On Minimal Autonomous Agency: Natural and Artificial

Alvaro Moreno
Department of Logic and Philosophy of Science
University of the Basque Country UPV/EHU
Avenida de Tolosa 70/20018 Donostia/San
Sebastián, Spain

The idea of minimal agency (MA) may be understood as the simplest agent within the known phenomenological domain, that is, the simplest agent as-we-know-it, and as the simplest agent that could exist, either synthetically or in a hypothetical process of biogenesis, that is, the simplest agent as-it-should-be. The second view is more radical, since it focuses on the simplest organizational and material conditions required for generating agential capacities, not on how they are in fact minimally instantiated. It searches for the simplest material building blocks that, either naturally or artificially, could achieve the simplest form of organization necessary to display agency. Yet, although synthetic methodologies may seem a more adequate strategy to generate minimal forms of agency, I argue that the study of how the biological domain has generated agents is ultimately necessary to understand paradoxical cases of minimal agents and shows us fundamental lessons for their artificial fabrication.

Keywords: minimal agency; autonomy; synthetic agents; natural agents; regulation

1. Introduction: What Is Agency?

The concept of agency is used in many domains, with a different sense in each case. Thus, what is meant by the term agent is understood differently in human sciences, robotics, computer sciences and biology, for example. Unfortunately, if we look for a fundamental account of agency, here too we find a wide diversity of responses. We can, however, limit this diversity (to a certain extent, at least) by initially adopting an eminently inclusive view. For example, broadly speaking, agency can be defined as the capacity to act or exert power (Webster's dictionary) or to have direct control or guidance over one's own behavior (Stanford Encyclopedia). However, since what we often mean by agency is ultimately an extension of human action (i.e., a robot), here we are interested in autonomous agency.

https://doi.org/10.25088/ComplexSystems.27.3.289

In this sense, different authors have tried to include certain forms of "autonomy" in the definition of agency. Simply by way of example, an agent is:

- a system that tries to fulfill a set of goals in a complex, dynamic environment [1]
- a system that can initiate, sustain and maintain an ongoing and continuous interaction with its environment as an essential part of its normal functioning [2]
- a system situated within an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to affect what it senses in the future [3]
- a system that can act on its own behalf in an environment [4]
- an entity that engages in normatively constrained, goal-directed interaction with its environment [5]
- a system doing something by itself according to its own goals or norms within a specific environment [6]

There are several aspects worth considering in these different definitions. First, an agent is the source ("the self") of a set of actions. This implies the existence of a relatively stable system clearly differentiated from the environment, namely, an "identity" or an "individuality." Second, the agent must not only interact with its environment, the actions should also be causally generated by the agent ("causal asymmetry"), modifying the environment. Last but not least, the actions should be functional, which means that they should be "purposeful." Thus, an action is always an oriented process, aimed at achieving a certain goal, and could therefore *fail* if it does not achieve that goal. In order to achieve all these conditions, an agent must be able to selectively distinguish some relevant features in its environment, so as to trigger functional responses to them. In sum, the action of an agent is thus defined as a set of processes in its environment, generated by the agent as an integrated whole (a "self"), which is the causal source of these processes, performed according to a certain goal or norm. This is why the concept of agency is attracting increasing attention both in biology and cognitive science, since it involves fundamental aspects such as function, purpose, sensing and acting, as well as the distinction between self and environment, and for some, it even constitutes the basis of cognition.

Yet, when we move from the realm of abstract definitions to the world of real systems, we face another type of problem: there is no general agreement on whether robots and other artificial systems can be considered true autonomous agents. We easily recognize autonomous agency in humans and animals, but it is much more difficult to determine what is a minimal autonomous agent. For example, for some authors, only certain biological organisms (such as animals) are autonomous agents, while others argue that a eukaryotic cell or even a virus displays a minimal form of autonomous agency. The debate is even fiercer in relation to artificial devices: is an artificially designed robot an autonomous agent? In fact, the only generally agreed-upon examples of agency are humans and (certain) biological systems. The reason for this lies in the problematic autonomy of any artificial device, which, for many, makes it a mere extension (or maybe a potentiation) of the autonomous agency of its human creators. This is why, having analyzed what constitutes an autonomous agent, I now address the question of what might constitute its minimal form. In Section 3, I critically review several proposals; then, based on the analysis of natural (biological) systems, I discuss what the necessary features may be for minimal agency (MA). Finally, I discuss the differences of the concept of MA in artificial and natural contexts and its implications.

2. Implications

Unlike a merely physical dissipative structure (e.g., a hurricane), which maintains its identity as long as certain specific boundary conditions are met, *a system is autonomous if it actively maintains its identity*: for example, by modulating its internal, constitutive organization, in accordance with environmental changes. The simplest forms of self-maintaining systems, such as hurricanes or candles, cannot really "do" anything in order to compensate for new environmental conditions; they have only a small margin of maintenance provided by a buffering or flexible structure. However, even a minimal autonomous system, such as a hypothetical cellular proto-organism, would deploy active self-maintenance by functionally modulating its own internal organization. But this active maintenance is not, by itself, a form of agency. Autonomous agency requires that the autonomous system deploy constraints that *functionally affect processes and reactions in the system's environment*. In other words, autonomous agency requires that the processes on which it exerts a causal influence belong to the external environment, which, in turn, means that these processes have not already been constrained by the autonomous system.[1]

This prompts the need to make a fundamental distinction between *constitutive* processes, which generate identity and largely delimit what the system actually is, and *interactive* processes, which are side effects of the former, but with the specific function of controlling interactions with the environment. Ultimately, the latter are also crucial to maintaining the identity of the system, but constitutive processes are faster and more fundamental than interactive ones. The reason is that since an agent is a precarious (i.e., far from equilibrium) self-producing organization, it is logically necessary that it exist (at least for a while) in order to have the possibility to display actions; otherwise, there would not be the machinery for displaying interactive processes. As a subset of constitutive processes, interactive processes are generated by and depend on the existence and stability of the whole autonomous organization. In turn, they contribute to the maintenance of that very organization by specifically managing its relations with its environment [7, 8].

