1, \ldots S_n$ of its constituents (its supervenient base B). The set of the (relevant) configurational properties of the constituents of the system is, at least in principle, equivalent to the supervenient property. Hence, if M can be functionally reduced to the set $S_1, \ldots S_n$ of configurational properties of its constituents, it follows that it cannot possess distinctive causal powers[9] since, in fact, they are equivalent.[10]

---

which the whole natural world emerge, are relational concepts, not definable separately (Alexander 1920). The opposition between atomistic and relational approaches is particularly evident in Lloyd Morgan's work. He opposes to the billiard balls model of extrinsic interactions, the idea of relatedness based on inherent relations, which contributes to specifying the properties of the terms involved in the relation (Lloyd Morgan 1923: 19). It is also worth noting that, according to some authors, Kim's reference to relations is still made in a fundamentally atomistic framework, and does not imply a clear commitment to relational mereological supervenience, which implies the idea that relations "do not simply influence the parts, but supersede or subsume their independent existence in an irreducibly relational structure" (Thompson 2007: 428).

[9]The interpretation of relational mereological supervenience in terms of constitution is consistent, we argue, with the position developed by (Craver and Bechtel 2007) within their mechanistic framework. As they suggest, relations between constituents located at different levels in a mechanism are better understood as constitutive relations (pp. 554–555). See Sect. 2.5.1. below for a detailed discussion.

[10]For simplicity, we will only refer, from now on, to "configurational properties $S_1, \ldots S_n$" (equivalent to "property M"), and to "configuration C" (equivalent to "supervenience base B").

Yet, as we will claim in the following section, a coherent account of emergent properties provided with distinctive causal powers can still be provided, even under the constitutive interpretation of whole-parts relations.

2.3.2 A Reply to the Exclusion Argument

Our reply to the exclusion argument consists of arguing that even though supervenient properties (M) have no distinctive causal powers with respect to the configurational properties $S_1, \ldots S_n$ of the constituents, $S_1, \ldots S_n$ themselves (which are equivalent to M, because of constitution) are irreducible properties which may generate distinctive causal powers. Accordingly, $S_1, \ldots S_n$ can be said to be genuinely emergent. In other terms, there is an interpretation of emergence that is compatible with a monist stance.

Here is our argument. A given configuration C of elements of a whole W is identified by a set of (possibly dynamic) distinctive constitutive and relational properties $S_1, \ldots S_n$. On the basis of this set of distinctive properties, a configuration is functionally irreducible to any entity that does not *actually*[11] possess the same set of properties. We claim that a relation of emergence holds between a configuration C and any emergence base P whenever C is irreducible to P, i.e. if C possesses some distinctive set of configurational properties that P does not possess, such that C does *not* supervene on P. The reader would immediately note that this characterisation of emergence is very general, and could in principle include a wide range of obvious and uninteresting cases of P, which would not be considered salient for the philosophical debate on emergence. This is correct, and we deal with this issue just below. Yet, let us point out here that, as (Campbell and Bickhard 2011: 18; see also Teller 1986) have highlighted, appealing to configurations seems to be a sufficient answer to the danger of causal drainage and epiphenomenalism. The crucial point, as mentioned above, is that configurations include relational properties, which cannot be reduced to intrinsic properties, i.e. properties of constituents taken in isolation. Relatedness is ontological novelty. Consequently, because of relatedness (again: *actual* relatedness), configurations may possess distinct causal powers that would not otherwise exist.[12]

[11] It is important to emphasise that configurational properties must be actually realised, and not just "dispositional". As a consequence, a configuration C is functionally irreducible, in this account, also to those entities that would possess the "potential disposition" to actualise these properties.

[12] It might be objected that not *any* relational property gives rise to distinct causal powers. For instance, spatial relations do not seem to be relevant candidates in this respect while, for instance, relations that express energetic bonds among constituents do. In our view, useful specifications could indeed be offered on this point. Yet, we do not know whether a distinction between relevant and irrelevant classes of relational properties could (and should) be established *a priori*: hence, we do not provide further details in this book. For a related account of emergence, see also (Hooker 2004).

To avoid confusion, it is important to stress again that this account, in contrast with most existing ones, defines emergence exclusively in terms of ontological irreducibility, leaving aside the issue of the epistemological non-derivability of C from P. C is emergent on P if it possesses some set of *new* (relational) properties $S_1, \ldots S_n$ which P does not possess, and which are then irreducible to the set of properties $N_1, \ldots N_n$ of P. A different issue, which is irrelevant here, is whether one can derive or predict $S_1, \ldots S_n$ from $N_1, \ldots N_n$. In particular, $S_1, \ldots S_n$ would be irreducible *even if* they were derivable, because of the novelty introduced by the relations between constituents.

At this point, given the constitutive relation between the whole and its constituents advocated so far, one may wonder what exactly configurations do emerge on. Following our definition, three main kinds of emergent base P can be logically identified. Firstly, the configuration C is not supervenient, yet is emergent on the properties of any proper *subset* P_{sset} of its constituents (its parts). A wheel has emergent properties and distinctive causal powers on any subset of itself (e.g., a half-wheel). Secondly, the configuration C is not supervenient, yet is emergent on its *substrate* P_{sstr}, i.e. the collection of its constituents taken separately, as if they were not constituents (so to speak, the "potential ingredients" of a configuration). A wheel is emergent on the collection of molecules taken as if they were not actually assembled as a wheel.

Thirdly, and most importantly, the configuration C is not supervenient, yet is emergent on its *surroundings* P_{surr}, i.e. each set of external elements that does not actually constitute C. The wheel is emergent on each set of external molecules or entities, which are not actual constituents of it. In particular, given that a very broad set of entities might be included in P_{surr}, only relevant instances will actually be considered: in particular, the reference to surroundings P_{surr} will be restricted to those P_{surr} on which the configuration C has causal effects, by virtue of its emerging properties. As we will discuss in the following section, this is precisely the relevant case with regard to biological systems.

At this point, we have all the elements required to formulate our reply to the exclusion argument. The argument claims that emergent properties cannot be such unless it can be shown that they possess distinctive causal powers; at the same time, it seems that, as supervenient properties, they do not possess new causal powers with respect to their supervenience base. Hence, they are epiphenomenal. To this, we reply that emergent properties do not need to be irreducible to their supervenience base to possess distinctive causal powers: what matters is that configurations, because of relatedness, possess irreducible properties with respect to their subsets, substrate and (relevant) surroundings. Supervenience and emergence are then *alternative* notions: either a set of properties is supervenient on another one (in which case there is constitution between them), or it is emergent (in which case there is irreducibility).

Let us stress again that this way of conceiving emergence, interpreted exclusively as ontological irreducibility, is indeed very general. For instance, all chemical bonds are configurations emergent on their parts, substrate and surroundings, since they realise new relations, and therefore possess distinctive configurational properties.

Yet, the fact that this definition covers also irrelevant or obvious cases is, we argue, the price to pay for making it compatible with the monist stance of the autonomous perspective, represented by the constitutive interpretation of the relations between the whole and the parts. More generally, we hold that this characterisation of emergence is *sufficient* to provide a justification for the appeal to distinctive and irreducible causal powers in the scientific discourse (Laughlin et al. 2000), specifically in biology. Emergence appears whenever scientists are dealing with a system, such as a biological system, whose properties are irreducible to those of its isolated parts, substrate and surroundings. In such cases, one must introduce new observables, relations and causal powers, which exist only within that very system, and not in its emergent base.[13]

2.4 Constraints and Closure as Emergent Determinations

By virtue of their relatedness, configurations possess emergent properties and may exert distinctive causal powers on their surroundings that can take different forms, in accordance with the kind of systems under consideration. Let us focus here on the case in which these causal powers are exerted as constraints, which can in turn be organised as closure.

As discussed at length in Chap. 1, constraints are those configurations that, while exerting a causal action on a set of physicochemical processes and reactions (involving the movement, alteration, consumption, and/or production of entities, in conditions far from thermodynamic equilibrium), can also be shown, at the relevant time scale, to be conserved with respect to them.

By using the labels introduced in this chapter, we can rephrase the conditions (see Chap. 1, Sect. 1.3) under which an entity can be taken as a constraint as follows. Given a particular P_{surr}, a configuration C acts as a constraint C_{constr} if:

1. At the relevant time scale τ, C_{constr} *exerts a causal action on* P_{surr}, i.e. there is some observable difference between P_{surr} and $P_{surr}{}^{c}$ ($P_{surr}{}^{c}$ is P_{surr} under the causal influence of C_{constr} by virtue of relational properties $S_1, \ldots S_n$);
2. At τ, C_{constr} *is conserved throughout* P_{surr}, i.e. there is a set of emerging properties $S_1, \ldots S_n$ of C which remain unaffected throughout P_{surr}.

In Chap. 1 we claimed that constraints constitute a distinct regime of causation because, by fitting these conditions, they are irreducible to the thermodynamic flow on which they exert a causal action. In particular, for a given effect, one can make a

[13]It is worth noting that the relation between the emergent properties and their emergence base can be interpreted both synchronically and diachronically. Being based on novelty, in fact, the irreducibility to any entity that does not belong to an actual configuration is in principle compatible with both dimensions of emergence.

conceptual distinction between two kinds of causes: the "material" ones (inputs or reactants), which do not meet condition 2, and constraints.

Now, this characterisation of constraints relies precisely on a justification of the emergent nature of those relevant properties in virtue of which they satisfy the above conditions. And – we submit – our conception of emergence provides such justification: constraints are irreducible to the thermodynamic flow insofar as they possess properties that emerge from that flow, because of the relatedness of their configurations. In other words, to explain why constraints are irreducible to the relevant P_{surr} on which they act (and then why the exclusion argument does not apply to them) one has to appeal to the ontological novelty of their emergent properties with respect to P_{surr}, a novelty generated by the relatedness of their configurations at the relevant time scale. Similarly, their emergent configurational properties support the distinctive causal powers of constraints: at biological relevant time scales, constraints – as discussed in Chap. 1 – are enabling, in the sense of causally contributing to the maintenance of the whole organisation, which would otherwise be a highly improbable (or *virtually* impossible) phenomenon.[14]

It is worth noting that, in line with our characterisation, constraints can be said to emerge on a wide spectrum of entities belonging to P_{surr}, which goes far beyond the case of processes and reactions in far-from-equilibrium thermodynamic conditions. Yet, as highlighted earlier, we restrict our analysis to the subset of P_{surr} on which constraints exert a causal action, because this is the case in which the (ir)reducibility of one regime of causation to another has explanatory relevance for biological systems.

Let us now turn to closure. As discussed in Chap. 1, closure generates an organisation in which the network of constraints achieves self-determination as collective self-constraint. In the terms of this chapter, hence, closed organisations are then *a specific kind of higher-level configurations* C_{org}, whose distinctive feature consists in the fact that their constituents are themselves configurations C_{constr} acting as constraints. As such, we claim that closure, because of the relatedness between the constraints, generates itself ontological novelty and new emergent properties, possibly supporting distinctive causal powers of the organisation. In particular, let us distinguish three aspects.

First, as we already mentioned in Chap. 1, closure inherits the irreducibility of constraints to the surroundings P_{surr}, the thermodynamic flow. To the extent that no adequate account of constraints can be provided by reducing them to the causal regime of thermodynamic changes, *a fortiori*, closure of constraints cannot itself be reduced to a closed network of processes and changes. Hence, a description of biological systems in terms of pure thermodynamics would not be able to account for their organisation.

[14] In Chap. 1 we argued that, in most cases, C_{constr} does not extend the set of possible behaviours of P_{surr}, i.e. P_{surr} could in principle (although it is highly unlikely) exhibit, at different time scales, the behaviour of $P_{surr}{}^{c}$ without the action of C_{constr}.

