To be published in *RIEV*

Emergence and Autonomous Agency

Xabier E. Barandiaran

Dept. of Philosophy, Univ. of the Basque Country (UPV/EHU)
https://xabier.barandiaran.net

Postprint revised version of the original published in *Revista Internacional de los Estudios Vascos, 70 (2)* in January 2026: doi.org/10.61879/riev702zkia202511

Short Abstract: Emergentism seeks to reconcile material reality with the complexity of biological and psychological phenomena, rejecting both dualism and reductionism while preserving scientific unity across levels. This paper traces the emergence of agency, from the self-maintaining organization of protocells to human autonomy, arguing that far-from-equilibrium autonomous systems ground a distinctive form of causality (where actions replace events, norms replace laws, and goals replace equilibria) that characterizes agents.

Abstract: The primary aim of emergentism is to find a balance between the undeniable material foundation of our reality and the complex characteristics displayed by biological and psychological phenomena. Emergentism disputes dualism by negating the need for separate substances or distinct domains of reality. Moreover, it contradicts reductionism by denying that all properties can be derived from lower or more fundamental levels of explanation. And against a proliferating pluralism, it preserves the coherence of a unified science by acknowledging systematic relations across levels that are accessible to scientific scrutiny. One of the greatest challenges to a scientific understanding of the world is *agency* as a manifestation of autonomous organized systems. Explaining agency (from the origins of life to the appearance of state intelligent agencies) involves explanatory resources unlike those of classical physicalist vocabulary: instead of *events* we speak of *actions*, instead of *laws* we talk of *norms*, instead of *equilibria* we talk of *goals*, instead of *forces* we talk of *intentions*. The most basic type of agent can be traced back to the recurrent interactions that a protocell establishes with its environment in order to stay alive. This minimal organization, that of *basic autonomous systems*, is used to illustrate the central features of organizational emergence and the distinctive form of causality that we attribute to agents. This way, in a world that is so often poorly understood as a mere collection of events, we can make room for a scientific understanding of the emergence of goal-directedness, reasons, and, ultimately, human agency.

Keywords: Agency, Emergence, Agent-causation, Organization, Basic Autonomy, Sensorimotor Autonomy, Origins of Life, Intentional Action.

Table of Contents

1. Introduction.....2
2. A brief history of emergence.....4
2.1. Dual/pluralisms.....4
2.2. Reductionisms.....6
2.3. Emergentism.....7
2.4. Structural and organizational emergence.....11
3. The emergence of basic autonomous agency.....13
3.1. What agency requires.....13
3.2. Conservative and dissipative structures.....13
3.3. Dissipative order as self-organization.....15
3.4. Towards organized complexity.....18
3.5. Basic autonomous agents.....21
4. Agent causation and the scientific understanding of the world.....22
References.....25
Acknowledgments.....29
Conflict of interest statement.....30

1. Introduction

We tend to subtract ourselves from the order of objects and artifacts. The study of objects is carried out by physicists and chemists, the study of humans and their assemblages by psychologists and sociologists. The manipulation of material objects falls into the realm of engineering, that of humans into ... politics (?!). The difference between disciplines is not merely a contingent frozen accident of the history of Western academia and its departmental divisions. Rather it responds to a fundamental split between two kinds of causal regimes. For instance, it might take but a whispered word carrying scarcely 10^{-9} joules of acoustic energy to set millions of human beings (and countless artefactual systems) into sustained motion toward another country (e.g. for an invasion, vacation, or exile). A similarly minuscule cue, such as the slight lengthening of daylight at springtime or a faint geomagnetic signal, can set billions of migratory birds into coordinated flight across continents. By contrast, it would take on the order of 10^{9}-10^{15} joules to move the same bodies by mere external causal physical forces (push, gravitation, magnetism, etc.). This difference of over 20 orders of magnitude demands some explanation. Whereas the latter case involves movement by *external force* (push, gravitation, magnetics, etc.) the second involves *reasons*[1] (and forces generated or stirred from within according to them).

1 When we invoke the "force" of reason we do so metaphorically. I have used the term "reason" here on its most liberal or generic form, less heavily loaded terms can be information, purpose, goals, or norms.

Not only the whisper or the magnetic signal need to be involved into the explanation, as triggers of the observed behavior. Why do birds migrate from A to B? Because it is warmer in B. Why do tourists move to B? Because it is warmer in B. The causes that bring about the agentive movement are not un-physical. And yet, they don't seem to be plausibly captured by standard physical causes alone.

To come out of this conundrum there have been three standard responses in the history of philosophical and scientific understanding of the world[2]: a) to split reality into different realms (dualisms or pluralisms), b) to deny the proper existence of agency and related concepts like teleology, intentionality, goal-directedness, etc. (reductionism), or, c) the acknowledgment that, despite reality been grounded on the materiality that physicists study, there are properties that emerge from it (agency among them) that deserve a proper level of study and the assignment of a proper level of existence. What is at stake here is more than a sophisticated theoretical complication. What we are dealing with here is the very possibility of attributing causal powers to specific individuals or ensembles, that could ultimately lead to explaining self-determination, free will and responsibility (although a proper treatment of these categories requires a trip much longer than what this journey permits).

2 By a scientific understanding of the world we mean that understanding that is sustained by the results of the scientific method(s), as opposed to understandings based on pure speculation, religious or spiritualist beliefs or other cultural or social foundations of our conceptual or otherwise cognitively structured dealing-with and symbolization-of the world.

2. A brief history of emergence

The concept of emergence (and its various manifestations) has accompanied western thought almost since its origins. Its role has been to mediate and bridge between reductive understanding of nature and our place in it (collapsing the observed world into a preferred domain: generally matter or physics), and dualist or supernatural understandings (invoking supernatural domains of existence to account for observed phenomena: the origins of humanity, the nature of consciousness or the soul). The Aristotelian notion of *form* was an early (and still vividly celebrated) attempt to reconcile both a materially grounded (i.e. not transcendental) reality that, in turn, can display finality or teleology, and it is often considered a forerunner of emergentism (O'Connor, 2020; for a contemporary hylomorphic approach to emergence see Jaworski, 2016). Ever since, philosophers and theoreticians alike (in physics, chemistry, biology and psychology) have struggled to make sense of the way in which nature (including ourselves within it) appears stratified in different levels or domains, each of them displaying their own properties and regularities, yet dependent on those of its constituents. Such differences have manifested in a split between disciplines and thus, also, justifying different methods of study and the difference between scientific disciplines.

These levels of manifestation of reality invite competing explanatory strategies. *Dualisms* (or pluralisms) posit two (or more) fundamentally distinct realms. *Reductionisms* seek to collapse diversity into a single explanatory base. The concept of *emergence* arises as a mediating framework that preserves the integrity of different levels without resorting to either absolute separation (dual/pluralism) or total subsumption (reductionism). Since emergentism owes so much to this tension between dual/pluralism and reductionism it is worth spending some time detailing what these two poles demand. We shall briefly navigate them attending to one main author. We have chosen René Descartes (1596-1650) for dualism and Rudolf Carnap (1891-1970) for reductionism.

2.1. Dual/pluralisms

René Descartes' epistemological path to the *Cogito* begins with a *methodological doubt*: any belief about the world can be false. The senses deceive, dreams confuse appearance and reality, and even a malicious demon (a technological corporation?) could be tricking us on a constant basis. What remains certain is the existence of the thinking self, a *cogito*. But this one is continuously confronted with its senses, that also deliver clear and distinct ideas. It could all be a dream, but God couldn't be so wicked. There has to be something out there that the senses actually sense. From here, Descartes distinguished between two fundamental substances, giving rise to Cartesian dualism. Each substance posits its own defining properties. *Res extensa*, on the one hand, names the external reality, captured by the senses and grounding our experienced reality of the world. Entities in this domain are

spatially extended, localized, divisible, determinate, passive, mortal, and perceivable by the senses. *Res cogitans*, on the other, is the realm of the mind, rationality and ideas. Entities in this domain are non-extended, non-localizable, indivisible, free, active, immortal, and not perceivable by the senses.

Descartes' doubtful sin[3], was to renew and sharpen the conceptual distinction that is not only pervasive in the inherited history of Western thought, already present in Western philosophy (from Pythagoras to the Scholastics), but also difficult to resist by looking at the gigantic gap that separates the dirt of material properties from the purity of the mathematical beauty that is accessible to consciousness. Yet, this ontological dualism, made clear and explicit, generated its own philosophical problems: How do mind and body relate? If the physical realm is deterministic, how can freedom and physically-effective human action be reconciled? How do I know my conscious perceptions correspond to external reality? How can one prove that other conscious beings exist?

Contemporary dualists continue (in one way or another) to face these challenges while enjoying the safety and tranquility of the conceptual divide between mind and matter. Karl Popper and John Eccles, for instance, defended an interactionist view between "World 1" (physical) and "World 2" (mental) (Popper & Eccles, 1977). Many Christian thinkers today defend substance dualism, often renewing Cartesian arguments in a contemporary philosophical context to support it (Moreland, 2011; Swinburne, 2019). And without any specific religious commitment, David Chalmers (1997), one of the most prominent philosophers of science alive, defends a strong form of property dualism (what he calls "naturalistic dualism") holding that conscious experience is a fundamental feature of reality; one that can not be treated as reducible to, or emergent from, the physical.