As said earlier, an autonomous system is able to actively modulate itself so as to functionally respond to environmental changes. Yet an autonomous system can functionally respond to environmental changes by either modifying its internal organization or functionally modifying the external environment (or both). Only in this last case is an autonomous system an agent. Therefore, agency requires more than self-maintenance, since an agent is an autonomous system that induces functional changes in its environment, which means that these changes will give a return on the agent's "investment" by contributing to its self-maintenance. A system is an agent if it is capable of deploying interactive processes in its environment in order to modify external conditions in such a way that these processes contribute to the ongoing maintenance of the system's constitutive organization. This requirement excludes processes that, while functionally adaptive, do not modify the external environment, but only the constitutive organization of the system.

The simplest forms of self-maintaining autonomous systems, such as hypothetical proto-metabolic organizations (probably encapsulated), may compensate for internal and external perturbations by means of feedback mechanisms integrated into their constitutive organization. A variation affecting a given component may propagate within the system and trigger the variation of one or several other components, which in turn compensate for the initial one. In this way, the system may achieve homeostatic stability.[2] However, since these adaptive mechanisms consist only of modifying the constitutive organization of the system, in accordance with environmental conditions but without affecting them, they hardly qualify as a form of agency. What, then, is lacking?

If we analyze carefully the concept of agency, we see that for the deployed interactions to be functional, the agent must be able to *control its relationship with the environment in accordance with the circumstances*. More specifically, agency requires a form of anticipation. So, if the action is to be functional, the agent must compare different possibilities and determine which one will lead it to a better (or less bad) fulfillment of its boundary of viability.[3] In other words, an agent is a system that is able to evaluate sequentially temporal situations and determine which possibility is functional at each moment in time. In other words, it is a system capable of assessing or measuring the specificity of each stimulus at a given moment. Thus, an agent has the ability not just to avoid negative tendencies, but to actively seek to improve its situation.[4] As formulated by DiPaolo [9], if the states are sufficiently close to the boundary of viability: (1) tendencies are distinguished and acted upon depending on whether the states 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 kind, thus preventing future states from reaching the boundary with an outward velocity.[5] To do this, the system should also be capable of switching between different alternatives or adjusting them, and so forth, in accordance with external changes.[6]

Organizationally, this requires a subsystem that operates in a relatively autonomous way from the basic constitutive processes. It is clear that only a dynamical decoupling from low-level constitutive processes will permit this higher-level subsystem to develop a free search for new functional interactive processes. We are therefore talking about two dynamic levels in the system: the constitutive level, which ensures ongoing self-construction, and the (now decoupled) interactive subsystem, which functionally modifies the operations of the former [10]. This is the essence of regulated adaptivity [11]. This decoupling of regulatory mechanisms from the basic constitutive network allows a selective choice from among a large number of not yet functional dynamical states of the constitutive self-maintaining network. The capacity to differentiate between and compensate for tendencies requires that whatever makes a distinction and generates a compensation be dynamically differentiated from what it distinguishes and acts upon. This presupposes that operational mechanisms are capable of distinguishing between the different implications of equally viable paths of encounters with the environment (Figure 1).

The simplest mechanisms of adaptive regulation could be of two different types: either they operate by modifying the constitutive organization of the system (as exemplified in the *lac* operon activation and deactivation of genes, as a switch between metabolic pathways according to certain environmental conditions), or they operate by modifying the external conditions of the system (as exemplified in bacterial chemotaxis, where a subsystem of biochemical pathways, not directly involved in the basic metabolic network, modifies through movement the environmental conditions of the system). The common characteristic of both cases is that some degree of dynamic decoupling from the basic constitutive processes is required [11, 12] (Figure 2).