Second, and crucially, organisations themselves possess additional emergent properties with respect to their substrate P_{sstr}, namely the collection of their constitutive constraints taken separately. When constraints actually realise closure, their relatedness generates new emergent properties that none of them would possess separately. One of these properties is, of course, self-determination itself: individual constraints cannot determine (or, more precisely, maintain) themselves, only their collective organisation can.[15] Hence, the capacity to self-determine is, in the biological domain, an emergent property generated by the specific relatedness of the closure of constraints. As mentioned in the Introduction, the whole autonomous perspective can be seen as an exploration of the distinctive features of biological organisms generated by their emergent capacity of self-determination. In a way, each chapter of this book focuses on a set of properties or capacities stemming from the realisation of organisational closure: this holds for agency, biological complexity, multicellular organisations, cognition, and so on . . . . The philosophical justification of the irreducible ontological novelty of closure, through the appeal to the relatedness between the constitutive constraints, is therefore a pivotal step toward the elaboration of the whole theoretical framework.

To make this point clear, let us mention a specific example, which introduces the following Chap. 3. As we will discuss at length, closure generates functionality. As has been recently argued (Mossio et al. 2009b; Saborido et al. 2011), when they are subject to closure, constraints correspond to biological functions: performing a function, in this view, is equivalent to exerting a constraining action on an underlying process or reaction in an organised system. All kinds of biological structures and traits to which functions can be ascribed satisfy the above definition of constraint, although at very different temporal and spatial scales. Some intuitive examples in addition to the vascular system mentioned above include, at different scales: enzymes (which constrain reactions), membrane pumps and channels (which constrain the flow of ions through the membrane) and organs (such as the heart which constrains the flow of blood), among others. The emergence of closure is then the emergence of functionality within biological organisation: constraints do not exert functions when taken in isolation, but only insofar as they are subject to a closed organisation. As a consequence, the defence of a naturalised account of functionality as a distinctive biological dimension (developed in Chap. 3) fundamentally relies on the justification of the emergent and irreducible nature of closure advocated here.

Third, closure is specifically defined with respect to the emergence base P_{surr}, constituted by a set of processes and changes occurring in conditions far from thermodynamic equilibrium. The two causal regimes, although mutually irreducible, realise a two-way interaction in biological systems, to the extent that constraints act on thermodynamic processes and changes, which in turn contribute to reproducing or maintaining these constraints. Hence, it might be tempting to

[15] Whereas, as we mentioned in Chap. 1, Sect. 1.4 above, individual self-determination does in fact occur in physics, in the case of self-organising dissipative structures.

conclude that closure (just like any form of self-maintenance) inherently involves not just emergent causation, but also inter-level causation, at work between the two causal regimes. Yet there are several reasons to resist this temptation, at least insofar as particularly controversial kinds of inter-level causation are concerned.

2.5 Inter-level Causation

The issue of inter-level (be it upward or downward) causation has been, explicitly and otherwise, a central aspect of the debate about emergence from its earliest beginnings,[16] since the very concept of emergence carries on the issue of the relations between properties at different levels.

In Kim's account, attributing causal powers to emergent properties necessarily implies downward causation. Let us recall his argument that, as we discussed, identifies the supervenience and the emergence bases. Let M and M* be two emerging properties, and suppose that M causes M* (a case of "same-level" causation). As an emergent property, M* has an emergence base, say P*. Given the supervenience relation, P* is necessary and sufficient for M*: if P* is present at a given time, then M* is also present. Accordingly, it is unclear in what sense M could play a causal role in bringing about M*: given P*, its role would be useless, unless M is in fact somehow involved in causing P*. In other words, the same-level causation of an emergent property makes sense only if this implies the causation of the "appropriate basal conditions from which it will emerge" (Kim 2006: 558). Consequently, causation produced by emergent properties seems to imply, in all cases, downward causation in the sense of a causal influence exerted by an emergent property on the basal conditions of another emergent property.

Yet, as Kim himself has argued (Kim 2010), this general form of downward causation, i.e. a causal influence exerted by an entity at a higher level on a different entity located at a lower level, is indeed widespread and unproblematic. In particular, this interpretation of downward causation applies straightforwardly to self-maintenance and closure, which inherently involve, as discussed above, upward and downward causation between constraints and dynamics, with each being located at different levels of description.

[16] According to Lloyd Morgan, "[. . .] when some new kind of relatedness is supervenient (say at the level of life), the way in which the physical events which are involved run their course is different in virtue of its presence-different from what it would have been if life had been absent. [. . .]. I shall say that this new manner in which lower events happen – this touch of novelty in evolutionary *advance depends on* the new kind of relatedness" (Lloyd Morgan 1923: 16). According to (Stephan 1992) Lloyd Morgan's passage could admit different interpretations, such as that of a logical claim about supervenience. On the contrary, McLaughlin asserts: "In Morgan one finds the notion of downward causation clearly and forcefully articulated" (McLaughlin 1992: 68).

The more controversial form of downward causation would be that exerted by a whole on its own constituents (in Kim's terms, "reflexive" downward causation, Kim 2010: 33). According to (Emmeche et al. 2000), there are various possible interpretations of reflexive upward and downward causation. In their view, the only non-contradictory versions of the concept are those that interpret downward causation in terms of "formal" causation (Emmeche et al. 2000: 31–32), such that the whole exerts *a constraining action on its own constituents*, by selecting specific behaviours from among a set of possible ones. This interpretation can be taken, as Emmeche and his co-authors claim, as the standard and possibly more compelling one of downward causation, and it is very close to the original proposal by (Campbell 1974).[17]

As an illustration, consider Sperry's classic example of the wheel rolling downhill (Sperry 1969). On the one hand, the various molecules generate the wheel as a whole, and on the other, as (Emmeche et al. 2000: 24) explain:

> none of the single molecules constituting the wheel or gravity's pull on them are sufficient to explain the rolling movement. To explain this one must recur to the higher level at which the form of the wheel becomes conceivable.

The set of configurational properties of molecules is assumed here to underdetermine their behaviour so that, in order to explain it, one needs to appeal to a property of the whole (in this case: the form of the wheel) that would generate a causal influence (a selective constraint) exerted on its own constituents.

Because of the (assumed) under-determination of constituents by configurational (intrinsic and relational) properties, constituents' behaviour is partly determined, in a functionally irreducible way, by the whole to which they belong. In particular, this train of thought seems to apply equally to biological systems, in which the behaviour and dynamics of the parts appear to be, in an important sense, determined (notably through regulation and control functions) by the downward causation exerted by the whole system to which they belong.

In what follows, we will examine whether self-maintenance and closure do indeed involve some form of reflexive inter-level causation, intended as a particular form of constraint exerted by the whole on its parts. As we will argue (Sect. 2.5.1), there seems to be no compelling argument in favour of a positive answer within our framework, at least under the monist assumptions adopted so far. Alternative conclusions could be obtained (Sect. 2.5.2) by rejecting some of these assumptions, or by shifting the analysis to an epistemological or heuristic dimension.

Before continuing, a terminological clarification: A possible objection might contend that this debate somehow forces a narrow understanding of inter-level causation in terms of reflexive whole-parts causal influence, whereas the usual meaning in the biological domain refers to the non-reflexive case, where higher-

[17] Campbell defines downward causation as follows: "all processes at the lower level of a hierarchy are restrained by and act in conformity to the laws of the higher level" (Campbell 1974: 180). More recently, (Vieira and El-Hani 2008) have proposed a similar view, although they refer to "formal determination" instead of formal causation.

level entities interact with lower-level entities, the latter not being constituents of the former. Indeed, this interpretation of inter-level causation applies straightforwardly to biological organisation, and is inherently involved in the very notion of closure. In this sense, biological discourse requires a general concept of inter-level causation. To avoid ambiguities, we propose using different terms to refer to the two ideas: in what follows "inter-level causation" will be therefore used for the general non-reflexive case, and "nested causation" for the reflexive whole-parts case. This way, biological descriptions would be able to refer to inter-level causation, while at the same time avoiding incongruities with philosophical analyses.

2.5.1 Why We Do Not Need Nested Causation in Biology

The account of emergence and supervenience developed so far has relevant implications for the conception of nested causation.[18]

Concerning the supervenience base B – insofar as the principle of inclusivity of levels is maintained (but see Sect. 2.5.2 below), and the relation between an emergent property M and the configurational properties $S_1, \ldots S_n$ of B is conceived as constitutive – the exclusion argument applies more cogently to relational supervenience than to its atomistic version. Consequently, as (Craver and Bechtel 2007) emphasise, no nested causation can exist between an emergent property and its own supervenience base: there is no justification for claiming either that $S_1, \ldots S_n$ "generate" or "produce" M, or that M exerts downward causation on $S_1, \ldots S_n$. In particular, the closed organisation C_{org} does not exert causation on the whole network of constitutive constraints, and the whole network of constitutive constraints does not produce the closed organisation. Under the monist stance adopted so far, there is therefore no room for nested causation in the autonomous perspective.

Let us now consider the emergence base P of C, and its different versions discussed in Sect. 2.4. Is there nested causation between the organisation C_{org} and any *subset* P_{sset} of its constituents? In our view, by assuming the principle of the inclusivity of levels, the answer is no, since the properties of each P_{sset}

[18] In the philosophical literature, nested causation comes in two variants, synchronic and diachronic (Kim 2010: 34–36). On the one hand, *synchronic* nested causation refers to the situation in which upward and downward causation would occur simultaneously. In more technical terms, a supervenient property M acts causally on its supervenient base $S_1 \ldots S_n$ *at the same time* as the supervenience base generates M. On the other hand, *diachronic* (or diagonal) nested causation refers to the situation in which M acts on its own supervenience base $S_1, \ldots S_n$, causing its modification, but only at a subsequent time with respect to the upward determination. In this chapter, however, we assume that the distinction is irrelevant, since we question the very idea of the causal influence of M on $S_1, \ldots S_n$, be it synchronic or diachronic. In particular, in line with Sartenaer's detailed argument (Sartenaer 2013: 240–250), we assume in this chapter that all cases of diachronic nested causation are also synchronic and vice-versa.

(which may refer, for instance, to each individual constraint C_{constr}) are by definition configurational, so that the appeal to some constraint exerted by the whole would be redundant: configurational properties are so precisely because an entity belongs to a whole. To put it more straightforwardly, local constraints are not so "because they are under the causal influence of the whole". Also, no nested causation occurs between the whole and its *substrate* P_{sstr} because, in our account, the collection of its constituents taken separately (without their configurational properties) is an abstract description that does not correspond to the way in which constituents are organised in the system. Since, in the system, there is no such thing as a collection of unrelated constituents, they cannot be, *a fortiori*, involved in nested causation, or indeed any causation at all.

The case of the third kind of emergence base, the *surroundings* P_{surr}, is somewhat different. As discussed in Sects. 2.3 and 2.4, emergent configurations do exert a causal action on their surroundings, notably in the form of constraints. Yet, surroundings are by definition external to the configuration, which means that the constraints exerted by C_{org} on P_{surr} *can by no means be interpreted in terms of nested causation*.

The claim according to which constraints, in our framework, do exert causal powers, but not in the form of nested causation, has crucial consequences for the interpretation of self-maintenance and closure.

In the case of physical self-maintaining systems, the fact that the emergent configuration acts to maintain itself does not appear to constitute, *per se*, a case of nested causation, since the constraining action is exerted on the surroundings of the configuration, not on its own constituents. Let us again take the example of Bénard cells. An interpretation appealing to nested causation would claim that each cell (i.e. the emergent configuration) exerts a constraint on its own microscopic constituents, in the sense that the fact of belonging to a given cell *determines* whether a molecule rotates in a clockwise or anticlockwise direction. As (Juarrero 2009: 85) puts it:

> Once each water molecule is captured in the dynamics of a rolling hexagonal Bénard cell it is no longer related to the other molecules just externally; its behaviour is contextually constrained by the global structure which it constitutes and into which it is caught up. That is, its behaviour is what it is *in virtue of the individual water molecules' participation in a global structure*.