Paradoxically, many dualisms operate also as reductionisms. This is particularly true of Descartes, since it demands that the diversity of regularities and properties observed in the universe be subsumed or reduced to either one of the two substances or domains that split reality. Since only the most rational and conscious aspects of the human mind were left to fall under *res-cogitans*, the rest of reality, including (importantly) the living, falls under *res-extensa* and the mechanical view of nature it implies. But we shall come back to this latter. It is now time to illustrate one of the most influential and even today widely (although often implicitly) spread form of reductionism, physicalism, that targeted directly the dualism between mind and bodies and that finds Rudolf Carnap among its founders and one of its greatest philosophical exponents.

3 We mean "doubtful" here is a double sense: because it is doubt as a method that brought him there, but also because it is hardly a sin, but a virtuous philosophical en-framing of an old and everlasting problem.

2.2. Reductionisms

Alongside dualism, philosophy has long cultivated reductionist explanations of nature. From the atomism of Democritus, to Empedocles' pluralist four elements, philosophers and scientists sought to explain the multiplicity of the world by reference to fundamental components.

In modernity, reductionism gained momentum. Its philosophical roots can be traced to early empiricists like Francis Bacon, who championed the nascent scientific method of reducing complex problems to observable, measurable facts, laying the groundwork for *positivism*. This spirit was further developed by René Descartes' analytical and decompositional method, which reduced the scientifically accessible material world to mechanical causes. This vision found its most compelling physical expression in Newtonian physics, which grounded explanations in particles, forces and motions. La Mettrie's materialist psychology and Hume's associationism extended reductionist strategies to minds, epistemology and morality. Laplace's determinism sealed the causal closure of the physical (with no room for the will) and Comte declared a definitive victory of positivism over any kind of metaphysics. Marx's scientific materialism reduced political structures to economic relations of production subsumed under historical determinism. More recently, logical positivism and behaviorism reduced psychology to observable stimulus-response pairing (Skinner, 1953; Watson, 1913). Reductionism also advanced (and advances) within the natural sciences: chemistry was claimed to be reduced to quantum mechanics (Dirac, 1929), thermodynamics to statistical mechanics (Boltzmann, 1964), biology to molecular biology (Crick, 1966) and genetics (Dawkins, 1976), and mind (even philosophy itself!) to neuroscience (Ramachandran, 2012).

Rudolf Carnap was a member of the Vienna Circle and what came to be known as *logical positivism* (or logical empiricism), a school of thought that, through the success of natural sciences and the newly develop solid foundation of logic (Russell, 1916; Wittgenstein, 1922) stated that all human knowledge should be reducible to logical statements grounded on observation. To say it with Carnap: "science is a unity in that (1) all empirical statements can be rendered in a single language, (2) all states of the world are of one type, and (3) they are known by the same method" (Carnap, 1934, p. 32). For Carnap, unity stems from three converging dimensions. Semantically and epistemologically, different theories, models, and even languages should be inter-translatable or reducible into a common formal language. Ontologically, various entities or processes are not fundamentally heterogeneous but reducible to a single kind; thus preserving a uniform ontology across disciplines. Methodologically, scientific knowledge should be acquired via a single privileged method (be it observation, manipulation, or experimentation) forming a coherent interface with the world. Mental or psychological statements are no exception:

> So-called psychological sentences—whether they are concrete sentences about other minds, or about some past condition of one's own mind, or about the present condition of one's own mind, or, finally,

> general sentences—are always translatable into physical language. Specifically, every psychological sentence refers to physical occurrences in the body of the person (or persons) in question. On these grounds, (...) *psychology is a branch of physics.* (Carnap, 1959).

If psychology is a branch of physics so is sociology and biology. Do nations, persons, or organisms exist? Well, for a pure reductionist there are two answers to this question[4]. Either these entities are simply higher level labels for long physicalist descriptions or ensembles of constituent aggregates, and add nothing to them except for a handy shortcut (like "a dozen" is no more than 12 = 1+1+1+1+1+1+1+1+1+1+1+1). Or they simply don't exist; like *ether* or *phlogiston* never existed; despite their systematic presence in early scientific theories. Talks of desires as motivating human behavior, or free will, are simply wrong; and will be substituted in the future by a more detailed neurobiological understanding of human behavior (Churchland, 1981). Psychological concepts of this type are no more different in scientific explanations ghosts moving objects, animal spirits taking over the body of a person committing a crime, or a curse explaining my failure to finish this paper. The same goes for organisms in biology, a concept that needs to be eliminated in favor of genes, which are nothing but differentially reproducible patterns of molecules (Dawkins, 1976). This form of strong reductionism takes the name of *eliminativism*, for it defends to eliminate those causal principles, substances, entities or properties from the catalog of what exists.

2.3. Emergentism

As we mentioned earlier, emergentism is an attempt to reach some sort of equilibrium between the undeniably material grounding of our reality and the challenging properties displayed by biological and psychological phenomena. Against dualism, it denies the necessity to separate substances or domains of reality. Against reductionism, it denies that all properties can be deduced from lower or more fundamental levels of descriptions. And against a proliferating pluralism, it preserves the coherence of a unified science by acknowledging systematic relations across levels that are accessible to scientific scrutiny. Emergence is both a historical response to unresolved debates between dualism and reductionism, and a contemporary methodological principle grounded on the sciences of complexity (as we shall see). It makes it possible to see novelty not as a threat to explanation, but as an irreducible feature of the multilayered reality in which we live.

---

4 To be fair to Carnap, my use of his reductionism here is rather instrumental, profiting from the clarity and boldness of his claims. However, Carnap's own position is better read as a semantic-methodological physicalism (unity via translatability and intersubjective testability) rather than as an eliminativist metaphysics. In this sense it leaves more room for higher-level explanatory autonomy than many other physicalists, particularly eliminativists.

In the context of the rise of Cartesian mechanicism, Newtonian mechanics, chemistry, and the success of automata as models and molders of the world (clocks, instruments, and industrial machinery) the debate on *vitalism* emerged. Vitalists argued that living beings could not be reduced to mere physical and chemical processes, but required a special *force* to account for their properties. Throughout the 17th up to the early 20th century a systematic resistance to simplify living phenomena to the available physicalist or mechanistic methods grew strong. Georg Ernst Stahl's (1659-1734) "anima", Caspar Friedrich Wolff's (1734-1794) "vis essentialis", Baron Carl von Reichenbach's (1788-1869) "odic force", or Henri Bergson's (1859-1941) "elan vital", they all postulate a fundamentally new or irreducible force, energy source or organizing principle to explain living phenomenology. In contrast, *emergentism* arose as an alternative. It sought to preserve the materiality and complexity of life without falling into the reductionism of materialism, while also avoiding the invocation of a non-physical "vital principle", or recurring to the growing spiritualism and idealism in philosophy.

Before the term emergentism was coined, in *A System of Logic* (1843), John Stuart Mill distinguished between two types of laws. *Homopathic* laws follow the "composition of causes," meaning that the joint effect of causes is simply the sum of their separate effects; for instance, the combination of forces in mechanics. Here, the whole is just the sum of its parts. By contrast, *heteropathic* laws do not follow this principle. In these cases, the whole exhibits properties that are not predictable from the mere aggregation of its parts. The much celebrated motto was born: "The whole is more than the sum of its parts". It was Mill's student, George Henry Lewes, who introduced the term *emergent* to describe precisely these heteropathic effects.

During the 1920s, there was a surge of publications on the concept of emergence, especially in philosophy of mind and philosophy of science on what would be known as British Emergentism. One of the most influential works was Charlie Dunbar Broad's *The Mind and Its Place in Nature* (1925). Broad addressed not only the problem of the reduction of properties (mental and vital) but also the question of whether sciences could be reduced to each other. We owe him one of the earliest (still useful) precise definitions of emergence:

> Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents A, B, and C in a relation R to each other; that all wholes composed of constituents of the same kind as A, B, and C in relations of the same kind as R have certain characteristic properties; that A, B, and C are capable of occurring in other kinds of complex where the relation is not of the same kind as R; and that the characteristic properties of the whole R(A, B, C) cannot, even in theory, be deduced from the most complete knowledge of the properties of A, B, and C in isolation or in other wholes which are not of the form R(A, B, C). The mechanistic [reductionist] theory rejects the last clause of this assertion. (Broad, 1925, p. 61)

Unfortunately, by the 1930s, emergentism began to lose influence. Scientific progress on some reductionist programs (we explained earlier) contributed to this decline. The development of quantum mechanics, molecular biology, and the Modern Synthesis in evolutionary biology provided powerful reductionist explanations that still permeate the natural and social sciences. Philosophical factors also played a role: the rise of logical positivism, particularly in psychology through behaviorism (see the previous subsection on reductionism), rendered emergentist approaches out of focus. If the reductionist program was succeeding there was no need of emergentism.