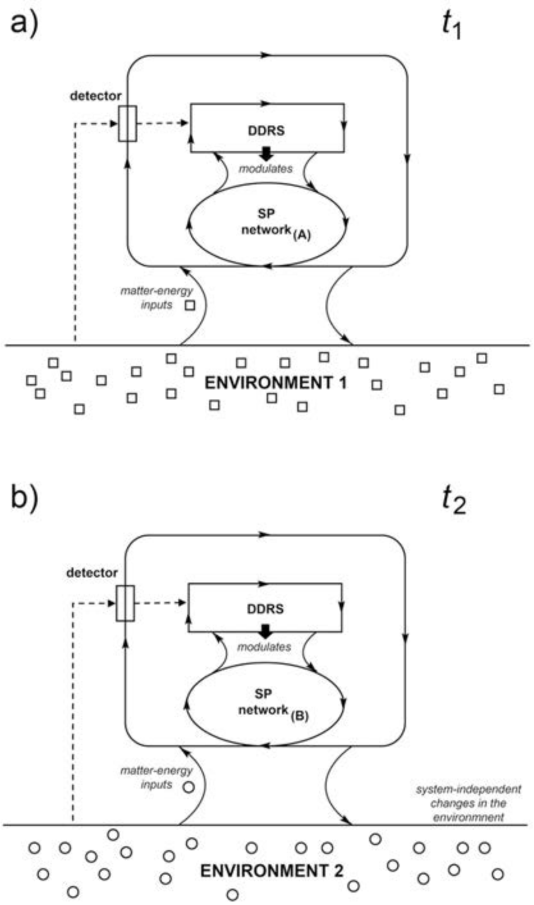

Figure 1. Schematic representation of an autonomous non-agential system. When the environment changes (below) the system is capable (within a certain range) of detecting these changes and functionally modifying its own internal organization through a regulatory subsystem, so as to maintain its own identity. SP network represents the cyclic processes of self-production that constitute and maintain all the parts of the system (black arrows); DDRS stands for a dynamically decoupled regulatory subsystem, exerting control actions (black bold arrows). Dashed arrows represent relevant information about states of the environment detected by the system.

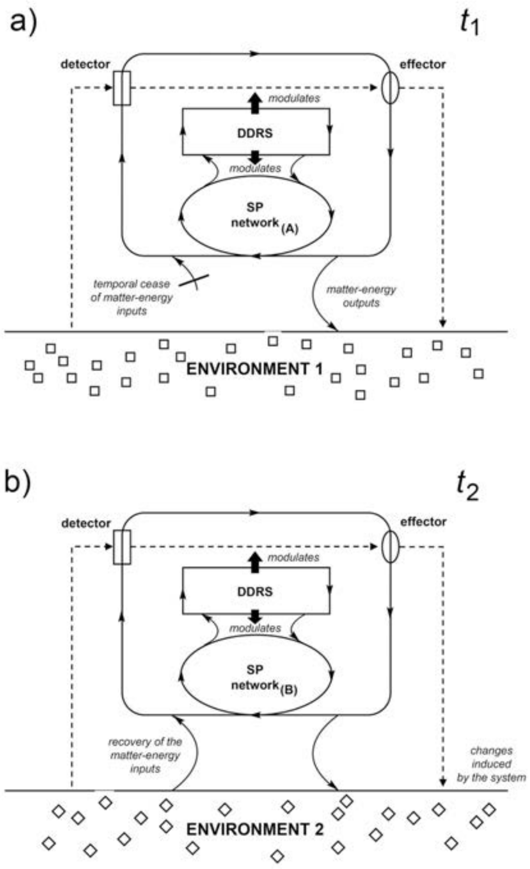

Figure 2. Schematic representation of an autonomous minimal agent. The system is capable of detecting the relevant features of its close environment and of triggering processes that, when needed, modify functionally the environmental conditions (for example, by moving toward better conditions). The figure shows the control exerted by the regulatory subsystem, which modulates detection and effector processes so as to achieve successful action-detection loops. SP network represents the cyclic processes of self-production that constitute and maintain all the parts of the system (black arrows); DDRS stands for a dynamically decoupled regulatory subsystem, exerting control actions (black bold arrows). Dashed arrows represent relations of detection and action.

Accordingly, the appearance of agency implies the emergence of detection and effector mechanisms by which the adaptive regulatory subsystem links the environmental conditions to the constitutive organization of the autonomous systems. It is necessary that the processes triggered to modify the external boundary conditions be correlated with the measurements of the relevant changes induced by them: this should be an ongoing process that therefore needs to be monitored to be functional. Hence, two regulatory processes emerge as being interlinked: the adaptive regulation of the essential variables through recursive interactions with the environment, and the regulation of this interactive cycle in accordance with its effects on the essential variables. This point is important: agency is necessarily a cyclical process, because it requires that the effector processes be modulated in accordance with the detected environmental conditions. The interactive subsystem works by measuring different conditions, triggering correlative effector actions, and monitoring its own constitutive processes so as to avoid or prevent dysfunctional situations. In sum, it is this intertwined regulation, along with the detection-response coupling, that gives rise to genuine agency[7] [10].

3. Candidates for Minimal Agency

Having clarified what exactly an agent is, let us now address the next question: what is the simplest or minimal organization necessary for a system to behave as an agent? By MA, I mean the simplest form, either naturally generated or artificially built, of a system that fulfills the aforementioned characteristics. There is a wide variety, even disparity, of proposals regarding MA. This disparity is partly due to the diversity of research perspectives and domains: synthetic biology and artificial life, computer science, the philosophy of biology and cognitive science, to name but a few. Consequently, there are not only different views on what can be conceived of as MA, but also different opinions regarding how to approach the question. Thus, some proposals emphasize bottom-up methodology and attempt to show how, starting from purely physicochemical principles and using relatively simple material building blocks, we can *synthetically* generate minimal agents. Or, even more radically, other authors propose that certain computer programs could be considered as minimal agents, too. This is the case with software agents, self-propelling droplets, nanorobots and synthetic protocells. Others instead focus on already constituted (natural) systems that are so simple that we can *analytically* understand their structure or organization. This is the case with viruses or prions.

Unfortunately, an overemphasis on organizational or structural simplicity with regard to possible candidates for agents tends to weaken the requirements for rigor and precision in connection with the analysis of what they are really capable of achieving. For their part, proposals that emphasize instead the analysis of the resulting properties and capacities, looking to rigorously satisfy all agency criteria, tend to be less clear concerning the structural and organizational conditions of their generation, and even less so concerning how they could be synthetically generated from scratch.