Yet, what we call the cell *is* the configuration of constituents, so that, as argued above, it is redundant to appeal to the whole set of constituents and relations to explain the behaviour of each constituent, whose characterisation already includes its relational properties as part of the configuration. Once a given molecule has been "captured" by the cell and has begun to rotate with the others, in what sense would it still be "constrained by the global structure"[19]? Let us have a closer look at this

[19] A satisfactory analysis of nested causation requires, then, a careful distinction between two ideas. One is the idea that a configuration is made up by a set of constituents, which have causal interactions between them. Explaining why a given molecule of water is rotating in a given manner at a given moment requires an appeal to its causal interactions with other constituents. And the

situation, the temporal steps being very important. At a given moment a macroscopic structure is formed. Once this structure is formed, it acts on the surrounding molecules, constraining their microscopic trajectories in such a way that they generate a thermodynamic flow that, in turn, contributes to the maintenance of the (otherwise decaying) macroscopic structure. The result is a causal regime in which the dissipative structure acts on its surroundings that, through this action, contribute to the maintenance of the very structure. Only in this loose sense the surrounding constrained processes could be said to being "parts" of the self-maintaining regime; yet, the causal action of the dissipative structure in each moment does *not* operate on its own constituents.[20]

Two reasons may explain why self-maintaining systems seem to be a case of nested causation. First, the description of the configurational properties of dissipative structures, which are available at a given moment, usually under-determines their behaviour. This is of course a crucial point; still, as discussed earlier, this should not be taken as sufficient reason for appealing to nested causal relations since, as pointed out in Sect. 2.3, it confuses epistemological non-derivability with ontological irreducibility (but see Sect. 2.5.2 below). Second, self-maintaining systems would not exist if they did not generate a causal loop between the whole configuration and its constituents. Yet, the crucial point is that, in our view, this loop is not a *direct* loop, but rather an *indirect and diachronic* one, realised through the action of the constraint on its surroundings. What might appear as an action exerted on the constituents is in fact exerted on the *boundary conditions* of these constituents.

In the light of these considerations, in particular, we do not think that the appeal to the supposed constraint exerted by the configuration on its own constituents in terms of *formal* causation is explanatory (again, under the monist assumptions adopted so far). The formal causation of the whole on its constituents would be in principle reducible to the constraining action exerted on the boundary conditions of these constituents, without loss of information or explanatory power.

---

reason why a set of constituents may exert a causal influence on other constituents is, of course, that all of them belong to the same system. The other idea, in contrast, is that the "whole system", including any specific constituent, would have a causal effect *on that very constituent*.

[20]This argument also applies to the relation between the whole and its parts in self-assembling structures. Let us take the example of protein folding. Protein folding is a process in which the parts – aminoacids – form small secondary (metastable) structures, which, once constituted, harness the surrounding interactions leading to the formation of new structures, and so on, till the global tertiary and quaternary folded structure is achieved. The whole folding process is a succession of formation of local wholes, acting as constraints on surrounding dynamical pieces that, later, will become wholes harnessing other pieces, and so on. Once the folding is achieved the whole has attained a relative thermodynamic equilibrium (it is a conservative structure) and it does not make sense to say that it acts constraining its parts. So, when we consider the temporal (diachronic) process of folding, the (local) "wholes" act on their surrounding, not on their constitutive parts; similarly, when we consider the protein already folded (synchronically) the whole does not act on its own parts either.

Let us now examine closure. Is there a characteristic aspect of closure that would justify, in contrast to simple self-maintenance, the claim according to which it realises nested causation?

The main difference between physical self-maintenance and closure is that, in the second case, self-maintenance is realised collectively by a network of mutually dependent constraints. In real biological systems, closure is realised through a very complex organisation of constraints, such that, in most cases, a given constraint exerts its action on surroundings that have already been subject to the causal influence of at least one other constraint. For instance, most enzymes act on reactions whose reactants are the result of the joint action of other constraints, including the membrane (through its channels and pumps). In these cases, it can be said that constraints act on entities that are already "within" the system, at least in the sense of having already been constrained by the system. This seems to be a clear difference with respect to simple self-maintaining systems, and one may then conclude that the closed organisation does act on its own constituents, and realises nested causation.

Yet we hold that this conclusion is incorrect, since it interprets those constrained processes and reactions as constituents of the organisation (which, we recall, is defined as a specific kind of higher-level configurations, as a closed configuration of mutually dependent constraints), whereas they are not. In biological systems, the constituents of the organisation are the constraints themselves, which realise collective self-maintenance. According to the constitutive interpretation of the relation between the whole and its constituents, the organisation as such does not possess emergent and distinctive causal powers with respect to the closed network of constraints which, in turn, exert causal powers on surroundings which are not themselves constituents of the network (although they are usually within the spatial borders of the system).[21] Accordingly, to use the terminological distinction introduced above, we maintain that closure does involve inter-level causation, but *not* nested causation.

A second reason why closure seems to inherently imply nested causation is that evoked by Kant (1790/1987), i.e. the fact that the existence of the constituents (the constraints) "depends on the whole". Indeed, the mutual dependence between constraints constitutes a fundamental difference between organisations and other configurations. In the second case the existence and maintenance of the constituents might not depend on their being involved in the configuration: one can decompose a wheel into its molecular elements, which would continue to exist as separate elements. The same holds for the microscopic constituents of a dissipative system. In contrast, closed organisations imply a more generative kind of relation between

[21]The physical processes on which the network exerts (constraining) causal powers can, in some cases, become members of the network itself, when they enter into configurations which act as constraints. Nonetheless, the network would exert causal powers on them for as long as they remained part of its surroundings, and would cease acting causally on them as soon as they started playing the role of constraints.

constituents (the constraints themselves), which exist only insofar as they are involved in the whole organisation. Actually, the appeal to formal causation advocated by several authors is essentially aimed, in our view, at capturing this distinctive feature of biological organisms.

Yet, these specific features of organisations do not require an ascription of distinctive causal powers to the whole, since closure can be realised through the network of mutual, usually hierarchical, causal interactions. Hence, "depending on the whole" could simply mean "depending on the whole network of interactions" without appealing to the whole as a causal agent emergent on its own supervenience base.

This interpretation of the whole-parts relation in biological organisation is particularly relevant because it applies to all those cases in which biological literature typically appeals to nested causation, i.e. all kinds of regulation and control mechanisms (see Chap. 1, Sect. 1.8 above) thanks to which organisms are able to (adaptively) compensate for internal and/or external perturbations (Piaget 1967; Fell 1997). What is frequently described as a causal action of the whole system on its own constituents, is in fact the result of the interaction among organised constraints (or subsystems of constraints) which can result, for instance, in the acceleration of the heart rate and glucose metabolism when the organism starts playing tennis (see Craver and Bechtel 2007: 559, for a detailed description of this example, and other relevant ones). In particular, inter-level control can be generally understood in terms of causal interactions among constraints located at different hierarchical levels of emergent organisation. In turn, as we claimed in Chap. 1, Sect. 1.8, regulation specifically concerns interactions among constraints at different orders of closure. Although both cases inherently require, as all biological functions do, the realisation of closure as well as inter-level (and inter-order) causation between hierarchically organised constraints, they do not involve nested causation exerted by the whole organism.

2.5.2 *Why We Might, After All, Need Nested Causation in Biology*

The rejection of nested causation depends on the constitutive interpretation of the supervenience relation adopted so far. It is an implication of our monist interpretation of the autonomous perspective. Indeed, the central goal of the analysis was to suggest that closure can be justifiably taken as an emergent and distinctively biological regime of causation *even* under a constitutive interpretation of supervenience. Yet, several strategies could be adopted to justify nested causation, and they might be successful and operational in some cases, including the biological domain, which is specifically under study here. To date, however, we believe that these strategies lack any compelling argument in favour of their adoption in Biology; their relevance is still under conceptual and scientific scrutiny. That is why, in our

view, the constitutive interpretation of the whole-parts relation is still the wiser one. Let us discuss these strategies.

The first strategy is ontological and advocates that a non-constitutive interpretation of relational supervenience should be adopted, in order to admit causation of the whole on the constituents. In this interpretation, emergent properties can be simultaneously supervenient on *and* irreducible to configurations. For this ontological stance to be coherent, one must accept the violation of the inclusivity of levels, hence accepting the idea that the very same entity (say: a constituent of a configuration) may possess different properties, and then obey different laws or principles, at different levels of description. In other terms, it consists in rejecting the monist stance advocated so far. For instance, each molecule constituting the wheel would have the property to behave in a given way when considering the whole configuration, but would *not* possess the same property when looked at individually. Even though we are looking at the very same molecules under the very same conditions, their properties would vary according to the level of description, since the relevant laws and principle would also vary.

In our view, rejecting the principle of the inclusivity of levels could indeed be an important tool for adequately accounting for natural phenomena that would therefore require an appeal to nested causation. We have no principled objections to this position. Yet, we maintain that its relevance for the biological domain is still uncertain. As (Craver and Bechtel 2007) have convincingly argued, many (or most) apparent biological examples of nested causation (in particular cases of downward regulation) seem to be adequately explainable by appealing to what they call "hybrid accounts" involving intra-level causal interactions *between* constituents and inter-level constitutive relations, or to what we dubbed inter-level (not nested) relations. In those cases, an advocate of the constitutive interpretation of mereological supervenience could argue that the appeal to nested causation seems precisely to stem from an inadequate understanding of the role of configurations: the behaviour of the constituents appears to be influenced by the whole because the description focuses only on the internal properties of the constituents, neglecting the relational ones.[22] In a word, there seems to be no clear case in the biological domain for

[22]In the case of the wheel, for instance, one may say that if we describe a given molecule as a constituent of a wheel, we are already including in the description all constitutive and relational properties, which make it a constituent ("being in such and such position", "having such and such interactions and links with neighbouring molecules" etc.), and which determine its behaviour under specific conditions. For instance, a force (i.e. gravity) applied to a part will generate a chain of causal interactions between the constituents that, because of their individual configurational properties, will behave in a specific way. We will then call the collective pattern the "rolling movement of the wheel". Each molecule of the wheel will move in a specific way because its configurational properties force it to do so, and a complete description of the configurational properties of the individual constituent will suffice to explain why it behaves as it does. The fact that the constituents collectively constitute a wheel, whose macroscopic behaviour can be described as a rolling movement, does not add anything to the explanation of the individual behaviour. There are indeed causal interactions here, but no inter-level causation.

which the appeal to nested causation is mandatory. To a first approximation, self-maintenance and closure are no exceptions in this respect.

The second strategy is epistemological, and consists of justifying nested causation by demonstrating that it would be impossible, *in principle*, to determine the behaviour of a system through a description of its configurational properties. On the basis of such a negative result, the appeal to nested causation would be justified in epistemological terms, since there would be, in principle, no alternative description.[23] Yet, while arguments of "inaccessibility" have already been formulated in physics (Silbersten and McGeever 1999), and might for instance be relevant for describing dissipative structures, this is not the case in biology.[24] Consequently, there seems, to date, to be no compelling epistemological argument for admitting nested causation for biological systems.

The third strategy is heuristic. As a matter of fact, there are many cases, especially in complex systems, in which the available description of the configurational properties is insufficient to adequately determine the behaviour of the whole system. In these cases, which are indeed widespread, it might be useful to appeal to the configuration as a whole *as if,* by virtue of its emergent properties*,* it were exerting nested causation on its constituents, so as to provide a description capable of sufficiently determining the behaviour of the system. Since it is not committed to a theoretical non-constitutive interpretation of supervenience, the heuristic appeal to nested causation can be adopted as a pragmatic tool even within a constitutive interpretation of supervenience. Yet, such a heuristic use of nested causation would not point to any specific feature of the causal regime at work in biology (which is the object of this chapter), but would simply correspond to a scientific practice common to several scientific domains. In particular, as mentioned above, the strategy can be adopted for self-maintaining, closed systems for which, mostly because of their internal complexity, complete descriptions of their organisation are difficult to elaborate.