However, emergentism experienced a revival in the 1970s and 1980s. Partly because of the failure and limitations of the reductionist programs, notably in psychology (Chomsky, 1959; Tolman, 1967), but also because of the success of the emergentist research programs, as we shall see. An influential physicist (yet not traditionally a "physicalist") turning point came with Philip W. Anderson's article *"More Is Different"* (1972). Anderson departs from the standard distinction between two kinds of science: *intensive science*, which seeks *fundamental* principles (mainly particle physics), and *extensive science*, which applies those principles to explain higher-level phenomena. He argued against some forms of reductionism and in favor of *constructionism*:

> The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe. (...) At each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other. (Anderson, 1972, p. 393)

Thus, not even physics was safe to discover properties that could not be deduced from its most fundamental principles. This paved the way to conceive of biological and psychological domains as constituted by emergent fundamental principles not deducible from the scientific understanding of their most basic constituent.

In science, new fields such as Artificial Life and Complexity Sciences (including nonlinear dynamics, self-organization, and cellular automata) reopened the debate on emergent properties (Bedau, 1997; Kauffman, 1993). In philosophy, new discussions of the mind–body problem arose facing the problems of behaviorism (Chomsky, 1959; Tolman, 1967). With the help of the computer metaphor, the mind came to be understood as supervenient on matter yet distinct from it, very much like software is understood, programmed and explained using different concepts and procedures to those used on the hardware in which it is implemented (Fodor, 1968; Putnam, 1965). But more radical emergentist views also grew out of complex systems approaches to the mind and the cognitive sciences, understanding mindful properties as emergent from the non-linear interaction between brain, body and environmental processes (Clark, 1998); irreducible not only to their material components but also the abstract computational representation (Chemero, 2013; Di Paolo et al., 2017; Varela et al., 1991).

Today, the concept of emergence plays a central role in diverse areas. In physics, emergence is studied in condensed matter systems (Laughlin & Pines, 2000), phase transitions (Cheung et al., 2018), and quantum phenomena (Lewis, 2017). In biology, it informs systems biology (Boogerd, 2007; Kitano, 2002), evolutionary developmental biology and theories of self-organization in life processes (Kauffman, 2000; Moreno & Mossio, 2015; Newman, 2018; Sole & Goodwin, 2002). In cognitive science and artificial intelligence, debates on the distributed and emergent nature of behavior from neural networks (Rumelhart et al., 1987; Smolensky, 1988; Wei et al., 2022), the nature of consciousness (Edelman & Tononi, 2001; Tononi et al., 2016), embodied cognition (Barandiaran, 2017; Chemero, 2013; Clark, 1998; Di Paolo et al., 2017), all heavily rely on and contribute to emergentist thinking.

In philosophy, the theory of emergence has developed in great detail (Bedau & Humphreys, 2008; Kim, 1999; for a systematic recent update see O'Connor, 2020). It is commonly accepted that the main challenge of emergentism is to balance two central thesis: the *dependence* of the emergent phenomena (domains, properties, substances, laws, etc.) on the lower level constituents (ultimately understood as material or physical particles, forces, etc.) and, at the same time, the *autonomy* of the emergent, its capacity to display higher order regularities or laws that cannot be deduced, predicted and/or understood from (the best knowledge of) the properties of the constituents on which the emergent depends[5].

Philosophically, debates now distinguish between weak and strong emergence. *Weak emergence* posits, at least, that emergent phenomena cannot be deduced (analytically) from the functioning of its constituents. The only scientific access to the relationship between constituents and emergents is through fine grained computer simulations that *reproduce* a significant part of the constituent dynamics (together with additional constraints). These complex models are needed to reconstruct the higher level emergent phenomena from its base, but, at the same time, the emergent phenomena might be subject to higher level descriptions that are more robust, accurate, informative or efficient than the lower level simulations (McGregor & Fernando, 2005; Seth, 2011). Moreover, it is often the case that, in practice (and sometimes in principle, like cases involving deterministic chaos), not even such simulations can render fully accurate, predictable or reliable information of the behavior of the emergent.

Proponents of *strong emergence* add that the emergent level displays *new* causal powers that influence the lower level dynamics (Bishop, 2008; Bitbol, 2012; Campbell, 1974). This process is called *downward causation*, and it is the object of strong debate for it pushes the problem of emergence to its limits. If an emergent has the power to cause changes on its constituents beyond the causal interactions that take place between constituents, in a manner that adds to the causal properties of its constituents, then it follows

5 Note that the term "autonomy" here is used in a different (yet not totally unrelated) sense to that used when we talk about biological autonomy or autonomous agency.

that there is no causal closure at the fundamental level. In other words, explanations at the level of the base of emergence (of the more fundamental level of explanation below that of the emergent phenomena) are incomplete or causally open or the emergent phenomena are overdetermined both by constituent causes and by the emergent ones (Kim, 1998, 1999).

2.4. Structural and organizational emergence

Emergent phenomena can be categorized along many dimensions: diachronic vs. synchronic emergence, epistemic vs. ontological emergence, weak vs. strong emergence, and others. For the purpose of this paper, however, I will suggest a distinction between *structural* and *organizational* emergence.

Structural emergence occurs when, given a set of boundary conditions and a particular type of constituents, a system produces higher-order patterns and properties that can be subsumed under a law or nomological structure. This type of emergence is typically studied in physics and chemistry. Examples include the emergence of solubility, superconductivity, laser coherence, crystallization, etc. Structural emergence often involves phase transitions, where new properties appear after the transition, but the transition itself is determined by constraints, constituents, and the system's global state (so that local states, specific initial conditions and trajectories, or historicity, are often negligible).

Structural emergence is not limited to physics. It also appears in special sciences such as biology and sociology. Whenever an emergent pattern can be expressed as a law-like or nomological regularity, it is determined by the interplay of boundary conditions and constituent properties. Examples include morphogenesis in biological tissues (Turing, 1952), predator–prey oscillations in ecological systems (Lotka, 1920), or collective behaviors such as human crowd dynamics and bird flocking (Reynolds, 1987). Other cases include traffic jams (Nagel & Paczuski, 1995), neural synchronization in the brain (Buzsaki, 2006; Thompson & Varela, 2001), and financial market bubbles (Sornette & Cauwels, 2015), where macro-patterns arise lawfully from many interacting units under specific and identifiable conditions.

Organizational emergence, by contrast, is marked by the appearance of *functional* differentiation and integration contributing to the self-maintenance of a system. An organized system is one in which a diversity of components or processes can be distinguished as differentially contributing to sustaining the system in a coordinated (integrated) manner . The hallmark of organizational emergence is not merely a new property that can be captured by a law, but rather the constitution of an autonomous, self-sustaining whole. Paradigmatic examples are living organisms, ecosystems, and human organizations, where subsystems play specialized roles that maintain the viability of the overall system. The organizational concept of *function* (Christensen et al., 2002; Mossio et al., 2009; for systematic contemporary revision see Barandiaran, 2025) plays here a crucial explanatory and differentiating role between the type of emergence studied by physics and that of living,

psychological and social sciences. And it is here where agency emerges (Ruiz-Mirazo & Moreno, 1998).

The distinction between structural and organizational emergence, however, is not sharp or binary. Although the transition is non-linear (yet continuous) in nature, and while most systems can generally be placed within one or the other category[6], there exist intermediate cases. These intermediate cases are usually captured by the notion of *self-organization*, in which local interactions among components spontaneously generate global order that, in turn, stabilizes and maintains itself. Such transitional forms give rise to emergent qualifications that combine aspects of both structural and organizational emergence. Some examples include super-cell thunderstorms (Davies-Jones, 2015) or reaction-diffusion spots (Virgo & Harvey, 2008). Both display distinctive parts, they are composed of different patterns, that altogether contribute to the phenomenon, we can identify them (sometimes we even give them names), local contingent differences between constituent spatial and interactive properties start to make a difference, and these systems tend to develop over time and display cumulative and idiosyncratic histories. However, these are typically short-lived, they display little adaptive capacity, information is easily lost and they tend to decay or dissipate leaving no reproducible traces.

Complex organizations (whether biological, social, or technological) are typically both the result of, and the condition of possibility for, evolutionary, developmental, and nested self-organizing assemblages. They posses a proper individuality and the external manifestation of such complex emergent organization is *agency.*

A crucial feature that distinguishes organizational from some forms of structural emergence (particularly those that give rise to new generic and lawful properties in physical and chemical systems) is that organizational emergence is predicated of specific individual (this particular cell under the microscope right now, the tornado that destroyed Miami last week, or yourself). But how can we define agency as something that doesn't presuppose the terms that typically characterize it (intentionality, reasons, etc.)? How does self-organization complexify to give rise to the emergence of autonomous agency?

6 This non-linear jump from structural to organizational emergence is partly due to the bootstrapping effect of protocellular formation, development and latter evolution, that pushes organizational identity away from generic physico-chemical self-organization (Ruiz-Mirazo et al., 2020).