More specifically, at one end of the scale there is a group of scholars who tend to view agency as an intentional, mind-based behavior and therefore consider MA to be, at most, present in low animals [13] or perhaps unicellular protozoa [14]. At the opposite end are those who believe that MA is already present in "infra-biological" entities, such as software agents, synthetic self-propelling droplets, autonomous molecular machines (e.g., nano-robots) and hypothetical prebiotic proto-organisms. In a similar vein, other authors argue that some biological infra-organismic entities, such as hypothetical prebiotic nude replicators or present-day viruses, plasmids or prions, constitute a valid example of MA. Let us now analyze these different cases.

3.1 Software Agents

Wikipedia defines a software agent as "a computer program that acts for a user or other program in a relationship of agency." Of course, in this context "acting" is understood as performing operations in computer media. Yet, these programs are called "agents" not only because they perform operations but also because, to a certain degree, they behave autonomously. For example, these computer programs are often seen as agents because they are capable of detecting certain features of their environment and act seeking their "own" goals, without being explicitly programmed for these operations. Besides, they are candidates for minimal agency because it would always be possible in principle to specify an agent—or at least a class of them—whose program is algorithmically shorter than that of the others. Since the capacities of these agents are specified by their program, one could argue that (provided their behavior fulfills the definitory requirements of agency) those agents whose programs have the minimal Chaitin–Kolmogorov description length would also be candidates for MA.

Cellular automata (CAs) constitute the most important type of software agent. CAs are networks of simple automata connected locally that produce an output from an input, modifying their state in the process, according to a transition function. More formally, a cellular automaton (CA) is defined as a cellular space or set of cells regularly distributed in an n-dimensional space. There is, for the automaton, a discrete measure of time that is called sequence of generations $t_1, \ldots, t_m$. Every cell c_j can be found, in a given generation t_i, in a state that belongs to a finite set of states. The set of states of all cells in a given generation is called the CA configuration. The neighborhood of a cell is composed of a finite set of cells, and generally, the neighborhood is defined in the same way for all space cells, although different neighborhoods can be defined for different cells (irregularly connected automaton). The state of a cell in a given generation depends on the states of the neighboring cells and on its own state in the previous generation. And the rule that determines the state of a cell from the states of its neighbors and from its own state in the previous generation is called transition function *F*.

The concept of CAs dates back to the pioneering work of von Neumann and Ulam in the 1940s and was later developed by researchers like Conway, Toffoli and Wolfram, among others. CAs are useful tools for modeling any system in the universe. They have been used to model physical systems as diverse as interactions between particles and galaxy formation; kinetics of molecular systems and crystal growth; biological systems, at the cellular, multicellular and population level; and in computer science, to model von Neumann's parallel processing systems and self-producing automata. And it can even be considered that networks of formal neurons are specialized CAs. Wolfram [15] has studied the types of behaviors that CAs could display and has demonstrated that a class of CA is capable of universal computation and therefore, of simulating any other automaton. Langton [16], for his part, has proposed another classification based on the value of a parameter λ that measures the relationship between the number of neighboring states that provide null states and the total number of neighboring states. According to this view, he found three classes of CA. Class 2, which corresponds to Wolfram classes 2 and 4, is for Langton the most interesting because it allows the appearance of virtual agents. Virtual automata are able to carry out structural construction processes in general, since they can write in the environment in which they are located. In the same way, they are able to erase, since they can write the null state. And they are capable of self-deleting. As a consequence of the two previous characteristics, they are able to build other automata, read them and modify their structure. For all this, according to Langton, in such automata, one can recognize the typical functions of biological organisms. In 1989, Langton [17] defended the idea that from this type of system all the functional features of living beings (hence agency) can be defined. Accordingly, CAs could display appropriate actions in each circumstance. They are able, for example, to detect certain features of their environment and act to change part of the environment or of its status, and influence what they sensed. Besides, insofar as—contrary to what is usual in standard AI—CAs are not explicitly programmed to do what they do, their behavior can be considered as "emergent." Based on these facts, many researchers—especially in the domain of

AI—consider certain CA-based programs as true autonomous agents that act in pursuit of their "own" agendas.

The problem with CAs as candidate minimal agents has two interrelated aspects. On the one hand, there is the problem of autonomy. As we have argued in Sections 1 and 2, the requirement of being autonomous is not so easy to fulfill. It implies the capacity to display normative actions, namely, that the action could be evaluated as a success or failure by the agent (and not by its creator). If what the agent seeks is ultimately a human decision or if it would be the human designer who will evaluate the success or failure of the agent's behavior, it may hardly be considered as truly autonomous. Defenders of the autonomy of software agents argue that they are capable of performing actions without being explicitly programmed to do what they do. Yet the problem is that ultimately it is the programmer or the user who evaluates what the CAs do. Hence, it is not clear at all that they could be considered as truly autonomous. On the other hand, there is the objection that CAs depend on a hardware structure that itself is neither self-produced nor self-maintained: the computer as a whole is a fully designed, built and maintained artifact.

That is why most of the research in the domain of autonomous agency is focused on the study of systems displaying actions in virtue of their physical properties. In these lines, there is nowadays an increasing acknowledgement of the role played by the so-called "active matter" (i.e., macromolecular structures capable of self-assembling) in the explanation (and artificial design) of autonomous agents. This is the case for most of the following examples.