2.6 Conclusions

In order for closure to be a legitimate scientific concept rather than merely an epistemic shortcut, philosophical arguments must be provided in favour of its emergent and irreducible nature with respect to the causal regimes at work in other classes of natural systems. To do this, we developed a twofold argument.

[23] See (Bich 2012) for an epistemological discussion of the relationship between emergence and downward causation.

[24] It should be noted, however, that the issue is currently being explored by several biologists and theoreticians. For instance, a relevant proposal in this direction has recently been developed by Soto et al. (2008).

On the one hand, we argued that constraints are configurations that, by virtue of the relations existing between their own constituents, possess emergent properties enabling them to exert distinctive causal powers on their surroundings, and specifically on thermodynamic processes and reactions. When a set of constraints realises closure, the resulting organisation constitutes a kind of second-level emergent regime of causation, possessing irreducible properties and causal powers: in particular, organisations are able to self-determine (and more precisely to self-maintain) as a whole (something which none of their constitutive constraints are able to do). As we will see in the following chapters, most of the distinctive features of autonomous systems specifically rely on closure and organisation, which therefore play, as an emergent causal regime, a pivotal role in the autonomous perspective.

On the other hand, we advocated the idea that a coherent defence of closure as an emergent and irreducible causal regime does not need to invoke nested causation. Closed organisations can be understood in terms of causal (possibly inter-level) interactions between mutually dependent (sets of) constraints, without implying upward or downward nested causal actions between the whole and its parts. Biological emergence, accordingly, is logically distinct from nested causation, and one can advocate the former without being committed to the latter.

Again, we by no means wish to exclude the possibility that biological organisation might involve nested causation. As discussed earlier, various strategies could be adopted to adequately justify this idea, and promising explorations are currently underway. Nevertheless, we believe that these attempts are, as yet, incomplete, and do not offer compelling arguments in the biological domain. That is why we argue that biological organisation can be shown to be emergent and irreducible *even though* nested causation is, by hypothesis, ruled out.

3
Teleology, Normativity and Functionality

According to the autonomous perspective, the constitutive organisation of biological systems realises an emergent regime of causation, which we labelled *closure of constraints*. One of the crucial implications of the realisation of closure is that, as we will argue in this chapter, it provides an adequate grounding for a distinctive feature of biological systems, namely their *functionality*.

The concept of function is widespread in the language of all life sciences. At the scale of individual organisms, functions are usually ascribed to a variety of structures, traits, or processes that constitute the whole, such as, for instance, systems, organs, cells, and molecules. Similarly, functions are invoked when considering larger scales, so that organisms themselves, as well as populations and species, may be the object of functional ascriptions. Moreover, as Gayon points out (Gayon 2006: 480), functional ascriptions mostly tend to have a nested structure: parts of a functional entity can also perform functions and, reciprocally, systems containing functional entities may also be described as functional.

What is the status of the concept of biological function? At first glance there seems to be a broad consensus regarding the idea that functions play a genuine explanatory role in biology and the other life sciences: functional ascriptions are by no means simple descriptions of a trait, but rather provide an understanding of some of its essential properties and activities. To be sure, the explanatory role of functions seems to be so fundamental in life sciences that one could argue that biological explanations are *essentially* functional: in contrast to those at work in, for instance, physics or chemistry, biological explanations would be specific in this, i.e. in that they appeal to functions.

Many of the ideas developed in this chapter, as well as several portions of the text, come from Mossio et al. (2009b), and Saborido et al. (2011). Section 3.3.2 is partially based on Saborido's account of malfunction, exposed in his PhD dissertation (Saborido 2012) and in Saborido et al. (2014).

A. Moreno, M. Mossio, *Biological Autonomy*, History, Philosophy and Theory of the Life Sciences 12, DOI 10.1007/978-94-017-9837-2_3

Even though it is not our aim to adopt a final position in relation to this last issue, it cannot be denied that the concept of function is at the very heart of scientific discourse in life sciences. Yet it generates a major epistemological problem, since it seems to be, at least at first sight, at odds with the ordinary structure of scientific explanation, because of its characteristic dimensions, i.e. its *teleology* and *normativity*. But what does this actually mean?

On the one hand, functions have an explanatory role in accounting for the existence of function bearers. Affirming that (to cite a classic example) "the function of the heart is to pump blood" does not correspond to a simple description of what the heart does; rather, in addition, it means that this effect has specific relevance in explaining the existence, structure and morphology of hearts (see also Buller 1999: 1–7). Hearts exist to some extent because they pump blood. Functional attributions thus introduce a teleological dimension into the structure of explanation, in the sense that the existence of a trait could be explained by appealing to some specific effects or consequences of its own activity, which reverses the conventional order between causes and effects.

On the other hand, the concept of function possesses a normative dimension, to the extent that it refers to some effect that the trait is *supposed* to produce (Hardcastle 2002: 144). Attributing functions to a trait implies a reference to some specific norm, against which the activity of the trait can be evaluated. The claim that "the function of the heart is to pump blood" implies also that the heart must pump blood. Whereas, usually, causal effects simply occur, functional causal effects must occur.

Because of its teleological and normative dimensions, the concept of function seems then to be in conflict with the accepted structure of scientific explanation. The central question is then: is the concept of function a legitimate and admissible scientific concept?

To answer this question, two alternative strategies are possible. The first is an eliminativist one, and consists of denying that functions do in fact play an explanatory role. All functional claims can be reformulated in terms of an ordinary causal claim, without losing information or meaning. In this case functions would constitute, at best, a linguistic shortcut. The second strategy, in contrast, claims that while functional statements cannot be reduced to ordinary causal ones, they are compatible with the structure of scientific discourse. In this case, a naturalisation of teleological and normative dimensions is required, i.e. a justification of the idea that these dimensions are grounded in some objective features and properties of biological systems and, consequently, can be analysed in adequate scientific terms.

In this chapter, we will suggest that the autonomous perspective adopts the second strategy, and puts forward a naturalised "organisational" account[1] of functionality, based on the emergent properties of closure.

[1]For terminological clarity, note that we will dub "organisational account" (OA) the account of functions stemming from the view of living beings as organisationally closed systems, and, in particular, from the autonomous perspective.

3.1 The Philosophical Debate

Broadly speaking, the philosophical analysis of the concept of function is very old, to the extent that it has always developed hand in hand with scientific research into biological phenomena. However, the debate on functions, in its contemporary form, has been framed during the last four decades, during which an increasing number of studies have been conducted in philosophy of science and philosophy of biology (several collections have published that survey the recent philosophical debate: see Ariew et al. 2002; Buller 1999; Allen et al. 1998; Gayon and de Ricqlès 2010).

The contributions that gave rise to the contemporary debate were formulated during the sixties by Nagel (1961, 1977) and Hempel (1965) who, by adopting an eliminativist stance, tried to reduce functional statements to the nomological-deductive model (Hempel and Oppenheim 1948). Because of the difficulties inherent in their approach (Saborido 2012: 51–59), the vast majority of subsequent literature has focused on justifying functional discourse through naturalisation.

Current philosophical accounts of functions are usually grouped into two main traditions, called "dispositional" (or "systemic") and "aetiological". As we will argue, the autonomous perspective advocates a third one, the "organisational" view, which aims to combine the previous accounts into an integrated framework. Before expounding our own view, we shall first provide a brief overview of the other two accounts, and describe their respective strengths and weaknesses.

3.1.1 Dispositional Approaches

In the philosophical debate on functions, several authors have, against the eliminativist stance, advocated the idea that functional attributions do indeed refer to current features of the system under examination. By explicitly discarding teleology as a constitutive dimension of the concept of function, these authors hold that functions do not refer to a causal process that would explain the existence of the function bearer by appealing to its effects. Rather, functional relations are interpreted as a particular class of causal effects or dispositions of a trait – means-end relationships contributing to some distinctive capacity of the system to which they belong.[2]

The philosophical agenda of dispositional approaches focuses on providing naturalised and appropriate criteria for identifying what counts as a target capacity of a functional relationship, from which the relevant norms can be deduced, and the different dispositional approaches have proposed various criteria to identify these target capacities.

[2]On the basis of this common theoretical stance, these approaches have been labelled "causal role", "dispositional" or "forward-looking", as opposed to "backward-looking" etiological ones. Here we will use the general label "dispositional" to refer to this class of theories.

The more classical dispositional approach is the "systemic approach" (SA), which defines a function *F* as the contribution of a process *P* to a distinctive higher-level capacity *C* of the system *S* to which it belongs (Craver 2001; Cummins 1975; Davies 2001). In the SA, explaining functions means analysing a given higher-level capacity of the system in terms of the capacities of the system's components, which jointly concur in the emergence of the higher-level capacity. The SA dissolves the problem of teleology of functions by reducing them to any causal contribution to a higher-level capacity. In turn, the normative dimension of functions is reduced to the fact that the causal effect must contribute to a higher-level capacity, with no reference to a "benefit" for the system.

The explanatory strategy adopted by the SA is subject to one major criticism, namely that it seriously *under-specifies* functional ascriptions, which in turn generates several problems (see also Wouters 2005). Firstly, the SA fails to draw a principled demarcation between systems whose parts appear to have functions and systems whose parts do not (Bigelow and Pargetter 1987; Millikan 1989). Secondly, the SA lacks a principled criterion for identifying the relevant set of contributions for which functional analysis makes sense. And thirdly, the SA is unable to draw an appropriate distinction between "proper" functions and accidental, useful contributions (Millikan 1993, 2002).

Because of these fundamental weaknesses of the SA, the "goal contribution approach" (GCA) has attempted to introduce more specific constraints on what makes causal relations properly functional, by linking the concept of function to the cybernetic idea of *goal-directedness*. In particular, the GCA restricts functional attributions to causal contributions to those (higher-level) capacities that constitute the "goal states" of the system (Adams 1979; Boorse 1976, 2002; Rosenblueth et al. 1943). In particular, biological systems can be described as having the essential goal of surviving (and reproducing). Hence, biological functions are dispositions that contribute to these goals.

The main virtue of the GCA is that it provides an interpretation of functions that, in contrast to the SA, recognises and substantiates their specificity as means-end causal relationships. Nevertheless, the characterisation of a goal-directed system introduces norms whose application is not restricted to the relevant kinds of systems and capacities. As Bedau (1992) points out, the cybernetic characterisation of the goal state is unable to adequately capture the frontier between "genuinely" goal-directed systems (supposedly biological systems and artefacts) and physical equilibrium systems, which tend to some steady state or state of equilibrium (see also Nissen 1980).

Moreover, as Bedau (1992) and Melander (1997) argue, cybernetic criteria may interpret the dysfunctional behaviours of goal-directed systems as functional and, also, the GCA account lacks the theoretical resources to distinguish between functions and accidental contributions to a goal state. In sum, the GCA still seems to under-specify functional attributions, and in some cases it appears even to be a less satisfactory account than the SA.

The third main dispositional perspective proposes the identification of functions with causal contributions of components to the life chances (or fitness) of the system

(Bigelow and Pargetter 1987; Canfield 1964; Ruse 1971). Bigelow and Pargetter, in particular, have proposed the "propensity view", according to which "something has a (biological) function just when it confers a survival-enhancing propensity on the creature that possesses it" (Bigelow and Pargetter 1987: 108).

By appealing to survival in terms of enhancing propensities as the goal of a functional relation, the propensity view succeeds in restricting functions to components of biological entities. Moreover, Bigelow and Pargetter's reference to survival-enhancing *propensity* is intended to avoid functional attributions to contingent and/or accidental contributions to survival, which would be contrary to intuition and common use. Yet, as McLaughlin perceptively argues (McLaughlin 2001: 125–8), the appeal to propensities does not fully succeed in restricting functional attributions to the relevant cases. Even by restricting propensity to the current environment (the "natural habitat", in Bigelow and Pargetter's terms), it is possible to imagine, for each specific effect produced by a trait, a situation in which that specific effect would confer a (possibly low) propensity that enhances survival, and thus have a function.