3. The emergence of basic autonomous agency

3.1. What agency requires

I have elsewhere argued in detail (Barandiaran, 2008; Barandiaran et al., 2009) that agency requires meeting three interrelated, necessary and sufficient, conditions: 1) there has to be a system [individuality condition], 2) doing something by itself [causal asymmetry condition], 3) according to a certain goal or norm [normativity condition]. If there is no well identifiable system there is nothing to which agency can be attributed. Agency is predicated of individuals (or individualized collectives), not of diffuse causal networks[7]. In turn, as we shall see the constitution of this individuality is relevant for the goals and the asymmetry that defines agency. Second, if the individual is not the source of the interaction it establishes with the environment then there it is not an agent, but an patient, a passive receiver of environmental influences. It is not the same to jump than to pushed by the wind. Finally, not every kind of active movement sustained by an individual can be considered an action. Parkinsonian movements or epileptic attacks are endogenously generated but they don't constitute actions. They are not attributable to the individual *as an agent.*

3.2. Conservative and dissipative structures.

A solid departure point to ground the explanation of the emergence of agency is to study the *persistence* of systems. Only relatively stable, enduring, entities will be able to support agency. Surely, this condition is not exclusive to agency, for it extends to almost any kind of entity that we can rely on as part of our world, yet, however obvious, this departure point already helps to define a number of key notions.

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

The first type (conservative structures) refers to spatially ordered forms of assemblage of material sub-units, where order is temporally instantaneous,

7 Although there is a growing liberal usage of the term that has spread its application to different kinds of assemblages, most prominently in neomaterialism (Barad, 2007) and actor-network theory (Latour, 2005). But we shall not enter to discuss them here.

like in rocks or some crystals or temporally unfolded, like in atoms or planetary systems. In both cases the cohesive form exhibited is just an expression of gross physical forces acting between components that, interacting under certain conditions, fall into a stable dynamic or structural stability (some of these structures fall under the category of structural emergence). These systems will stay as they are indefinitely once created: i.e. they are energetically conservative or quasi-conservative. Either chemical bonds (in molecules) or gravitational forces (in planetary systems) lump parts together. The cohesion of such systems will only be destroyed if internal or external fluctuations/perturbations can counteract these forces. However, and precisely because of their form of cohesion, conservative systems are not good candidates to support agency since: 1) they are not capable of generating variable internal states by themselves; if left alone they tend to maximal entropy (i.e. maximum level of disorder) and all their ordered complexity has been externally pre-defined (in terms of initial and boundary conditions, not as a result of the activity of the system), and 2) these systems can perform no work (which, as will be argued latter, becomes central to characterize agency)[8].

The other form of stability is that shown by dissipative structures (Nicolis & Prigogine, 1977). It appears in far-from-equilibrium thermodynamic conditions (FFE hereafter). In a dissipative structure a set of interacting elements generate a cohesive dynamical pattern under an energy gradient in FFE conditions (i.e. far from the state of maximum disorder to which the second law of thermodynamics brings the system in the absence of the continuous flow of matter and energy). Examples of this type of system are whirls, hurricanes, Bénard cells or lasers. In all these different systems a huge amount of microscopic elements adopt a global, macroscopically ordered pattern in the presence of a specific flow of matter and energy. Interestingly, their internal dynamic cohesion is not only a consequence of the material features of their components but also, and most importantly, of a process of circular causality (Haken, 1977). The resulting macroscopic pattern itself contributes to maintain the dynamical cohesion at the microscopic level: by dissipating energy, the pattern contributes to its own stability and cohesion. Thus, these systems are able to generate and maintain, through recursive dynamics, a new way of correlation among their constitutive elements that otherwise would remain disconnected.

8 The case of human made machines, such as robots, that present an organized conservative order capable of channeling energy to produce changes on their environments is a particular case that will receive the attention it deserves at the end of this work. By now it shall be enough to note that these systems require other agents (humans) for their design and continued existence, which shall prevent us from taking them as a departure point. Note also, following Collier and Hooker (1999) on this topic, that such conservative systems are fully determined by externally imposed constraints on their design and thus escape what can be said to be a genuine source of autonomous agency.

3.3. Dissipative order as self-organization

A dissipative structure is often called a self-organizing system. Yet, the term self-organization has been controversial since its very origins in cybernetics (Ashby, 1962) and it is important to note that models and definitions of self-organization are generally domain specific (collective behaviour, chemistry, Artificial Life, biology, ...). In our case we are interested on the domain that deals with the lower boundary of physical sciences and we shall adopt, for our purpose, a preliminary definition that is due to Ruiz-Mirazo: "By self-organization we mean a phenomenon by which local non-linear interactions between elementary units generate a global behaviour (e.g., a spatial/temporal pattern) that is maintained through a certain number of constraints, among which—at least—one is a product of that very phenomenon" (Ruiz-Mirazo, 2001). In this sense self-organized systems constitute an emergent phenomena.

A paradigmatic example of this type of self-organization are Bernad's convection cells (see Figure Error: Reference source not found). These cells appear spontaneously in a liquid layer to which heat is applied from below (Bishop, 2012; Given & Clementi, 1989; Pandey et al., 2018). The initial state of the layer is that of equilibrium and a thermal conduction that extends from the hot area (bottom of the layer) to the top. As the heat increases, however, the uniformity of the fluid breaks, small fluctuations are amplified and convection cells start to form. This cells remain stable and dissipate heat more quickly than thermal conduction does in the equilibrium phase.

Note that the global pattern (the convection cell) is not instructed (dynamically specified) from the outside, nor can it be reduced-to or predicted-from the activity of any of its local components (molecules) alone or to a statistical average of them. As Collier notes (2004), for self-organization to occur three conditions must be met: a) an energy gradient across the boundaries of the system must be present so that internal order can be generated without violating the second law of thermodynamics (this is an external boundary condition)[9], b) components must interact recursively and non-linearly so that correlations between components can be established (this is an internal microscopic or local property); and c) these local interactions need to be able to create global attractors, defined by the critical points that separates different dynamic phases of the system (internal macro-global property). Under an appropriate combination of these three conditions a self-organizing pattern can emerge spontaneously: local fluctuations of interactions between components are amplified to achieve the global configuration that minimizes local entropy production and energy dissipation (Nicolis & Prigogine, 1977).