3.2 Self-Propelling Oil Droplets

Self-propelling oil droplets [18] consist of a combination of chemical reactions, self-assembly processes and convective phenomena that together trigger the spontaneous global movement of an oily system in an aquatic environment. There are a great variety of self-propelled droplets, but most of them move during a certain time without an external source of energy (the driving mechanism is a Marangoni flow due to gradients in the interfacial energy on the droplet interface [19]). Thus, their movement does not depend on their obtaining "nutrients" (matter and energy) from the environment. Rather, they move by consuming their already available internal oleic anhydride (Figure 3).

According to the researchers that synthesized these systems, self-propelling oil droplets can be seen as "models of autonomy and minimal cognition based on physicochemical principles" [20]. Self-propelling droplets certainly do move by themselves, and one could argue that their self-sustained movement (although not the droplets themselves) can be considered self-maintaining in a minimal sense, insofar as the movement is co-driven by internal processes and gradients and exhibits a certain degree of robustness. However, these systems lack an internal organization (there is no internal organizational differentiation, nor any modular units) that enables them to display an inherent capacity to modify the conditions at the system-environment interface. Because of this simplicity, the droplet's movement does not lead the system to a different environment that, eventually, contributes to its maintenance.

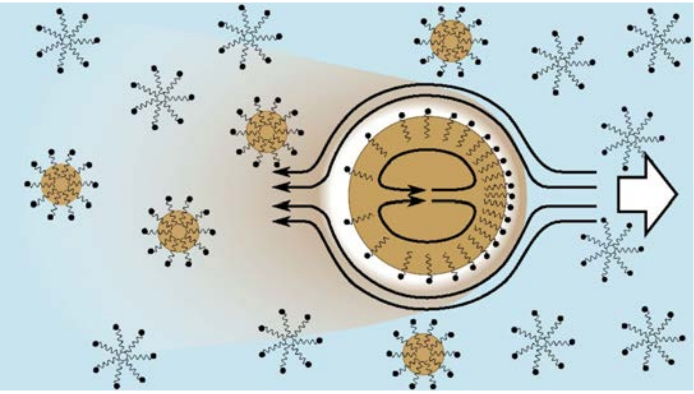

Figure 3. Schematic representation of the proposed mechanism for self-propelled motion of oil droplets. Reproduced from [21]. More details and a discussion of the mechanism can be found in the original publication.

Nevertheless, some of these systems (see, for example [22]) are steps on the path to constructing more interesting synthetic molecular machines, which we analyze next.

3.3 Autonomous Molecular Machines

As mentioned earlier, much more complex molecular devices (e.g., nano-robots) have been developed within a similar research direction [23, 24]. A molecular robot is an artificial molecular device capable of performing a variety of tasks. Molecular robots can be programmed to move and build molecular cargoes or build other molecules, for example, using a tiny robotic arm capable of manipulating a single molecule [25]. Molecular robots are essentially self-assembling macromolecular structures synthetically shaped and located so as to affect the relative motion of other component parts, so that together, they harness the energy flow in a desired way. It is the *interlocked architecture* of the components (i.e., their design and their structural codependence) that permits the overall system to transform an energy input into work performing a desired function. A molecular robot is usually built using mechanically interlocked molecular architectures (i.e., molecules linked according to their shape, such as rotaxenes and catenanes).

The basic principle underlying the construction of artificial molecular robots is the control of noncovalent interactions that govern the relative mechanical movement of the building blocks so as to create a *functionally integrated structure* able to perform work, transport cargoes or signal molecules through molecular shuttles, and so on [26].[8] There are different examples of "agents," made up of one or several (macro)molecules. For example: a robotic ribosome that assembles peptides [27], a four-wheeled molecule that moves on a metal surface [28] and a molecular propeller that pumps water by means of its hydrophobic surface [29].

These systems are actually much more complex and perform more interesting tasks than self-propelling droplets. Moreover, they may be designed to serve a function similar to that of their biological counterpart, thus fully fitting into a biological organization. And yet, the main conceptual problem with these devices is that they are really nothing more than parts of a larger system (the cell), and the actions they perform are functional for that larger system, not for themselves. Compared with viruses, for example, they are capable of performing more interesting and complex tasks, but on the other hand, they are also less autonomous.

Although considerably complex, these systems, by themselves, can only trigger purposeless processes. Indeed, these—eventually complex—processes could become functional; but this happens only as long as they participate in the larger organization of a cell. Thus, the point is whether something less complex than a cell could provide the requirements for functionality and purpose. As we will see next, this point will drive us to the right research line.

3.4 Replicators, Viruses and the Like

Here we will discuss two types of candidates for MA. First, we will consider the case of some hypothetical prebiotic entities, namely nude replicators; and second, we will analyze the case of present-day parasitic replicators (such as viruses, plasmids and prions). Both are self-replicating structures, but whereas the former are assumed to replicate by themselves, the latter can only do so within their infected hosts. On the other hand, the former can only replicate, whereas the latter may perform many other functional activities.

3.4.1 Nude Replicators

Several authors consider nude replicators, that is, molecules capable of replicating by themselves, without the aid of a metabolic (or at least a proto-metabolic) organization,[9] to be agents capable of "performing actions." Insofar as these hypothetical prebiotic entities would have been able to select the "building blocks" required for their own continuous replication from the molecular environment, some authors such as Dawkins [30] or Pross [31] often talk about them as if they were agents.[10] However, it was the philosopher Dennett [32] who explicitly claimed that a minimal agent is actually a Darwinian system (which he understood as a set of self-replicating molecules), blindly generated by natural selection and possessing different hardwired phenotypes.