The problem is that propensities to enhance survival in virtual (but not impossible) situations correspond, in a forward-looking approach, to *actual* functions of the existing trait. Moreover, to the extent that the specific contribution of the trait would presumably change in accordance with the particular condition in question, each trait in fact possesses an indefinite list of actual functions. Again, the propensity view fails to provide an adequately restricted definition of what counts as a functional relation. All (biological) functions are survival-enhancing contributions, but not all survival-enhancing contributions are functions. Appealing to propensities does not solve the problem.

To summarise, the main virtue of the dispositional approaches is their capacity to capture the fact that the concept of function points to something more than mere causal relations: functions refer to current means-end relationships, and more specifically to current contributions of components to the emergence of a target capacity of the containing system. Yet, dispositional approaches in the end fail to provide a satisfactory grounding for the normativity of functional attributions, and dispositional definitions turn out to be systematically under-specified, allowing functional ascriptions to irrelevant systems and/or capacities. In a word, the price paid for excluding the teleological dimension as a proper *explanandum* is not compensated for by a satisfactory foundation of the normative dimension.

In fact, most of the existing literature has favoured a different approach, according to which an adequate understanding of functional attributions has to deal with the problem of teleology. In particular, both the teleological and normative dimensions are conceived as being inherently related to the *aetiology* of the functional trait.

3.1.2 Aetiological Theories

The mainstream philosophical theory of functions is the aetiological approach (Wright 1973, 1976; Millikan 1984, 1989; Neander 1980, 1991; Godfrey-Smith

1994). The aetiological approach defines a trait's function in terms of its aetiology (i.e. its causal history): the functions of a trait are past effects of that trait that causally explain its current presence. In sharp contrast with dispositional accounts, the aetiological approach explicitly takes the issue of teleology as the central problem of a theory of functions.

The first aetiological approach was proposed by Wright, who defined functions as follows:

The function X is Z means:

1. X is there because it does Z.
2. Z is a consequence (or result) of X's being there (Wright 1976: 48).

Wright's definition explicitly appeals to a form of causal loop, in which the effect of a trait helps to explain – teleologically – its existence. The scientific validity of Wright's definition has been questioned and, moreover, several obvious counterexamples have been formulated (see, for instance, Boorse 1976).

In order to ground the teleological dimension of functions without adopting an unacceptable interpretation of the causal loop described by Wright, mainstream aetiological accounts, usually called "selected effect (SE) theories", have appealed to the Darwinian concept of Natural Selection as the causal process, which would adequately explain the existence (or, more precisely, the maintenance over time) of the function bearer by referring to its effects. The gist of SE theories is that functional processes are not produced by the same tokens whose existence they are supposed to explain. Instead, the function of a trait is to produce the effects for which past occurrences of that trait were selected by Natural Selection (Godfrey-Smith 1994; Millikan 1989; Neander 1991). Selection explains the existence of the *current* functional trait because the effect of the activity of *previous* occurrences of the trait gave the bearer a selective advantage. The main consequence of this explanatory line is its historical stance: what makes a process functional is not the fact that it contributes in some way to a present capacity of the system, but rather that it has the right sort of selective history.

By interpreting functions as selected effects, SE theories are able not only to deal with the problem of teleology, but also to ground the normativity of functions. By defining functions as effects subject to an evolutionary causal loop, SE theories identify the norms of functions with their *evolutionary conditions of existence*: the function of a trait is to produce a given effect because *otherwise*, the trait would not have been selected, and would not therefore exist.

Several virtues of SE theories are often emphasised, including their capacity to exclude functional attributions to traits of physical systems, and their ability to unambiguously identify functions from among the whole set of all processes occurring in a system and to draw a boundary between functions and accidental useful effects. Nevertheless, SE theories have their own weaknesses, which have been extensively discussed in the literature (see, for instance, Boorse 1976; Cummins 2002; Davies 1994, 2000). We will focus here on one specific weakness of the theories, which Christensen and Bickhard (2002) have labelled their *epiphenomenalism*. The crucial drawback of SE theories' explanatory line is the implication that

functional attributions bear no relation to the *current* contribution of the trait to the system, since they point solely to the selective history of the trait. This is at odds with the fact that functional attributions to biological structures do seem to bear some relation to what they currently do, and not only to what explains their current existence.

To solve some difficulties inherent to previous formulations of aetiological theories (mainly that they attribute proper functions to effects that are, in fact, no longer functional in the current system), Godfrey-Smith (1994) has proposed a "modern history theory" of function. In his approach, functions are "dispositions or effects a trait has which explain the recent maintenance of the trait under natural selection" (Godfrey-Smith 1994: 199; See also Griffiths 1993). While it successfully counters several objections raised against previous versions of the theory, Godfrey-Smith's account is no better placed to deal with the problem of epiphenomenalism. More precisely, as McLaughlin (2001: 116) points out, by reducing the cases in which it attributes functions to currently non-functional traits, Godfrey-Smith's account (which is explicitly an historical one) possibly reduces "uncooperative cases", but does not provide a principled solution to the problem.

Accordingly, SE theories provide an account that is problematically epiphenomenal, in the sense that it maintains that the attribution of a function does not provide information about the current system being observed. From the perspective of SE theories, a function does not tell us anything about the current organisation of the system being analysed.

3.2 The Organisational Account of Functions

The outcome of this brief critical survey is that current theories of functions seem to face a dilemma, arising from the way in which they deal with the two main issues related to the concept of function, i.e. its teleology and its normativity. Dispositional theories try to account for functions in terms of current contributions to some target capacity of a system, and discard the teleological dimension, but seem unable to provide fully adequate normative criteria for functional attributions. Aetiological theories, on the other hand, try to account for both the teleological and normative dimensions of functions, but appear inevitably historical and are unable to justify how functional attributions may refer to features and properties of the current system.

According to some authors, the solution to the dilemma consists of concluding that there is no unified account of functions, and that aetiological and dispositional approaches provide two different yet complementary concepts of function (Allen and Bekoff 1995; Godfrey-Smith 1994; Millikan 1989). Other authors, such as Kitcher (1993), Walsh (1996), and Walsh and Ariew (1996), have claimed that there is, in fact, a single concept of function, in which the aetiological and dispositional formulations can be subsumed as special cases. In this section, we argue that, from the autonomous perspective, there is indeed room for a unified account of functionality, based on the properties of self-determination of biological organisation.

The core of the organisational account (OA) is the idea that functional ascriptions do account *at the same time* for both the existence of functional traits and their current contribution to a system capacity, since functions make sense only in relation to the specific kind of organisation which is characteristically at work in biological organisms. In particular, as we shall argue, functions correspond to those causal effects exerted by the constraints subject to closure that contribute to maintaining the organisation.

Before expounding our own version of the OA, it should be mentioned that, very recently, a considerable amount of work has been done in this direction by Bickhard (2000, 2004), Schlosser (1998), Collier (2000), McLaughlin (2001), Christensen and Bickhard (2002), Delancey (2006), Edin (2008), and more recently by ourselves (Mossio et al. 2009b; Saborido et al. 2011; Saborido and Moreno 2015). In spite of some differences between the various formulations, there seems to be substantial convergence[3] regarding the fundamental tenets of the OA, which makes it a credible philosophical alternative to both aetiological (mainly in its "selected-effects" version) and systemic-dispositional accounts.

3.2.1 Teleology, Normativity and Self-Determination

The OA relies on an understanding of biological systems as sophisticated and highly complex examples of natural self-maintaining systems. In particular, the first claim of the OA is that self-determination, as characterised in Chaps. 1 and 2, constitutes the relevant emergent causal regime in which the teleological and normative dimensions of functions can be adequately naturalised.

On the one hand, the causal regime of a self-maintaining system provides a naturalised grounding for the teleological dimension. Since the activity of the system S contributes, by exerting a constraint on its surroundings, to the maintenance of some of the conditions required for its own existence, the question "Why does S exist?" can be legitimately answered by "Because it does Y". This justifies explaining the existence (again, in the specific sense of its *maintenance* over time) of a system in "teleological" terms by referring to its causal effects.

On the other hand, self-maintenance grounds normativity. The activity of a self-maintaining system has an intrinsic relevance for itself, to the extent that its very existence depends on the constraints exerted through its own activity. Such intrinsic

[3]Christensen and Bickhard (2002) have suggested, relying on their own work on the notion of biological autonomy, that the organisation of autonomous systems provides an adequate grounding for the normativity of functional attributions. In a similar vein, McLaughlin (2001) has developed an account in which both the teleology and normativity of functions can be naturalised in the organisation of self-reproducing systems. Despite some terminological differences, the central idea of these approaches (i.e. that the organisational closure instantiated by living systems provides an adequate basis for naturalising functions) fundamentally coincides with that defended here, and we explicitly recognise this theoretical relationship.

relevance generates a naturalised criterion for determining what norms the system is supposed to follow: the system must behave in a specific way, otherwise it would cease to exist. Accordingly, the activity of the system becomes its own norm or, more precisely, its conditions of existence are the intrinsic and naturalised norms of its own activity.

Note that, so far, we have been generally referring to self-maintenance, and not closure. Hence, we acknowledge that the grounding of the teleological and normative dimensions goes beyond the biological domain, and includes some kinds of physical and chemical self-maintaining systems. Let us take the simple example of a candle flame. As Bickhard (2000: 114) points out, by constraining its own surroundings, the flame makes several contributions to the maintenance of the conditions required for its own existence. Indeed, the flame keeps the temperature above the combustion threshold, vaporises wax and induces convection (which pulls in oxygen and removes combustion products). Accordingly, to the question "Why does the flame exist?" it is legitimate to answer "Because it does X": the existence of the combustion reactions (the flame itself) is explained (at least in part) by taking into account the effects of its constraining action. Moreover, what the flame does is relevant and makes a difference for itself, since its very existence depends on the specific effects of its activity. The conditions of existence of the flame are the norms of its own activity: the flame must behave in a specific way, otherwise it would disappear.

One may object that, if self-maintenance as such provides the relevant grounding for teleology and normativity, then the OA should allow functions to be ascribed to physical dissipative systems. But of course, this implication seems unsatisfactory since, usually, no one ascribes functions to physical systems. Hence – the objection could continue – the OA clearly fails to restrict functions to the relevant kind of systems, just as dispositional approaches do. To this objection, we reply by formulating the second claim of the OA, according to which self-maintenance is a necessary but not sufficient condition for grounding functions in a naturalised way. Functions emerge when the self-maintenance is realised in the specific form of closure.

3.2.2 Closure, Organisation and Functions

The second claim of the OA is that when self-maintenance is realised as closure, then the causal effects of the constraints subject to closure are functional. Accordingly, as we claimed in Chap. 2, functionality is an emergent property of closure. Closure of constraints is therefore closure of functions.

Before providing a more precise definition and exploring some implications, let us clarify what is behind the conception of functionality advocated by the autonomous perspective.

The central idea is that functionality, in addition to teleology and normativity, includes a third dimension, that of *organisation*. Functions, we submit, involve the fact that self-determination is achieved through the interplay of a network of

mutually dependent entities, each of them making *different* yet *complementary* (and also *hierarchical*, as in the cases of regulation and control, discussed in Chap. 1, Sect. 1.8) contributions to the maintenance of the boundary conditions under which the whole system can exist. In other words, to ascribe functions we must distinguish between different causal roles in the system, a division of labour among the parts. And, of course, this is precisely what happens when closure of constraints is realised. As clarified above, closure is realised as the mutual dependence of the whole set of constraints which collectively achieve self-determination. But the very idea of mutual dependence presupposes that the various constraints produce different yet complementary causal effects: if all constraints produced the same effect, they would not depend on each other, and each constraint would be able to self-maintain individually. That is why, in our view, functions are not ascribed to dissipative structures. As discussed earlier, in this case there is only a single entity (the macroscopic structure itself) that acts as a constraint on the surroundings, and contributes to maintaining the conditions of its own existence. Since there is no need to distinguish between different contributions to self-determination generated by different constraints, functional ascriptions are meaningless.