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

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

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

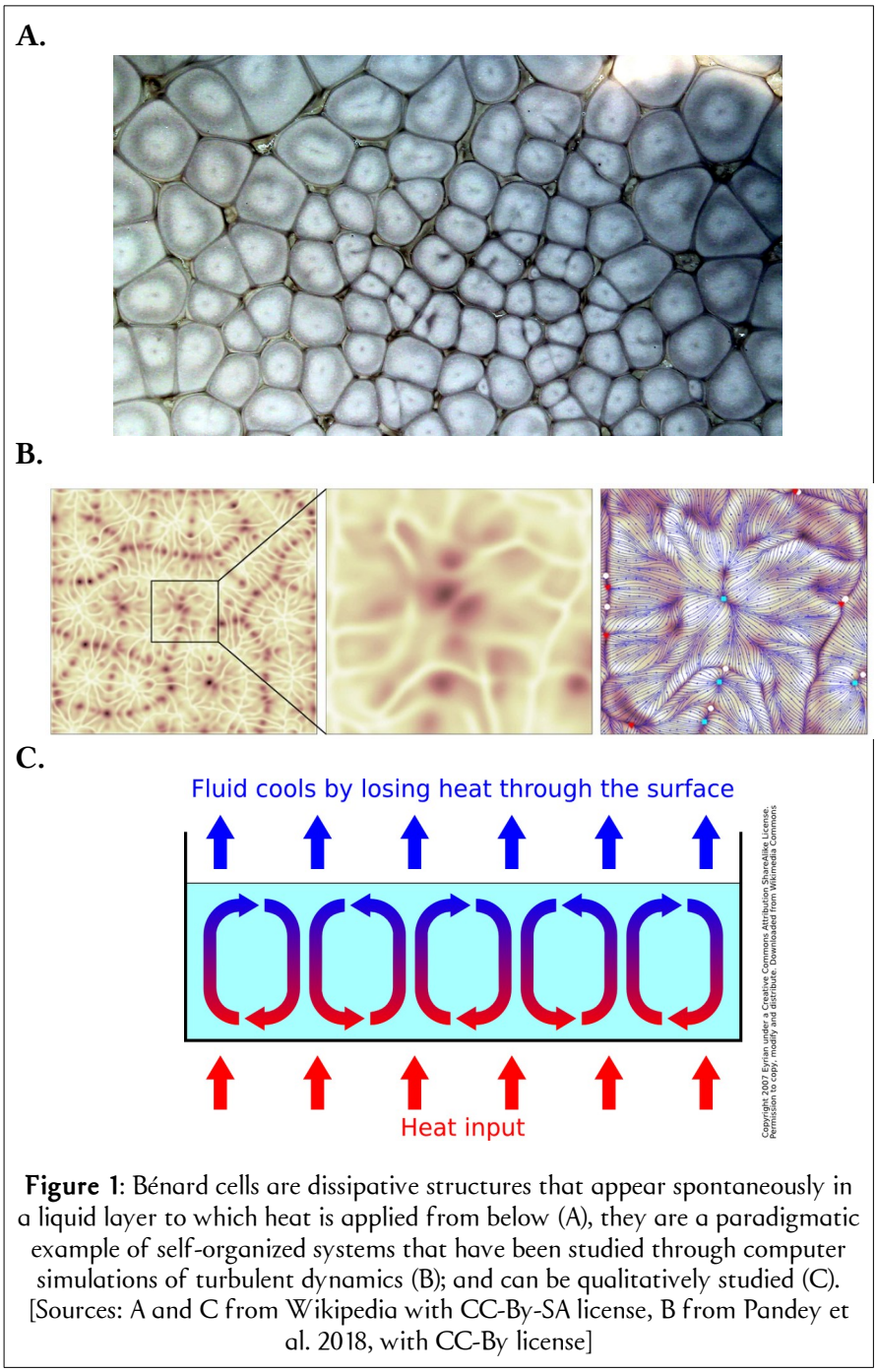

Figure 1: Bénard cells are dissipative structures that appear spontaneously in a liquid layer to which heat is applied from below (A), they are a paradigmatic example of self-organized systems that have been studied through computer simulations of turbulent dynamics (B); and can be qualitatively studied (C). [Sources: A and C from Wikipedia with CC-By-SA license, B from Pandey et al. 2018, with CC-By license]

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

3.4. Towards organized complexity

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

Although generally considered self-*organized* it is not easy to find a proper organization in these physical dissipative structures. The concept of organization usually refers to an integrated disposition of parts or processes within a system, each contributing differentially to different *functions*. In the most simple instances of dissipative structures there is only one part (the global or macroscopic pattern) and only one function (that of constitutive self-maintenance) and the concept of organization loses its meaning[10]. It is in self-organized chemical component production networks where this uniformity is broken and parts (reactions) can be distinguished and may come to contribute differentially to the maintenance of the whole. This differential contribution leads to a dissipative organization with different levels of kinetic and thermodynamic constraints (see Figure 2).

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

1. The shape of their components (the molecules) determines their reactive capacities; they can thus generate different *types* of reactions that may contribute *differentially* to the macroscopic self-maintenance of a network of reactions.
2. The molecules can change or be created combinatorially thus permitting the *creation of new types* of reactions.

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

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

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

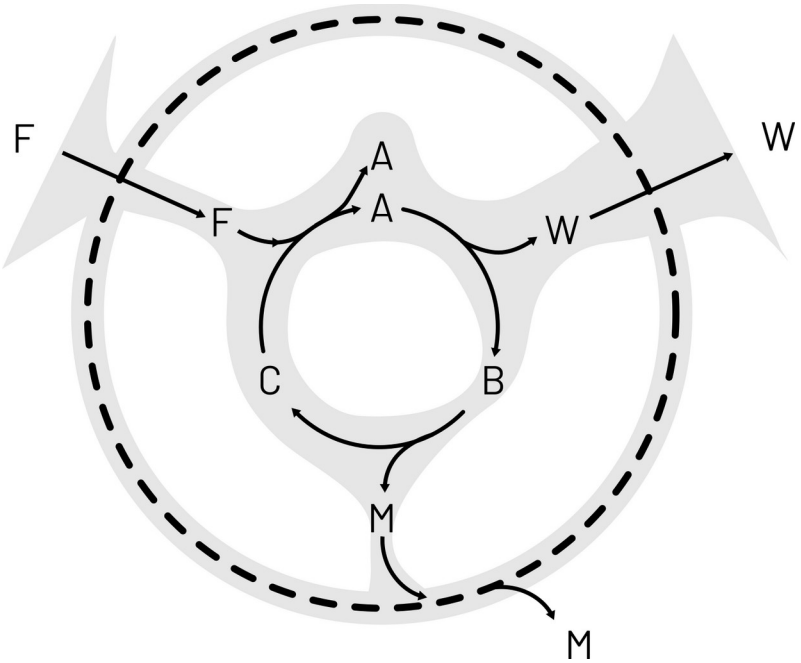

Figure 2: Minimal functional diagram of a self-producing and self-mantaining system composed of metabolites A, B, and C, encycled in a self-producing loop fed by food molecules F and generating a waste product W, while producing a membrane molecule M that encapsulates the whole system and constraints food inflow and waste outflow. The diagram captures [from Barandiaran (2025) with permission and licence CC-By]

As Prigogine and Stengers (1984) emphasized, in non-linear chemical reactions (those that appear under catalytic loops, which are ubiquitous in biological systems) microscopic fluctuations may give rise to highly specific behaviours of the system and permit generating more complex dissipative structures than those found along the "purely physical" domain. Not only is

temporal stability broken, leading to cyclic or strange attractors (defining an intrinsic rhythm), but also spatial homogeneity is broken, leading to spatial structuring. Thus the system determines, intrinsically, its own spatial structure. An interesting intermediate example between chemical systems and living agency are Belousov-Zhabotinsky reactions (Zhabotinsky, 2007) and reaction diffusion spots (Krischer & Mikhailov, 1994; Virgo, 2011). These systems escape from what Prigogine and Stengers have called Boltzmann's principle of order (i.e. that the behaviour of a macroscopic system is equal to the mean behaviour of its constituents, a typically reductive stance). The system appears organized and we can start speaking in terms of local specific states that propagate across the full systems giving rise to differentiated global states. Mesoscopic relationships appear in which micro-macro relationships are not just many-to-one (many components interact to produce a single macroscopic pattern), but many-to-many (differentiated parts generate differentiated global states). We are thus closer to systems capable of having self-defined internal states while collectively generating a global cohesion that may ultimately be expressed in terms of agency.

The idea of a network of reactions that, altogether, maintains a dissipative order through the production of part of the components of the very network is considered by some at the heart of the origin of life and living organization (Dyson, 1982, 1999; Kauffman, 1993; Maturana & Varela, 1980; Morowitz, 1992, 1999; Ruiz-Mirazo et al., 2004), often grouped under the label “metabolism first school”. Hans Jonas beautifully expressed the emergent capacity of metabolism as the continuous material re-generation of an individual (Jonas, 1966, 1968). Thus, the notion of the individual prevails over its renewing constituents. The identity of a cell (or a human) across time cannot be equated with its underlying material constitution because this is continuously changing. It is instead the organization that persists, the web of self-producing relationships (metabolism). A view that, although independently (re)conceptualized, came to have a strong influence on biological and cognitive science through the notion of *autopoiesis* (Maturana & Varela, 1980). Additionally, there's an issue of normativity: a network of interdependent, dynamic presuppositions between processes sustains the individual, determining its precariously self-sustaining nature and the emergence of norms that must be satisfied without any law-like guarantee (Barandiaran, 2025; Barandiaran & Moreno, 2008; Christensen et al., 2002; Weber & Varela, 2002). Jonas beautifully expressed the double dependence-autonomy dialectics that define emergence in contemporary debates in terms of *needful-freedom*. The increase on living agentive capacity is accompanied by a increasing distance between the need and its satisfaction (from animal motility to human imagination). The freer the agent the more dependent it becomes on specific material configurations and environments, more needs have be satisfied, the biggest its precariousness and fragility.

3.5. Basic autonomous agents

Basic autonomous agents are those that satisfy the conditions for agency on the most basic metabolic sense, in the frontier between chemistry and biology (Ruiz-Mirazo & Mavelli, 2008; Ruiz-Mirazo & Moreno, 2004). Protocells (Rasmussen et al., 2008), understood as simple precursors of early living cells, are a good example of basic autonomous agency. A minimal conceptual characterization of a protocell can be depicted in Figure 2. A self-sustaining chemical reaction A → B+W, B→ C+M, and C+F → 2A. The network produces itself in a circular fashion in the presence of F intake and waste output W. While doing so membrane molecules M are produces, and self-assemble encapsulating the reaction network into an *individual* that is distinguishable from its environment. A sense of function and normativity emerges:

> Within this organization, component parts or traits A, B and C and their specific concentrations dynamically presuppose the "right" concentration of the other components and the specific reaction rates. Metabolite B functions (operates normatively) when its concentration and reaction rates match the dynamic presupposition of the rest of components for the self-maintenance of the organization. So, for instance, B needs to produce M at a certain rate for the membrane to grow or replace decaying M molecules at the speed required to avoid bursting or disintegration. M molecules, in turn, need to control the inflow of F at rate sufficiently slow to avoid burst but sufficiently fast to avoid the A→B→C→A reaction to run down, and so forth in a circular manner. (Barandiaran, 2025, p. 5)

When we speak of agency we assume that an identifiable system is *acting*: that it is modulating its coupling with its environment following a certain goal or norm defined by itself. We have the system and we have the norms (dictated by the conditions of self-maintenance). We now need a case of this proto-individual being the source of the interaction with the environment. A minimal candidate is the membrane's modulation of food input and waste output to ensure self-maintenance and avoid osmotic burst respectively (Ruiz-Mirazo & Mavelli, 2008). But most identifiable as agentive is usually motion. This is the case of (proto)cellular (or bacterial) chemotaxis (Barandiaran & Egbert, 2014). When a complexly organized metabolic system propels itself with the net effect of selecting higher concentrations of nutrients it is *acting*[11]. It has compelling "reasons" to do so. It needs to eat.