Yet this claim is difficult to maintain. First, a self-replicating structure cannot be considered an agent, because agency requires a constitutive organization that includes different classes of functions [33]. And second, although it may be possible that (in a prebiotic scenario) nude replicators could replicate without using the machinery of proto-metabolic systems, nude replicators could hardly do a great deal. In the absence of a metabolic organization, a replicator would have lacked any functional activity other than that of self-replicating. As the famous Spiegelmann experiment [34] has shown, when a virus was able to replicate in the absence of the machinery of a cell, it suffered an evolution toward an extremely short chain, losing the sequences that codified for proteins or catalytically active RNAs, which were *functionally relevant only in the context of the invaded cell,* thus shrinking into a minimal degree of complexity. In other words, "involving" rather than evolving, the nude replicator would have become the exact opposite of what we usually mean by a Darwinian system.

3.4.2 Viruses and the Like

The case of viruses, plasmids and prions is different. Unlike synthetic molecular robots, these entities are the result of a natural—that is, not designed—process; but, more interestingly, they act "on their own behalf," in the sense that they autonomously pursue their own replication. Although they need the host machinery, they are able to penetrate the membrane of their prey. Once within the cell, they are able to deploy self-assembling and self-replicating capacities, which means that they have the capacity to modify their environment (i.e., the organization of their hosts) so as to generate copies of themselves. Thus, they are able to "autonomously" evolve, not in the sense that they can evolve *by* themselves (they need the organization of the host), but rather in the sense of evolving *for* themselves: they maximize their own fitness, not that of the host.

Moreover, in addition to being replicators, these systems (especially viruses) also perform other "functions" when infecting living cells. We are faced here with a very interesting case. On the one hand, these entities can perform functional activities, insofar as they become embedded in the autonomous organization of the host; yet on the other, they usually deploy activities that are functional not for the host, but rather for the infectious "agent." How can an invading entity deploy its own functional domain? Paradoxically, by becoming, in some way, a part (usually a pathological one) of the host organization, modifying it for its own benefit, and in so doing, deploying new functions (such as the capacity to compensate for the obstacles deployed by the host). Viruses infect the host cell and force it to fabricate copies of the invading entity, while prions act as a template that guides the unfolding of more proteins into prion form. In turn, these newly formed prions go on to convert more proteins, thus triggering a chain reaction that produces new copies of the original prion. But viruses, in addition, are quite complex molecular systems, constituted by several catalytically active macromolecules whose secondary structures are organized in rigid parts and may display relative movements, thus generating different interactive effects. These effects may include small displacements or other mechanical effects that are often described using a list of terms borrowed from the description of machines: lever and spring, ratchet and clamp, and others. In some cases, the interactions are triggered by the potential energy of the virus itself. For example, viruses attach to the host membrane and, traversing the host cell wall, inject their DNA inside the cell. In other cases, the interaction also requires some external supply of energy, which is provided by the machinery of the infected cell. Due to these interactive capacities and because, apparently, they "act" on their own behalf, viruses (and prions) can appear as minimal agents.

Ultimately, however, these entities cannot be considered examples of true MA. They are indeed much simpler than cells, yet their activity cannot be deployed except within the more complex organization of the cells: they do not possess a constitutive organization that allows them to deploy, by themselves, functional active behavior. Their activity requires, at the very least, a genetically instructed metabolic organization (see note 12), which is far more complex than that of either the virus or the prion.

3.5 Artificial Protocells as Hypothetical Proto-organisms

Other possible candidates for minimal agency are natural or artificial protocells. By this I mean an encapsulated *proto-metabolic* system that is organizationally much simpler than present-day prokaryotic cells. Strictly speaking, a protocell is any experimental or theoretical model that involves a self-assembling compartment linked to chemical processes taking place around or within it. Protocells are used as compartmentalized systems showing some lifelike properties, such as growth, autocatalytic activities or reproduction [35]. Although some of these systems are *in vitro* experimental models, here I will refer only to computational models of hypothetical prebiotic organisms. These models aim to explain how more complex biological cells or alternative forms of cellular organization may have come about. Figure 4 shows one model of a protocell that "does" some activity in its environment [36].

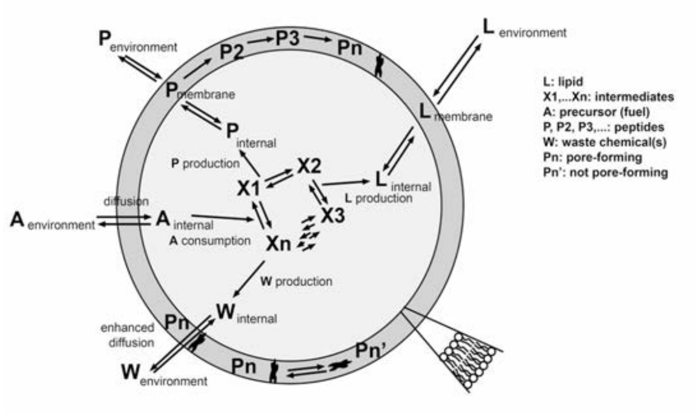

Figure 4. A model of a prebiotically plausible protocell whose membrane consists of both lipids and small peptides. The model simulates real membrane processes coupled to chemical autocatalytic reactions. Redrawn in simplified way after [37]. More details and discussion on the mechanism can be found in the original publication.