At this point, it is important to set out one general implication of the autonomous perspective. The concepts "closure" and "organisation" are inherently linked. In the technical sense defined in Chap. 1, an organisation appears precisely when a set of constraints realise closure. Here, we add a third dimension. To the extent that closure is taken as the naturalised ground of functions, it follows that the concept of functionality itself is theoretically linked to that of closure and organisation. "Functionality", "closure", and "organisation" are then *mutually related concepts*, which refer to the very same causal regime; in other words, in the autonomous perspective an organisation is by definition closed and functional.[4]

Functional ascriptions and explanations are relevant as soon as the kind of organisational complexity realised by closure comes into being. Accordingly, it might be useful to focus on the distinction between the organisational and what could be labelled the "material" complexity of a system, i.e. the variety of its internal components. Minimal self-maintaining systems may indeed differ considerably with respect to their material complexity. Whereas many physical dissipative systems possess a rather homogeneous nature in terms of the variety of molecules of which they are made up (e.g. whirlwinds and Bénard cells), other systems, including chemical dissipative systems such as candle flames, have many different molecular components. Certain types of dissipative chemical systems (the Belousov-Zhabotinsky reaction, for instance) may even possess a high degree of material complexity.

Even high material complexity, however, has nothing to do with organisational closure, and therefore does not imply functions. In the case of the flame, for instance, the different chemical components all "converge" to generate a single macroscopic

[4]Of course, the reciprocal equivalences hold equally: closure refers to a functional organisation, and functionality indicates a closed organisation.

pattern (the flame), which in turn constrains the surrounding dynamics. Accordingly, it is not possible in this case to distinguish between the different ways in which the various components contribute to the self-maintenance of the system. The flame, although materially quite complex, is organisationally simple: in fact it has no organisation at all. Hence, functional attributions to components of the flame, as well as to all physico-chemical dissipative structures, are not meaningful. The realisation of closure requires not only that different material components be recruited and constrained to differentially contribute to self-maintenance but, in addition, that the constraints which contribute to self-determination be generated, and maintained, within and by the organisation of the system.

Let us now give an explicit and formal definition of function. According to the organisational account, a trait T has a function if, and only if, it exerts a constraint subject to closure in an organisation O of a given system. This definition implies the fulfilment of three different conditions (Saborido et al. 2011):

C_1. T exerts a constraint that contributes to the maintenance of the organisation O;
C_2. T is maintained under some constraints of O;
C_3. O realises closure.

Let us apply this definition to the classic example of the heart. The heart has the function of pumping blood since (C_1) pumping blood contributes to the maintenance of the organism by allowing blood to circulate, which in turn enables the transport of nutrients to and waste away from cells, the stabilisation of body temperature and pH, and so on. At the same time, (C_2) the heart is maintained under various constraints exerted by the organism, whose overall integrity is required for the ongoing existence of the heart itself. Lastly (C_3), the organism realises closure, since it is constituted by a set of mutually dependent structures acting as constraints, which, by contributing in different ways to the maintenance of the organisation, collectively realise self-maintenance.

It should be underscored that this characterisation of functions is consistent with the one proposed by Wright. In this example, the heart is there because it pumps blood (otherwise the organism, and thus the heart, would disappear), and pumping blood is a consequence of the heart's being there. This consistency stems from the fact that the organisational account, by appealing to a causal loop at work in the organisation of the system, provides an argument for naturalising both the teleology and normativity of functions, which, at an organisational level, mirrors the explanatory strategy adopted by the aetiological approaches. The resulting account represents an integration of the aetiological and dispositional perspectives, since it may at the same time explain the existence of the trait and its current contribution to the maintenance of the system.[5]

[5]In a recent contribution, Artiga (2011) offered a detailed critical analysis of the organisational account. Some of his remarks have been taken into account in the present formulation of the OA, while others (with which we do not agree) would require a full reply; but we will leave this for a future analysis.

The organisational definition given above is very general, and aims at encompassing all particular cases. Yet, actual functional ascriptions would take into account the complexity of autonomous organisation. This means, first of all, that functional ascription could vary according to the specific instance of closure that the system is realising at a given moment (what we called a "regime of self-maintenance"). Also, for each specific constitutive regime, as discussed in Chap. 1, Sect. 1.8 above, autonomous systems can realise different *orders* of closure, in particular insofar as regulation is involved. Moreover, in Chap. 6 we will discuss how different *levels* of closure (and then of organisation) can be described in certain classes of biological organisms, in particular multicellular ones.

Each specific regime, orders and levels of closure generate, as argued in this chapter, a distinct set of norms and functions. For instance, a given function could be related either to an individual cell (first-level) or to the whole multicellular organism (second-level) to which that cell belongs; in each of these cases, that very function could be either constitutive (first-order) or regulative (higher-order). And that function could be at work only within a specific regime of maintenance of the considered system, realised, for instance, only in some particular conditions or at a given moment. As a consequence, adequate functional ascriptions should make explicit, in each specific case, which are the regime, order, and level of the closure involved in C_3.

Lastly, it should be noted that, in principle, for each constraint subject to closure, functional ascriptions may concern either the *structure* itself (the trait) or the *effects* produced by that structure. Although the second option would possibly be more precise, here we mainly refer to the functions of traits and structures, which is consistent with the typical use of functional ascriptions in the relevant literature, as well as in ordinary language (see also Wimsatt 2002: 179).

3.3 Implications

The organisational account of functions has several relevant implications for the philosophical debate. Some of them[6] have already been spelled out in a previous study (Mossio et al. 2009b), and shall not be discussed here. In this section, we will focus on two main issues that are of crucial importance for assessing the scope and prospects of the OA: the ascription of cross-generation functions and the characterisation of malfunctions.

[6]For example, the distinction between functionality and usefulness; or the relationship between the concept of primary functions and the aetiological concept of proper functions.

3.3.1 Cross-Generation Functions

A major theoretical challenge facing the organisational account concerns, as Delancey (2006) has argued, the capacity to ground "cross-individual functions", i.e. those functions which go beyond the boundaries of individual biological systems. Let us explain what exactly this challenge consists of.

In the OA, functions are characterised as contributions of parts to closed organisations, and since closed organisations are typically realised by individual organisms, the OA appears to have trouble grounding those functions involving several individuals and their interactions. In particular, it is unclear whether and how the organisational approach would account for what Schlosser (1998) calls "cross-generation functions", for instance, the function of reproductive traits (e.g. the function of semen to inseminate the ovum). In these cases, in fact, the trait seems to contribute to maintaining the organisation of a system that is different from the system of which it is a component. Hence, the trait does not contribute either to the maintenance of an organisation or to its own self-maintenance. Still, we do ascribe cross-generation functions, just as we do, for instance, in the case of the reproductive function of semen. At first sight, then, cross-generation functions constitute a major group of counterexamples within the organisational approach.

As we explained in a previous work (Saborido et al. 2011), some of the authors who advocated the organisational account were of course aware of this issue, and proposed (following very different paths) solutions, which were designed to enable the account to embrace both intra- and cross-generation biological functions. Broadly speaking, the existing formulations can be regrouped into two main versions. The first version, advocated by Schlosser (1998) and McLaughlin (2001), tends to characterise reproductive functions as states or processes, which are causally required for the reproduction of the trait that causes them. The emphasis is therefore on the self-reproduction of the trait, rather than specifically on the whole system that, nevertheless, must possess the adequate properties to enable trait self-re-production. The second version, proposed by Collier (2000), Christensen and Bickhard (2002), shifts the focus onto the organisation of the system, and interprets reproductive functions as contributions to a higher-level self-maintaining organisation.

Delancey's analysis criticises all these "unified accounts" by pointing out their weaknesses and drawbacks. As an alternative, he proposes a "splitting account", according to which intra- and cross-generation functions are in fact two different kinds of biological functions, requiring different conceptual treatment within an organisational account.[7]

[7] We do not describe Delancey's account here. For details, see Saborido et al. 2011.

It is our contention, however, that the OA may provide a unified definition applying to both intra- and cross-generation functions. The essence of our argument will be that cross-generation functions contribute to the maintenance of systems, which realise a closed self-maintaining organisation in the very same sense as that of systems whose parts are ascribed intra-generational functions. To the extent that the two kinds of systems do not differ in terms of organisational self-maintenance, there is no need to invoke two kinds of functions, and the ontological problem is therefore overcome.

Before developing our own formulation, let us briefly discuss another proposal, put forth by Christensen and Bickhard (2002), which also tries to provide a unified account of intra- and cross-generation functions within the autonomous perspective. Their central move consists of appealing to higher-level organised self-maintaining systems, composed of individual organised self-maintaining organisms, in which reproductive traits could be subject to closure. In particular, Christensen and Bickhard explicitly grant systems like populations or species the status of autonomous[8] systems, making them relevant supports for functional ascriptions, just like individual organisms:

> Living organisms in general are autonomous systems, as are reproductive lineages, species, and some kinds of biological communities (Christensen and Bickhard 2002: 3).

As a consequence, intra- and cross-generation functions are simply contributions to the maintenance of different specific systems, sharing the same kind of organisation at different scales. Whereas intra-generation functions would contribute to the autonomous organisation of individual organisms, cross-generation functions would contribute to the autonomous organisation of the lineage, the species or the biological community in question.

Christensen and Bickhard offer an elegant alternative to the splitting account by admitting the idea of higher-level autonomous systems, namely, systems that would include individual organisms as parts, and that would ground the ascription of cross-generation functions. Accordingly, the heart is functional because it contributes to the autonomy of each individual vertebrate organism, while semen is functional because it contributes to the autonomy of the species.

Yet, this solution is problematic, as Delancey's lucid criticism (Delancey 2006) shows. As he points out, considering those higher-level systems that are relevant for grounding cross-generation functions as autonomous systems does not come without a price. Whereas an individual organism is a paradigmatic case of an autonomous system, "the sense in which the species or some population is a complex system of the appropriate kind is much more difficult to discern" (Delancey 2006: 90). For instance (and the list could be longer), such higher-level systems

[8]Christensen and Bickhard use the term "autonomy" in a somewhat weaker sense than the one developed in this book. Note that, in our account, closure is a *sufficient* requirement for grounding functions: in other words, functional systems are not necessarily autonomous systems.

have no clear boundaries, no stable form and, above all, it is very hard to see how their own "internal" organisation would realise closure, as is the case for individual autonomous systems.

According to Delancey, the organisational account has not explored these radical differences with sufficient accuracy, which means that the interpretation of higher-level systems as autonomous (or at least closed) systems appears, to say the least, to be an ad hoc hypothesis to cover reluctant cases.[9] In particular, to the extent that Christensen and Bickhard appeal to the idea of autonomy in a fairly general sense, we assume that Delancey's criticism applies equally to an interpretation of higher-level systems as organised self-maintaining systems, which could be put forward within our own conceptual framework.

A possible reply would consist of arguing that other biological supra-organismal systems do possess the properties required to be considered self-maintaining organisations. Let us briefly explore another possibility, not mentioned by Delancey's analysis: the ecosystem. Compared with species, lineages or populations, there do indeed seem to be better reasons for considering ecosystems higher-level closed systems, relevant for functional ascriptions, especially if one adopts our characterisation in terms of self-maintaining organisations realising closure, rather than the more demanding terms of autonomy. Although there are clear differences (just to mention one: the ecosystem has no physical boundaries), ecosystems share several organisational properties with individual organisms. For instance, the various components (be they individual organisms or groups of organisms) contribute to maintaining a global organisation (the ecosystem itself), which in turn is a general condition for their own continuous existence. Similarly, the various components seem to be mutually dependent, so that the disappearance, death, or anomalous behaviour of one may provoke the collapse of the whole ecosystem.