Protocells, bacteria or plants are a long way from holding genuine intentions[12], but they need to be treated as agents, not as a mere collection of col-

11 Fundamental to this acting are a number of thermodynamic properties of agency that are particularly relevant at the basic autonomous or metabolic level. Agents are capable of channeling energy to produce work, that in turn contributes to generate the constraints that make possible to channel energy in work-constraint cycles that make it possible to speak of an endogenous source of effort and control of behavior (Kauffman, 2003).

12 See Barandiaran and Rama (2025) for a detailed discussion on what sensorimotor teleol-

liding particles (although they are made of them). Their constantly renewed material organization is disposed so that its behavior is directed to the satisfaction of metabolic needs. And this organization operates in a goal-directed fashion, unlike mechanisms. As Denis Walsh puts it:

> "The relation between a goal and the means to its attainment has the same invariant structure as the relation between a mechanism and its effect, but with a twist. In a mechanistic system the effect counterfactually depends upon the cause; in a teleological system, the cause (or means) counterfactually depends upon the effect (goal). (…) In each case—i.e. in mechanistic and teleological invariances—the counterfactual relation is robust. This means that holding a mechanism constant, and intervening on the initial or background conditions, produces a systematic difference in the effect. Similarly, holding the goal of a goal-directed system constant and varying the initial or background conditions, produces a systematic change in the means. (…) In a mechanistic system effects counterfactually depend on causes. In a goal-directed system, the causes counterfactually depend on the goals." (Walsh, 2012, p. 178)

What holds these goal-directed counterfactual relationships together is the complex metabolic organization that produces, selects and modulates different means to achieve a given goal (for a similar account of non-metabolic sensorimotor goal directedness see Barandiaran & Rama, 2025).

4. Agent causation and the scientific understanding of the world

Agentive talk and explanation of what agents *do* (from living cells to intelligence agencies) is scientifically necessary. And justifiably so. It involves explanatory resources unlike those of classical physicalist vocabulary: instead of *events* we speak of *actions*, instead of *laws* we talk of *norms*, instead of *equilibria* we talk of *goals*, instead of *forces* we talk of *intentions*. Agentive vocabulary is handy and useful, mostly because we are agents ourselves and can project our agentive self-understanding to other systems. We sometimes "incorrectly" apply this explanatory framework (such is the case of attributing intentions to moving shapes on a screen when they seem to "try to catch" another shape[13]), sometimes "correctly", attributing agency to systems that do actually posses agentive powers (from dogs to governments, to other humans). But how do we sharply justify the distinction? I have suggested (together with a myriad of philosophers and scientists) that genuine agents exist as emergent autonomous organizations that produce and sustain themselves in FFE conditions, as organized and complex networks of

ogy and intentionality requires.

13 This attribution of intentionality, finality or teleological properties has been documented for a long time (Heider & Simmel, 1944) and starts as early as 12 months of age (Csibra et al., 2003).

functionally distinct yet integrated wholes. When these systems modulate the interaction with their environments to satisfy the needs or goals determined by their own organization, they become agents.

Eliminativists will hold that there is "really" not such a thing as agency: our world is simply populated by physical interactions that deserve no genuine distinction beyond that which is perhaps convenient for our limited minds to enjoy some sort of cognitive economy (as a shortcut for a much complex fine grained explanation). But, they shall argue, agentic explanations depict no real difference on what lies out there. (It would rest to explain why, under which circumstances and to which degree, does this convenience hold in some cases and not in others.) Alternatively, one can take a stronger stance and claim that agency marks a real difference between types of systems, forms of causation and the accompanying epistemic strategies we need to deploy to understand them.

When we have a physical filter (like when you sift flour with a sieve) and we want to predict the final distribution of particles at one and the other side of the filter, we see that it is both the variation in particle size and motion of the filtered particles, *and* the structure of the filter that matters to the explanation. However, the most important part of the explanation is not about the motion of different sized particles, but the *structure* of the filter. It is the *selector* of lower level variations that is the cause of certain particles crossing to the other side, and therefore creating an asymmetric distribution of particles. Also, we gain little by attending to the position of each particle on a filter, like a sieve. It is enough to fix a number of assumptions (e.g. that the size of sieve covers the flour, that the sieve will not brake, etc.) and the size of the holes in the sieve (the structure of the filter) to accurately describe and predict the system. Shake it for enough time and those particles, and only those particles, that are smaller than the holes in the filter will pass. What is the main causes operating to explain the uneven distribution of particles? The structure of the filter. We gain nothing by eliminating the filter from the lower level analysis of the particles. And we gain a lot of explanatory power.

This logic also applies even more strongly,to emergent autonomous agency. Again, self-organizing systems can provide an illuminating intermediate illustrative step, like in Bénard cells. There, the emergent pattern selects the motion of the constituent particles. Yet, unlike a simple sieve filtering flour, here the selector and the selected are part of the same system, and it is the holistic emergent organization that acts as the cause. That emergent organization must be invoked to explain the resulting selective effect. Is the whole organization different from the detailed understanding of all its constituents and all their relationships? No. Is it reducible to any aggregate mean, individual property of constituents, or part of them, other than the totality? No. Is there a macroscopic pattern that is explanatorily relevant, regular and responding to a "logic" of its own macroscopic phenomenology (at an observable and, at least in principle, manipulable level)? Yes. Then we have a genuine case of emergence. Evolutionary, developmental and ecological factors increase the complexity of basic forms of autonomous organizations. Inter-

nal and external fluctuations are selected and functionally integrated in organizations they contribute to sustain in a circular, ever more structured, whole, that progressively subtracts itself from the physicalist reductive explanations of the laws governing its constituents[14].

This is true of a bacterium moving up a sugar gradient, and of yourself seeking a glass of water. Local lower level fluctuations, endogenous or environmental, get selectively amplified by the dynamic organizations they contribute to. This happens at the scales of metabolic reaction networks, genetic regulatory networks, multicellular tissues, ant colonies or electrochemical brain dynamics in embodied interaction with the environment (including other agents). It is in this last domain where we find and identify ourselves as cognitive and intentional agents (Barandiaran, 2007, 2008; Di Paolo et al., 2017; Barandiaran & Rama, 2025).

When cognitivism adopted a mechanistic worldview, assuming the primary function of the mind was to represent nature, the problem of agency surfaced in the philosophy of mind (Davidson, 1980; Frankfurt, 1978). How are mental representational mechanisms, as mere sequences of causal events, capable to ground agency? What is their relationship with physical mechanisms of other sorts? Where is the agent if only a sequence of causal events produces an action? These problems dissolve if we assume that emergence pervades the physical world and that agency emerges from organized physicochemical processes, in a manner that is not reducible to mechanistic terms, while remaining open to further complexification.

Human agency, free will, or personal autonomy are not anchored in an abstract rational space, but deeply materially embodied on nested and meshed domains of emergent organization: psycho-chemical, biological, neurodynamics, social, etc. Surely human brain-bodies and societies are, to some degree, capable of materially abstracting relations and reasons. But such abstraction admits no clean ontological cut from their substrates. If we look at it from the lenses of complexity and self-organization, the human mind emerges from the physical interaction between brain, body and environmental processes (Chemero, 2013). A detailed account of human autonomous agency (or freedom) is out of the scope of this paper. But all forms of agency (including the human one) respond to the same underlying logic. In line with our explanation of the emergence of autonomous agency, we can argue that environmental, bodily and neural (probabilistic) microscopic fluctuations are selected by the agent as a macroscopic emergent entity, bringing about an act of self-determined agency (we might call free-will), that is latter integrated on its own organization. In fact, turbulent-like dynamics, similar to those we previously used to characterize emergence of Bé-

14 Note that when we apply physicochemical laws to explain a given phenomena we don't *just* do it. This application is often accompanied by a set of conditions and constraints that pass inadvertently: scales, boundary conditions, space-time limits, etc. In organizational emergence the amount of this kind of auxiliary information and its interactions is gigantic compared to the role taken by the physical laws at play.

nard cells, are also present in the brain and have been shown to importantly contribute to brain function (Deco & Kringelbach, 2020).

What matters is not the antecedent event, nor (all) the physical laws that are applicable. But the way in which a given recursive organization, selects local (deterministic, stochastic or quantum undetermined) fluctuations of its constituents to channel them into a goal-directed action. Instead of antecedent and/or physical determination, it is rather *self*-determination that agency calls for: the fact that we need to attend to this *self*, its organization, its mode of constitution, its needs and precarious existence, its *raison d'etr*e, to understand what is to happen next.