Ultimately, it is difficult to assess whether a protocell is an example of MA, because there is no precise delimitation of the organizational complexity of these types of systems (aside from saying that they are less complex than any present-day cell). For example, a system that modulates its constitutive organization in accordance with its internal needs but without modifying the external environment (e.g., pumping ions to avoid an osmotic crisis)[11] is not a true agent. Nor does a slow modification of the environment that may emerge as functional only at a phylogenetic timescale count as agency. At most, we could classify some of these processes as forms of "proto-agency." As discussed in Section 2, being an agent requires an offline regulatory mechanism, which is unlikely without a genetically based metabolic organization.[12] Most synthetic protocells are not able to deploy regulated adaptive responses (protocells respond only through feedback to the external changes). In fact, what these systems actually do is to functionally modify their own constitutive organization (e.g., pumping ions to avoid an osmotic crisis, as in our previous example).

4. Conclusion: What Is Minimal Agency?

As we have seen, the term minimal agency (MA) encompasses many different approaches. Globally, however, the concept of MA may be understood in two different ways: (1) as the simplest agent within the known phenomenological domain, that is, the simplest agent as-we-know-it; and (2) as the simplest agent that could exist, either synthetically or in a hypothetical process of biogenesis, that is, the simplest agent as-it-should-be. For example, when authors such as Burge argue that unicellular eukaryotes are the simplest example of agency, they are interpreting MA in the first sense. And when other scholars such as Ikegami or Hanczyc argue that their self-propelling oil droplets are indeed agents, they have in mind the second definition of the term.

The second view is more radical, since it focuses on the simplest organizational and material conditions required for generating such capacity, not on how it is in fact minimally instantiated. In other words, it searches for the simplest material building blocks that, either naturally or artificially, could achieve the simplest form of organization necessary to display what we mean by agency. Thus, although the concept of minimality is usually understood as the minimal organization (i.e., the minimal type of network, the minimal number of functions, and so on) *ceteris paribus*, a system with fewer components is simpler than another—functionally equivalent—but with a larger number of similar components. From this perspective, the question of MA should be formulated as if we were fabricating an agent from scratch. Moreover, this perspective of MA leads us to a molecular scenario,[13] moving very close to the synthetic-engineering and analytical-biological views.

On the other hand, the first perspective reveals some interesting points, which may serve to clarify certain puzzling aspects of MA. Indeed, one may think that since our interest is *minimal* agency, the history of life is pointless, because it can only show how agency has become more complex and diverse. Yet biology may also reveal to us intricate processes of agency minimization, along with a host of other borderline cases. Since biological agents can only exist as the result of: (1) a long and cumulative process of entailments covering a huge number of individuated lifespans; and (2) being embedded in a synchronic, spatially larger network (including also a huge number of individuated systems) of metabolic complementarities (i.e., ecosystems), they exist in an entangled multidimensional organization, with very different types of codependent systems. This makes clarifying the concept of MA even more complex. For example, the fact of being embedded in evolutionary history prompts an apparent form of MA: proto-organisms may have slowly modified their environment in a functional sense, yet since this slow change is functional only at a phylogenetic scale, this process of environmental modification can hardly be considered a form of agency attributable to some proto-organism. If we want to determine *what* is the source of this "action," we should focus on a given population during a specific evolutionary time period, when it is in fact the organization of each individuated system that ultimately explains the change in the environment. Even more puzzling is the case of certain forms of "minimal" agency derived from synchronic entanglements. Since biological agents form intimate associations, they share many organizational mechanisms, also forming complex nested hierarchical organizations, in which many functions of the individuated parts are transferred to the higher collective level. These facts often lead to an ultra-simplification of certain agents (e.g., endosymbionts).

Thus, in the biological framework, some systems may emerge as extremely MAs, sometimes even operating on their own behalf, and yet, paradoxically, requiring much more complex forms of organization in order to act (viruses, for example). In other cases, the appearance of an organized set of agents displaying a collective form of agency raises the question of whether the source of the agency is indeed a new integrated agent (such as a multicellular animal), or simply a coordinated sum of individual actions (such as a swarm).[14] In the first case, the action of the collection of agents becomes a constitutive process. But a set of individual agents could also lead to the formation of a new integrated entity devoid of agential power (e.g., an ecosystem). In sum, the biological domain is full of counterintuitive examples that challenge the attempt to define MA: on the one hand, a host of extremely simplified entities that apparently behave as agents (e.g., viruses, prions, plasmids), and on the other, much more complex systems that cannot so easily be classed as agents (e.g., ecosystems, colonies and certain plants).

How can we understand these paradoxes? And what lessons can be learned for the synthetic fabrication of agents? The first lesson is that the *natural* process for generating agency has been a long and arduous one, involving many temporally entailed steps and a huge number of synchronic systems. The second lesson is that this scenario necessarily generates a wide variety of entities, some increasingly complex and others specialized and simplified. The third lesson is that in turn, these different systems constitute a complex web of nested organizations and dependencies between very different types of entities. And yet, what I have tried to show is that this complex, multilevel and multidimensional phenomenology pivots around an organizational core, which I have described in Section 2. The message, then, is that a lower form of organization will not allow the emergence of maximally simple, yet derivative, forms of agency.