For these and other reasons, the ecosystem has some features in common with an organism, and in fact it does not seem unreasonable, despite being somewhat uncommon, to use a functional discourse to describe it. So, for instance, we could describe and explain the organisation of an ecosystem by attributing to its various components functions such as the regulation of air, climate, water, water supply, disturbance prevention, soil formation and erosion, nutrient cycling, waste treatment, pollination, biological control of pests and diseases, and so on (De Groot et al. 2002; Nunes et al. 2014). Specifically, cross-generation traits would have the function of regenerating the various components of the ecosystem, which would tend to decay because of their dissipative nature.

In our view, the idea that the ecosystem is, at least, a closed self-maintaining system is an attractive one, and deserves further investigation. Indeed, we discuss

[9] Delancey's remark is fundamentally correct. As a matter of fact, we try to make some preliminary steps towards an account of higher-level closed organisations at the end of Chap. 4, and then of higher-level autonomous systems in Chap. 6.

this issue in more detail in Chap. 4, Sect. 4.5.[10] Yet, the search for higher-level closed organisations would be largely irrelevant for solving the problem of cross-generation functions within the organisational account since, we submit, *the reason why we ascribe functions to cross-generation traits is not related to their contribution to the maintenance of some higher-level system*. Cross-generation functions, we argue, do not require an account of higher-level closed systems in order to be adequately naturalised within an organisational account. Let us then turn to our proposal.

The gist of our account consists of arguing that the apparent difficulty in integrating cross-generation functions into the definition does not stem from an ontological difference between intra- and cross-generation functions but rather from an inadequate understanding of what a closed self-maintaining organisation actually is. Cross-generation functions constitute a "recalcitrant" class of functions only if the boundaries of the self-maintaining organisation are confused with the boundaries of the individual organisms themselves, whereas, in fact, they are conceptually distinguishable. Once this confusion has been cleared up, the ontological problem disappears.

In our account, functional traits are those traits that, by being subject to closure, contribute to the maintenance of an organisation, which in turn exerts some causal influence on the production and maintenance of the traits. The whole system, as discussed in Chap. 1, realises a self-maintaining organisation through closure. The first remark is that a self-maintaining organisation occurs in time, and can be observed only in time. Now, as we have mentioned in Chap. 1, Sect. 1.6, biological organisms undergo various material, structural and morphological changes and modifications over time. If, due to these changes, one were to consider the various temporal instances O_1, O_2, . . .O_n, as different organisations, then functions could not exist. A trait would be produced by a given organisation O_1, and would contribute to maintaining another organisation O_2. No organisation would actually self-maintain, no trait could be subject to closure, and functions could not be ascribed.

The crucial point is that, in the organisational account, these changes are irrelevant with regards to functional ascriptions, because what matters is the *continuity of organisational closure*.

[10]Besides, the claim that certain supra-organismal organisations could harbour functional relations does not undermine our previous proposal of grounding functions in the causal regime of organisms. Since obviously any supra-organismal organisation requires the existence of organisms, it implicitly supposes the (intra)organismic organisation in order to ground the existence of functions. For example, the constraints that ensure the maintenance of an ecosystem are generated by the specific metabolic organisations of different types of species in a given ecosystem. In this sense, the requirement that the constraints be generated within the system – if by the system we understand the supra-organismal organisation– is only satisfied partially (Nunes et al. 2014).

As discussed in Chap. 1, Sect. 1.6, the realisation of closure requires considering a *minimal* temporal interval (say, τ_n), wide enough to include the specific time scales[11] at which all constitutive constraints and their mutual dependencies can be described. As a consequence, the various temporal instances (at time scales τ_1, $\tau_2 \ldots < \tau_n$) of a system can be considered – in spite of any changes that may occur – instances of the *same* encompassing self-maintaining organisation, to the extent that their constitutive organisation realises closure at τ_n. In particular, this implies that the system in which a trait x performs an enabling function at time τ_1 is the *same* system in which, at τ_2, that function of x is dependent, if both τ_1 and τ_2 are included in τ_n (at which closure is realised) .[12]

In other terms, for the purposes of ascribing functions, the continuity of closure (and thus the maintenance of the system) takes precedence as a criterion of individuation over other criteria on the basis of which the various instances of the organisation would possibly *not* be taken as instances of the same system. If there is a causal dependence between two temporal instances of a system, such that their conjunction realises closure, then it could be claimed that, in this respect (and possibly *only* in this respect) the two instances are temporal instances of the same encompassing organisation.

Our central thesis is that self-maintaining organisations, which ground the ascription of cross-generation functions, and specifically reproductive functions, comply with the very same characterisation as those organisations, which ground intra-generation functions. While they may (and actually do) differ in important ways, the two classes of self-maintaining organisations do not differ with respect to the relevant properties that ground functional ascriptions.

Cross-generation functions are subject to closure within those self-maintaining organisations whose extension in time goes beyond the lifespan of individual organisms. For instance, by inseminating the ovum, mammalian semen contributes to the maintenance of the organisation by contributing to the production of a new individual organism to replace the previous one. In turn, the organisation (which consists in the conjunction of both the reproducer and the reproduced system) exerts several constraints under which the semen is produced and maintained. The crucial point is that the organisation of the system constituted by the conjunction of the reproducing and reproduced organisms (in this specific case, a minimal lineage with two elements) has exactly the same status, in terms of self-maintenance, as that of the individual organisms themselves. The fact of considering the organisation

[11] Of course, time scales may greatly vary according to the specific function: the function of the lung is subject to closure in a very short period of time (one cannot stop breathing for more than a few minutes) whereas, for instance, the function of the stomach is subject to closure over a longer period of time (one can stop eating for days).

[12] See Chap. 1, Sect. 1.5, for the definition of dependence among constraints (functions), as well as the distinction between enabling and dependent.

of individual organisms (at τ_n) or their conjunction (at τ_{2n}) as the relevant self-maintaining organisation depends on the explanatory exigencies for functional ascriptions.

Since what matters in the case of organisational self-maintaining systems is the fact that they use their own constitutive organisation to exert a causal influence on the maintenance of (at least part of) their own conditions of existence, then the organisation of the "encompassing system" made up by a reproducer and a reproduced system itself fits the characterisation of a closed self-maintaining organisation. Reproduction, in this sense, simply constitutes one of the functions through which the organisation succeeds in maintaining itself beyond the lifespan of individual organisms. Since the encompassing system made up by the reproducer and the reproduced organism possesses a temporally wider self-maintaining organisation, reproductive traits are subject to organisational closure, and their functions are correctly grounded in the organisational account.

Why do cross-generation functions appear problematic? Intuitively, the point seems to be that reproduction involves a dramatic transition from the reproducer to the reproduced organism, so much so, in fact, that it cannot be maintained that they constitute the same organised system. Given that reproduction may involve phenomena like embryogenesis and development, such causal and phenomenological discontinuities prevent us from considering these systems as temporal instances of the same self-maintaining system. Only individual organisms are genuine self-maintaining organised systems.

In our view, this objection is not compelling, since it is based on an insufficient understanding of what matters for considering that an organisation is self-maintaining. The crucial requirement, as discussed above, is the functional dependence of the temporal instances of an organisation. Two systems which realise closure at a time scale τ_n, may be said to constitute, at a longer time scale τ_{2n}, two temporal instances of an encompassing self-maintaining organisation if it can be shown that the conjunction of the two instances realises itself closure (which includes more functions, in particular cross-generation functions). The relevant question is: is there a causal dependence between the two instances, such that the encompassing organisation can be said to realise closure? Or, to put it another way, is there continuity in the realisation of closure across the successive instances of the self-maintaining organisation? Since the answer to these questions is, in a fundamental sense, affirmative for the case of the relationship between the reproducer and the reproduced system, we claim that the encompassing organisation including them is itself a closed self-maintaining organisation that maintains itself also through reproduction.

As Griesemer has pointed out, the reproduction process does indeed involve the material connection between the reproducer and the reproduced system:

> Reproduction, ... is the multiplication of entities with a material overlap of parts between parents and offspring. Material overlap means that parts of the parents (at some time) become parts of the offspring (at some other time). Thus reproduction is no mere transmission or copying of form– it is a flow of matter (Griesemer 2002: 105).

Rather than a flow of matter as such, the autonomous perspective emphasises the continuity of the functional organisation, which maintains itself over time, also because of reproduction. As it has been argued (Zepik et al. 2001) the occurrence of reproduction may be explained in terms of the time relation between the production and decay of the constitutive components in a far-from-equilibrium organisation. If the rate of replacement of the constitutive components is faster than their decay, the self-maintaining cycles of the system will prompt it to establish reproductive cycles: the system will grow and reproduce; otherwise, it will disintegrate. Only in the very unlikely case of coincidence between the rates of replacement and decay will the self-production cycles of the system realise self-maintenance without reproduction.

The macroscopic transition produced by the reproductive process can then be seen as the way in which the organisation actually manages to self-maintain beyond the temporal boundaries of individual organisms. Just as the various temporal instances of an individual organism are considered, despite changes and modifications, a single self-maintaining organisation to the extent that the organisational properties are causally linked throughout the various instances, so too are the various instances of the inter-generational organisation considered a single self-maintaining organisation due to causal dependence between the instances. The conceptual operation is exactly the same, the difference lies only in the level of temporal "zoom" through which self-maintenance is observed.[13]

This is why development is an essential feature of the self-maintaining organisation of living organisms. Once we see reproduction as a process that causally connects the reproducer and the reproduced organisations, development appears as a necessary step in this continuous process of complex self-maintenance. Indeed, self-maintenance of biological individuals can only be ensured through a continuous unfolding of changes, including reproduction and development. Chapter. 6 will further elaborate on the place of development within the theory of autonomy.

Since the only relevant ground for functional ascriptions is organisational closure, all other criteria of distinction between biological systems may be considered as irrelevant for this specific purpose. This is why reproductive traits can be said to be subject to organisational closure and why, then, we ascribe functions to them.

3.3.2 Malfunctions

A second major implication of the organisational account is the characterisation of malfunctions. It is often claimed (see for instance Neander 1995; McLaughlin 2009; Krohs 2010, 2011; Christensen 2012) that a satisfactory theory of functions

[13]The fact that self-maintenance, in the form of closure, spans beyond the lifetime of individual organisms is an important aspect related to the historical dimension of autonomy. See Chap. 5 for a detailed discussion.

should be able to ground both functions and malfunctions,[14] since a function can be performed well, or defectively, or even not at all. Yet, in spite of the fact that the concept of malfunction is widely used both in everyday language and in scientific disciplines such as physiology or medicine,[15] the theoretical grounding of malfunctions has received little attention in the philosophical debate, which has mainly focused on the concept of function.

What is the gist of a philosophical account of malfunction? Claiming that a trait can function "well" or "poorly" implies a reference to a norm, which may or may not be fulfilled. Malfunctions, then, have a normative dimension, just as functions do. But, and here comes the central philosophical issue, the norms grounding functions and malfunctions are not the same, and an independent justification must be provided for each.

The closure of biological organisation provides the relevant grounding in which the concept of function can be adequately naturalised. In particular, it generates the norms that the traits subject to closure must fulfil in order to be functional: as we claimed, the organisational approach identifies these norms as the conditions under which the whole organisation (or, more precisely, each specific regime of organised self-maintenance, see Chap. 1, Sect. 1.8.1), and consequently each of its constituents, can exist. Thus, functional traits are all those whose causal effects contribute to the maintenance of the whole organisation.