Were we not to pay attention to this specific agentive organization, its identity and goals, we would have to simulate or compute the whole universe to predict its next move. What any given agent is potentially sensitive to and seeks to achieve is inherently open to its environment in multiple ways (Barandiaran & Etxeberria, 2026). And these potential sensitivities and goals are only specified by its own organization: they are non-deducible from any set of generic physical or chemical laws, nor from the constituents that are continuously being renewed and (re)produced and modified by the very organization.

After all, a simple whisper, or the shine of a star (the sun), can trigger a persistent intentional journey. We don't need to postulate the existence of an immaterial will, and we can't skip the whole organism, its particular mode of organization (functional differentiation and integration) by reducing it to physico-chemical laws. We, being agents ourselves, need to acknowledge that in this universe, autonomous agency *emerged*. And that is a good reason to care for it. Something that only agents can do.

References

Anderson, P. W. (1972). More Is Different. *Science*, *177*(4047), 393–396. https://doi.org/10.1126/science.177.4047.393

Ashby, W. R. (1962). Principles of the self-organizing system. In Jr. Von Foerster H. andZopf (Ed.), *G. W. Principles of Self-Organization* (pp. 255–278). Pergamon Press.

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press.

Barandiaran, X. E. (2007). Mental Life: Conceptual models and synthetic methodologies for a post-cognitivist psychology. In B. Wallace, A. Ross, J. Davies, & T. Anderson (Eds), *The World, the Mind and the Body: Psychology after cognitivism* (pp. 49–90). Imprint Academic.

Barandiaran, X. E. (2008). *Mental Life: A naturalized approach to the autonomy of cognitive agents.* [PhD Thesis, University of the Basque Country (UPV-EHU)]. https://xabier.barandiaran.net/phdthesis/

Barandiaran, X. E. (2017). Autonomy and Enactivism: Towards a Theory of Sensorimotor Autonomous Agency. *Topoi*, *36*(3), 409–430. https://doi.org/10.1007/s11245-016-9365-4

Barandiaran, X. E. (2025). Organizational Accounts of Malfunction: The Dual-Order Approach and the Normative Field Alternative. *Biological Theory*. https://doi.org/10.1007/s13752-025-00500-z

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action. *Adaptive Behavior*, *17*(5),

367–386. https://doi.org/10.1177/1059712309343819
Barandiaran, X. E., & Egbert, M. D. (2014). Norm-establishing and norm-following in autonomous agency. *Artificial Life*, *20*(1), 5–28. https://doi.org/10.1162/ARTL_a_00094
Barandiaran, X. E., & Etxeberria, A. (Eds). (2026). *Outonomy: Fleshing out the Concept of Autonomy Beyond the Individual*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-05501-9
Barandiaran, X. E., & Moreno, A. (2008). Adaptivity: From Metabolism to Behavior. *Adaptive Behavior*, *16*(5), 325–344. https://doi.org/10.1177/1059712308093868
Barandiaran, X. E., & Rama, T. (2025). *Sensorimotor teleology and goal-directedness. An organismic framework for normative behaviour*. https://philsci-archive.pitt.edu/id/eprint/25369
Bedau, M. A. (1997). Emergent Models of Supple Dynamics in Life and Mind. *Brain and Cognition*, *34*(1), 5–27. https://doi.org/10.1006/brcg.1997.0904
Bedau, M. A., & Humphreys, P. (Eds). (2008). *Emergence: Contemporary readings in philosophy and science*. MIT Press. https://epdf.tips/emergence-contemporary-readings-in-philosophy-and-science.html
Bishop, R. C. (2008). Downward causation in fluid convection. *Synthese*, *160*(2), 229–248. https://doi.org/10.1007/s11229-006-9112-2
Bishop, R. C. (2012). Fluid convection, constraint and causation. *Interface Focus*, *2*(1), 4–12. https://doi.org/10.1098/rsfs.2011.0065
Bitbol, M. (2012). Downward causation without foundations. *Synthese*, *185*(2), 233–255. https://doi.org/10.1007/s11229-010-9723-5
Boltzmann, L. (1964). *Lectures on Gas Theory*. Dover Publications.
Boogerd, F. C. (Ed.). (2007). *Systems biology*. Elsevier.
Broad, C. D. (1925). *The Mind And Its Place In Nature*. Kegan Paul, Trench, Trubner & Co., Ltd.
Buzsaki, G. (2006). *Rhythms of the Brain* (1st edn). Oxford University Press, USA.
Campbell, D. T. (1974). 'Downward Causation' in Hierarchically Organised Biological Systems. In F. J. Ayala & T. Dobzhansky (Eds), *Studies in the Philosophy of Biology: Reduction and Related Problems* (pp. 179–186). Macmillan Education UK. https://doi.org/10.1007/978-1-349-01892-5_11
Carnap, R. (1934). *The Unity of Science*. Routledge.
Carnap, R. (1959). *Psychology in physical language*. na. https://web.stanford.edu/~paulsko/papers/Carn.pdf
Chalmers, D. J. (1997). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
Chemero, A. (2013). Radical Embodied Cognitive Science. *Review of General Psychology*, *17*(2), 145–150. 2025-10-30. https://doi.org/10.1037/a0032923
Cheung, H. F. H., Patil, Y. S., & Vengalattore, M. (2018). Emergent phases and critical behavior in a non-Markovian open quantum system. *Physical Review A*, *97*(5), 052116. https://doi.org/10.1103/PhysRevA.97.052116
Chomsky, N. (1959). A Review of B. F. Skinner's Verbal Behavior. *Language*, *35*, 26–58.
Christensen, W. D., Bickhard, M. H., & The Hegeler Institute. (2002). The Process Dynamics of Normative Function: *Monist*, *85*(1), 3–28. https://doi.org/10.5840/monist20028516
Churchland, P. M. (1981). Eliminative Materialism and Propositional Attitudes. *The Journal of Philosophy*, *78*(2), 67–90. https://doi.org/10.5840/jphil198178268
Clark, A. (1998). *Being there: Putting brain, body, and world together again*. MIT Press.
Collier, J. (1988). Supervenience and Reduction in Biological Hierarchies. *Matthen, M. & Linsky, B. (Eds.) Philosophy and Biology: Canadian Journal of Philosophy Supplementary*, *14*, 209–234.
Collier, J. (2004). Self-organization, individuation and identity. *Revue Internationale de Philosophie*, 151–172.
Collier, J., & Hooker, C. A. (1999). Complexly organised dynamical systems. *Open Systems & Information Dynamics*, *6*(3), 241–302.
Crick, F. (1966). *Of Molecules and Men*. University of Washington Press.
Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111–133. https://doi.org/10.1207/s15516709cog2701_4

Davidson, D. (1980). *Essays on Actions and Events* (Underlining). Oxford University Press, USA.
Davies-Jones, R. (2015). A review of supercell and tornado dynamics. *Atmospheric Research*, *158–159*, 274–291. https://doi.org/10.1016/j.atmosres.2014.04.007
Dawkins, R. (1976). *The selfish gene*. Oxford University Press.
Deco, G., & Kringelbach, M. L. (2020). Turbulent-like Dynamics in the Human Brain. *Cell Reports*, *33*(10), 108471. https://doi.org/10.1016/j.celrep.2020.108471
Di Paolo, E. A., Buhrmann, T., & Barandiaran, X. E. (2017). *Sensorimotor Life: An enactive proposal*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198786849.001.0001
Dirac, P. (1929). Quantum mechanics of many-electron systems. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *123*(792), 714–733. https://doi.org/10.1098/rspa.1929.0094
Dyson, F. J. (1982). A model for the origin of life. *Journal of Molecular Evolution*, *18*(5), 344–350. https://doi.org/10.1007/BF01733901
Dyson, F. J. (1999). *Origins of life*. Cambridge University Press.
Edelman, G., & Tononi, G. (2001). *A Universe Of Consciousness How Matter Becomes Imagination*. Basic Books.
Fodor, J. A. (1968). *Psychological Explanation; An Introduction to the Philosophy of Psychology*. Random House Inc.
Frankfurt, H. G. (1978). The Problem of Action. *American Philosophical Quarterly*, *15*(2), 157–162.
Given, J. A., & Clementi, E. (1989). Molecular dynamics and Rayleigh-Benard convection. *Journal of Chemical Physics*, *90*, 7376–7383. https://doi.org/10.1063/1.456217
Haken, H. (1977). *Synergetics—An introduction: Nonequilibrium phase transitions and self-organization in physics, chemistry and biology*. Springer.
Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243–259. https://doi.org/10.2307/1416950
Jaworski, W. (2016). *Structure and Metaphysics of the Mind: How Hylomorphism Solves the Mind-Body Problem*.
Jonas, H. (1966). *The Phenomenon of Life. Toward a Philosophy of Biology*. Chicago-London.
Jonas, H. (1968). Biological foundations of individuality. *International Philosophical Quarterly*, *8*(2), 231–251.
Kauffman, S. A. (1993). *The origins of order*. Oxford University Press US.
Kauffman, S. A. (2000). *Investigations*. Oxford University Press US.
Kauffman, S. A. (2003). Molecular autonomous agents. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, *361*(1807), 1089–1099. https://doi.org/10.1098/rsta.2003.1186
Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*.
Kim, J. (1999). Making Sense of Emergence. *Philosophical Studies*, *95*, 3–36.
Kitano, H. (2002). Systems Biology: A Brief Overview. *Science*, *295*(5560), 1662–1664. https://doi.org/10.1126/science.1069492
Krischer, K., & Mikhailov, A. (1994). Bifurcation to Traveling Spots in Reaction-Diffusion Systems. *Physical Review Letters*, *73*(23), 3165. https://doi.org/10.1103/PhysRevLett.73.3165
Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.
Laughlin, R. B., & Pines, D. (2000). The Theory of Everything. *Proceedings of the National Academy of Sciences*, *97*(1), 28–31. https://doi.org/10.1073/pnas.97.1.28
Lewis, P. J. (2017). Quantum mechanics, emergence, and fundamentality. *Philosophica*, *92*(2). https://doi.org/10.21825/philosophica.82111
Lotka, A. J. (1920). Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences of the United States of America*, *6*(7), 410.
Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition*. D. Reidel Publishing Company.
McGregor, S., & Fernando, C. (2005). Levels of Description: A Novel Approach to