Now, of the whole set of traits that fulfil the norms of functionality, some do so well and others poorly. Yet the norms generated by closure are blind with respect to the distinction between these two types of effects, because both of them contribute to the maintenance of the organisation (albeit in some cases poorly), and both are therefore *functional*. Hence, the distinction between (well-)functions and malfunctions requires an additional set of norms, on the basis of which it might be possible to discriminate between different ways of contributing to the maintenance of a closed organisation.

One important implication of this line of thought is that functions and malfunctions are by no means alternative kinds of entities; rather, malfunctions are a subset of functions that, while fulfilling the norms generated by closure, fail to comply with the norms of *well*-functions. This enables, among other things, a straightforward conceptual distinction to be made between *malfunctions* and *nonfunctions* (often confused both in ordinary use and specialised literature): while the former are indeed a class of functions, the latter do not. Nonfunctions refer to the effects of traits which do not comply with the norms generated by closure, and do not therefore contribute at all to maintaining the organisation. A kidney that does not filter blood, for instance, is nonfunctional rather than malfunctional. The distinction between

[14]We prefer to use the term *malfunction*, because *dysfunction* is usually used to refer both to malfunctional and nonfunctional behaviours.

[15]The concept of malfunction has often been used to justify the conceptual distinction between health and disease: some of the most influential groundings of the concept of disease have specifically interpreted diseases as malfunctions (Boorse 1977, 2002; Schramme 2007).

nonfunctions and malfunctions also serves to highlight the fact that malfunctionality is a *matter of degree* (Krohs 2010: 342). While functions are all-or-nothing concepts (a trait is either functional or nonfunctional), malfunctions admit degrees and a given trait can contribute more or less well (or poorly) to the maintenance of the organisation.

How does the organisational account deal with the concept of malfunction? Although no fully-fledged organisational definition of malfunction has been proposed so far, several authors whose approach could be considered within, or at least close to, the organisational account have pointed to a link between malfunction and adaptivity. For instance, Edin (2008) refers to malfunctions in terms of deviations from the "optimal self-maintenance" of a living system:

> Organisms are typically endowed with multiple, extensive and complex feedback systems, many of which have a set point that, when considered from the standpoint of the maintenance of the organism, is close to optimal. For this reason, physiologists talk about events or circumstances that cause the variable magnitude of such a system to deviate from the set point as disturbances or challenges. (Edin 2008: 206)

Christensen and Bickhard (2002) also consider malfunctionality to be related to the adaptive properties of organisms:

> There are a number of reasons why understanding the relative significance of dysfunction is an important adaptive issue. It is important to understand the wider systemic implications of failure in order to understand whether and how the system can compensate. It is also important to know how the system can recognise failure as part of its compensatory abilities. These are surely important issues for understanding functional organisation (Christensen and Bickhard 2002: 18).

In what follows, we will elaborate on this very idea, by relying on our characterisation of regulation exposed in Chap. 1, Sect. 1.8.2 above. As we discussed, biological organisms have to modulate their organisation to cope with the changes that they continuously undergo, be they internally or externally generated (for instance, in this second case, by a variation of the environmental conditions). Regulation is a specific form of modulation, such that a functional subsystem (a dedicated mechanism) of the organisation induces the establishment of a different and more adequate constitutive regime of self-maintenance, among a set of possible ones. Regulatory functions are, then, second-order functions (subject to second-order closure and norms) that modulate the constitutive set of functional traits and their interrelations.

In a nutshell, our account of malfunction is the following. The whole dynamic repertoire of the constitutive organisation on which regulation is exerted is limited by its physical and material structure, which implies, in particular, that each trait can only operate within a given potential range of activity. For each specific regime of self-maintenance that the system may adopt, a specific *admissible* range of activity, included in the potential one, can be determined.

If, because of some structural defect, a particular trait (1) does not modulate its activity in spite of the triggering of a regulatory mechanism and (2) as consequence, it is unable to operate within the admissible range determined by some of the

regimes of self-maintenance among which regulation governs the shifts, then the trait malfunctions in organisational terms. Let us explain this idea in more detail.

Within each specific realisation of a closed organisation (i.e. each regime of self-maintenance), functional traits *presuppose*[16] each other, which means that the whole set of mutual interactions among them determines the range of admissible functional effects, defined as a subset of all potential effects that the trait may possibly produce, given its own structure. For example, a human heart can pump blood within a certain range of potential frequencies, among which a range of admissible frequencies are determined by each ongoing realisation of the organisation. Similar ranges apply of course to the lungs, kidneys . . . and to all other organs and functional traits.

Suppose that, in some circumstances, a regulatory mechanism is triggered to shift an organism form a given regime of self-maintenance to a different one. For instance, the autonomic nervous system (the regulatory subsystem, in this case), in a situation of danger, can send signals to move from a regime "at rest" to another one "under stress" in which the organism runs. Suppose also that, for some structural reason, one functional part of the organism does not modulate its activity and, as a consequence, it is unable to match the functional presuppositions of the regime induced by the regulatory functions. For instance, the coronary artery might not be able to increase its diameter sufficiently to match the higher rate of blood flow pumped by the heart: as a consequence, its range of activity is not in accordance with the functional presuppositions of the other functional traits and organs in this specific circumstances. Regulatory functions might therefore fail in modulating the defective trait's activity, so to match the new functional presuppositions (fall within the admissible ranges). For that specific trait, in a word, regulation had no effect.

In these specific situations, in which an unresponsive trait does not modulate its activity as required by the intervention of regulatory functions and therefore prevents adaptive regulation to shift to a different first-order organisational regime, so that the whole system can only remain in a specific organisational regime in which the trait match the functional presuppositions, that trait is malfunctional.

[16]The idea of functional presupposition was originally put forward by Bickhard (see for instance Bickhard 2000; Christensen and Bickhard 2002). We can understand the idea through the following example: "As everybody knows, the function of the heart is to pump blood, or more accurately to pump blood as a contribution to an ensemble of activities that result in blood circulation. The function that this serves, however, is to provide fluid transport for delivering nutrients to cells and removing metabolic end products. In this respect heart activity and cellular metabolism are interdependent processes. Without heart activity, fluid transport stops, and with it cellular metabolism. And if cellular metabolism ceases then heart activity also ceases, and subsequently fluid transport. In addition to heartbeat, cellular activity also produces other motor action that contributes to interaction processes such as breathing, food acquisition, eating and excreting. In turn these processes provide the resources required for cellular metabolism and expel waste products, thus contributing to the cellular processes that subserve them., . . . , These patterns of process interdependence in biological systems are (. . .) what determine the nature of organisms as viable (cohesive) systems" (Christensen and Bickhard 2002: 16–17).

The organisational account, therefore, interprets malfunction as any functional activity with respect to which there has been a *failure*[17] of regulation. In other terms, malfunctions are a subset of functions that fit first-order norms (of the first-order ongoing organisation in which they match functional presuppositions), but *not* second-order ones (since they do not obey to second-order regulatory functions, and prevent the shift to another first-order organisation). In this respect, the degree of malfunction of a trait could be assessed in terms of the set of first-order organisations of which it prevents the realisation. The degree of malfunction is, therefore, inversely proportional to the degree of adaptivity of the organism (see also chap. 4).

Malfunction occurs when the autonomous system fails in regulating the activity of a trait, including the specific case in which regulation aims at compensating for a "defective" activity of the trait in a given organisational regime. This is a crucial implication, because were a given first-order regime of self-maintenance capable of compensating for the apparently malfunctional activity of a trait, it would be impossible, from an autonomous perspective, to contend that the trait is malfunctioning. In such a case, its contribution to the system would, in principle, be indistinguishable from another contribution within the presupposed range of functioning. Indeed, if there were no higher-order regime with respect to which the behaviour of the trait is unfit, we would simply be faced, from the autonomous perspective, with a *different* organism (i.e. an organism that would function in a different, equally viable, way), and not with a malfunctioning one. For example, certain organs of the mole (i.e. those involved in sensory-motor activities) presuppose that its eyes provide very limited visual capacity, and that is why this animal is perfectly viable despite being almost blind (in fact some moles, like the star-nosed mole, display a remarkable foraging ability thanks to the star-shaped set of appendages that ring their nose). If there were no (failed) regulatory intervention, there would not be organisational criteria to interpret the behaviour of the trait as malfunctional. On the other hand, if regulation were able to compensate for the operations of a defective trait – by shifting to a regime of self-maintenance in which the trait would match the functional presuppositions – there would not be organisational reasons either to contend that the trait is malfunctioning. In such a case, its contribution to the system would match both first-order and second-order norms, and therefore it would be theoretically indistinguishable from any other contribution within the presupposed range of functioning.

A trait that malfunctions is, first of all, a functional trait, in the sense that it contributes to the maintenance of a self-maintaining organisation. What happens is that this contribution is not made *according to certain second-order norm* and that is why we say that it is a "bad" or "poor" contribution. Malfunctional traits show

[17]In technical terms, the very possibility to detect a failure of regulation supposes that the admissible ranges of the ongoing organisation and the alternative ones (to which regulation should move the system) are, at least *partly, non* overlapping. This means that the regulatory intervention must result in an observable *change* of the defective trait's activity.

a degree of malfunction rather than an "all-or-nothing", "function-no function" dichotomy. And the effects of a functional trait are deemed "good" or "bad" according to the norm that lies in the action of a regulatory subsystem (Saborido et al. 2014; Saborido and Moreno 2015).

It could be argued (Artiga 2011) that, ultimately, the norms to which any organism is subject have been set through evolution by natural selection, which shapes the species it belongs to. In particular, each given set of second-order norms could be defined at the populational level, because it has been selected in relation to the conditions of a stable existence (over a long period of time, covering many generations) in a given niche. Thus, it is because of its contribution to the self-maintenance of a class of organisms that this particular normative mechanism exists. And this happens too with the shaping of the structure and organisation of functional traits.

Yet, there are two aspects in which the current organisation of individual organisms matters. First, though the mechanism of adaptive regulation of a given organism is set through an historical-collective selective process – because only those forms of modulation that ensure viable organisations (in specific environments) can be selected – ultimately the regulatory mechanism would not exist if it did not make a contribution to the self-maintenance of each individual system in which it operates. The second, and even more important aspect is that although the *origin* of the norm according to which something is deemed malfunctional is ultimately an evolutionary matter, this does not mean that we cannot define, in the current organisation of each individual organism, whether or not a given trait is well-functioning or not. As Christensen has recently pointed out:

> The aetiologist may point out that, living systems have infrastructure for self-perpetuation largely as a result of an evolutionary history., … nevertheless, … the key perspective for normative evaluation of function is the current system rather than past selection. Regulation does not succeed by making parts function as they did in the past, it succeeds by making the system work well in present conditions. (Christensen 2012: 107)

In this respect, we consider that the organisational account of malfunctions can include evolutionary considerations without falling into epiphenomenalism, i.e., an understanding of functional attributions appealing to something other than the traits' current performance (see Sect. 3.1.2 above).

To conclude, we wish to emphasise that the organisational account to malfunctions does not rely on the subjective criteria of an external observer. What matters is what happens operationally within the system itself, and whether or not there is failure in adaptive regulation. Moreover, the normativity to which obeys the adaptive subsystem of a given organism is not defined with respect to a *type* of organisms but, rather, in relation to the current organisation of this organism and, more precisely, to the second-order closure to which regulatory functional traits are subject. This way of understanding the concept of malfunction is quite different from the most predominant notion of malfunction used in the philosophy of medicine, namely the bio-statistical conception, expounded by authors such as Boorse (1977, 1997), which claims that a malfunction is a deviation from

"normal" (i.e., the statistically more common) functional behaviour. The biostatistical conception has been fiercely criticised (Amundson 2000), and numerous problems and counterexamples have been put forward, so that its influence within the philosophy of medicine is declining (Khushf 2007).

The implications of an organisational account of malfunction are still to be explored and critically assessed. Yet, this account might open new directions in the search for a theoretical grounding of the notion of physiological disease, within an alternative naturalistic perspective.