Dynamical Hierarchies. *Artificial Life*, *11*(4), 459–472. https://doi.org/10.1162/106454605774270615
Moreland, J. P. (2011). *Bioethics, Substance Dualism and the Argument from Self-Awareness*. 11.
Moreno, A., & Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Springer.
Morowitz, H. J. (1992). *Beginnings of Cellular Life*. Yale University Press.
Morowitz, H. J. (1999). A theory of biochemical organization, metabolic pathways, and evolution. *Complexity*, *4*(6), 39–53.
Mossio, M., Saborido, C., & Moreno, A. (2009). An Organizational Account of Biological Functions. *The British Journal for the Philosophy of Science*, *60*(4), 813–841. https://doi.org/10.1093/bjps/axp036
Nagel, K., & Paczuski, M. (1995). Emergent traffic jams. *Physical Review E*, *51*(4), 2909–2918. https://doi.org/10.1103/PhysRevE.51.2909
Newman, S. A. (2018). Inherency. In L. Nuno de la Rosa & G. Müller (Eds), *Evolutionary Developmental Biology* (pp. 1–12). Springer International Publishing. https://doi.org/10.1007/978-3-319-33038-9_78-1
Nicolis, G., & Prigogine, I. (1977). *Self-organization in non-equilibrium systems: From dissipative structures to order through fluctuations*. Wiley, New York.
O'Connor, T. (2020). Emergent Properties. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/properties-emergent/
Pandey, A., Scheel, J. D., & Schumacher, J. (2018). Turbulent superstructures in Rayleigh-Bénard convection. *Nature Communications*, *9*(1), 2118. https://doi.org/10.1038/s41467-018-04478-0
Popper, K. R., & Eccles, J. C. (1977). *The Self and Its Brain*. Springer International.
Prigogine, I., & Stengers, I. (1984). *Order out of chaos: Man's new dialogue with nature*. Bantam books.
Putnam, H. (1965). The mental life of some machines. In *Philosophical Papers: Volume 2, Mind, Language and Reality* (pp. 408–428). Cambridge University Press.
Ramachandran, V. S. (2012). *The Tell-Tale Brain: Unlocking the Mystery of Human Nature*. Random House.
Rasmussen, S., Bedau, M. A., Chen, L., Deamer, D., Krakauer, D. C., Packard, N. H., & Stadler, P. F. (Eds). (2008). *Protocells: Bridging Nonliving and Living Matter* (1st edn). The MIT Press.
Reynolds, C. W. (1987). Flocks, Herds, and Schools: A Distributed Behavioral Model. *Computer Graphics, Proceedings of SIGGRAPH'87*, *21*(4), 25–34.
Ruiz-Mirazo, K. (2001). *Condiciones físicas para la aparición de sistemas con capacidades evolutivas abiertas* [PhD Thesis]. Physical conditions for the appearance of autonomous systems with open-ended evolutionary capacities] San Sebastián.
Ruiz-Mirazo, K., & Mavelli, F. (2008). On the way towards 'basic autonomous agents': Stochastic simulations of minimal lipid–peptide cells. *BioSystems*, *91*(2), 374–387.
Ruiz-Mirazo, K., & Moreno, A. (1998). Autonomy and emergence: How systems become agents through the generation of functional constraints. *Acta Polytechnica Scandinavica, Ma91*, 273–282.
Ruiz-Mirazo, K., & Moreno, A. (2004). Basic Autonomy as a Fundamental Step in the Synthesis of Life. *Artificial Life*, *10*(3), 235–259. https://doi.org/10.1162/1064546041255584
Ruiz-Mirazo, K., Pereto, J., & Moreno, A. (2004). A universal definition of life: Autonomy and open-ended evolution. *Origins of Life and Evolution of the Biosphere*, *34*(3), 323–346.
Ruiz-Mirazo, K., Shirt-Ediss, B., Escribano-Cabeza, M., & Moreno, A. (2020). The Construction of Biological 'Inter-Identity' as the Outcome of a Complex Process of Protocell Development in Prebiotic Evolution. *Frontiers in Physiology*, *11*. https://doi.org/10.3389/fphys.2020.00530
Rumelhart, D. E., McClelland, J. L., & Group, the P. R. (1987). *Parallel Distributed Processing, Vol. 1: Foundations*. The MIT Press.
Russell, E. S. (1916). *Form and Function: A Contribution to the History of Animal Morphology*.

John Murray. http://www.gutenberg.org/ebooks/20426
Seth, A. K. (2011). Measuring Autonomy and Emergence via Granger Causality. *Artificial Life*, *16*(2), 179–196. https://doi.org/10.1162/artl.2010.16.2.16204
Skinner, B. F. (1953). *Science and human behavior* (2005 B.F. Skinner Foundation). Macmillan. http://www.bfskinner.org/BFSkinner/PDFBooks.html
Smolensky, P. (1988). On the Proper Treatment of Connectionism. *Behavioral and Brain Sciences*, *11*(01), 1–23. https://doi.org/10.1017/S0140525X00052432
Sole, R., & Goodwin, B. (2002). *Signs Of Life: How Complexity Pervades Biology*. Basic Books.
Sornette, D., & Cauwels, P. (2015). Financial Bubbles: Mechanisms and Diagnostics. *Review of Behavioral Economics*, *2*(3), 279–305. https://doi.org/10.1561/105.00000035
Swinburne, R. (2019). *Are we bodies or souls?*
Thompson, E., & Varela, F. J. (2001). Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences*, *5*(10), 418–425. https://doi.org/10.1016/S1364-6613(00)01750-2
Tolman, E. C. (1967). *Purposive Behavior in Animals and Men*. Appleton-Century-Crofts.
Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44
Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *237*(641), 37–72. https://doi.org/10.1098/rstb.1952.0012
Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press.
Virgo, N. (2011). *Thermodynamics and the structure of living systems* [Thesis, University of Sussex]. http://sro.sussex.ac.uk/6334/
Virgo, N., & Harvey, I. (2008). Reaction-diffusion spots as a model for autopoiesis. *Artificial Life*, *11*, 816.
Walsh, D. (2012). Mechanism and purpose: A case for natural teleology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 173–181. https://doi.org/10.1016/j.shpsc.2011.05.016
Watson, J. B. (1913). Psychology as the Behaviorist Views it. *Psychological Review*, *20*. http://psychclassics.yorku.ca/Watson/views.htm
Weber, A., & Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*, *1*(2), 97–125.
Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models* (arXiv:2206.07682). arXiv. https://doi.org/10.48550/arXiv.2206.07682
Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus* (C. K. Ogden, Trans.). Routledge.
Zhabotinsky, A. M. (2007). Belousov-Zhabotinsky reaction. *Scholarpedia*, *2*(9), 1435. https://doi.org/10.4249/scholarpedia.1435

Acknowledgments

I acknowledge IAS-Research group funding IT1668-22 from Basque Government, grant PID2023-147251NB-I00 for project “OUTAGENCIES: Varieties of autonomous agency across living, humanimal and technical system” funded by MCIU/AEI/10.13039/501100011033 and FEDER/UE.

The content of this paper has reused materials from my teaching of reductionism and emergence, within the course of Philosophy of Science at the UPV/EHU master degree in “Filosofía, Ciencia y Valores” (mostly section 2) and parts of my (unpublished) PhD thesis (mostly in section 3).

I am deeply thankful to Manu Barandiaran and Kepa Ruiz-Mirazo for a detailed revision of this manuscript. In particular KRM's work has been, for more than 20 years, a guide for me into the emergence of autonomous agency from physics (together with many other aspects of life).

Conflict of interest statement

The author has no competing interests to declare that are relevant to the content of this article.