Rightly, research into synthetic biology and bioengineering is focused on the bottom-up design and fabrication of specific types of systems, in the hope that they will one day be capable of behaving like the simplest biological agents. Yet the fact that this research deals essentially with isolated systems masks our understanding of the aforementioned paradoxes. Thus, it is important to avoid an overly loose characterization of agency that could lead to counterintuitive cases. In such a complicated scenario, the requirement of autonomy, as well as all the requirements discussed in Section 2, is of paramount importance for understanding what exactly a MA is. This may place the organizational requirements for MA at a relatively high level of complexity, at least in terms of artificial engineering standards. However, they are necessary if we want to avoid the paradoxical cases of MA that exist in the biological domain.

Acknowledgments

The author acknowledges the Research Projects of the Basque Government IT 590-13, and of MINECO FFI2014-52173-P, as well as a Salvador de Madariaga Fellowship PRX17/00379. The author also thanks the IHPST (Paris) for hosting his research stay during the first semester of 2018. Also, the author wants to acknowledge the collaboration of Pasquale Stano and Luisa Damiano for their helpful comments on the revision of this paper and suggestions about the inclusion of the figures.

Notes

1. A clarification is required here. An interactive process is, as we have pointed out, a functional modification, exerted by an autonomous system, of the boundary conditions of said system. This implies that the system is already defined and that this is done by its constitutive processes. Yet, since an autonomous system is, by definition, capable of functionally *modulating* these constitutive processes, this modulatory activity, insofar as it does not affect environmental conditions, cannot be considered an alteration of the system's boundary conditions. For example, ion pumping by the cellular membrane in order to reduce ion concentration is a functional modulation of the basic metabolic processes of the cell, but it does not affect the system's boundary conditions. On the other hand, the secretory activity of a bacterium in a biofilm, which affects the density of its neighboring bacteria, does.
2. In this sense, Deamer [38] has proposed that a primitive form of adaptivity in prebiotic systems may have consisted of a negative feedback

mechanism that controlled the transport of molecules across the membrane boundary, which in turn controlled the overall process of growth. As he points out: "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."

3. This term is taken from Ashby [39]. According to this author, from a dynamic point of view, we can define a set of boundary conditions and an essential parameter value region necessary for the maintenance of a far-from-equilibrium system. We call these parameters and boundary conditions "essential variables," and the range within which the system's organization can be maintained "viability boundaries" [10]. A more detailed discussion of how this concept is related to the anticipation in agency can be found in DiPaolo [9] and Barandiaran and Egbert [40].

4. What I mean by this is that an agent is in a better situation at instant T_2 than at the previous instant T_1, if at T_1 it was closer to the limits of its boundary of viability than at T_2.

5. Actually, DiPaolo formulated this idea not to define agency, but rather as a definition of adaptivity. We use it here to characterize regulated adaptivity, which, in our account, while not equivalent to agency, is nevertheless a requirement of it.

6. Strictly speaking, this implies the existence of an embodied normativity (regulatory normativity), which of course, is of human origin in an artificial device. In a natural system, however, this norm could not be originated in the individuated agent, and it is in fact at a much longer timescale (phylogenetic time) that higher-level norms are originated. In a spatially and temporally wide scenario, certain interactive patterns that contributed to the survival of the organizations where they were implemented are selected. Thus, from an organizational perspective, these norms are also functions, because they also ultimately achieve a causal loop.

7. A very basic case of agency is bacterial chemotaxis. For example, in *E. coli*, the directional search for nutrients requires that the bacterium measure the temporal difference of attractant concentrations in its

environment and change the frequency of its flagellar rotation accordingly. This operation is achieved thanks to the so-called two-component signal transduction subsystem, which acts as a memory and inner connection between "sensors" and "effectors" [41]. The interaction is functional because the transformations induced in the environment contribute to the self-maintenance of the bacterium: the concentration of sugar increases in relation to the system, and the system operates recursively on these environmental changes (sugar detection and frequency of flagellar rotation are correlated).

8. One example is provided by DNA nanotechnology, which combines rotaxanes, catenanes and related structures to create interlocked DNA structures that can be generated from both double-stranded and single-stranded DNA [26, 42, 43].
9. Dawkins proposed the idea of the "replicator," defined as "anything in the universe of which copies are made" [27, p. 83]. In the context of the origin of life, "replicators" are hypothetical molecules that first managed to reproduce themselves and thus gained an advantage over other molecules within the primordial soup.
10. For Pross, replicators are "purposeful" entities. Yet, to be fair, he does not claim that a nude molecular replicator is already an agent, only that replication is the basis of the teleological nature of life.
11. This is the case, for example, of Ganti's chemoton [44] and of Ruiz-Mirazo and Mavelli's protocell model [37, 45].
12. A detailed argument showing why a dynamically decoupled subsystem is necessary to implement a genuine regulatory capacity can be found in [11, 46], and why the appearance of such a mechanism is likely related with the appearance of genetically instructed metabolisms in [47].
13. Suppose, for example, that somebody builds a macroscopic robot that instantiates the simplest organizational structure that permits agency to be deployed. Even so, one could still argue that this is not a true MA, since it would always be possible to build a replica with fewer molecules: a macroscopic arm, for example, is a rigid structure in which billions of molecules redundantly perform the same function. Hence, from this perspective, one could argue that a MA would be a molecular version of such a robot. One example that illustrates this is a molecular robot built recently by a group of researchers that is endowed with an arm made up of just 150 carbon, hydrogen, oxygen and nitrogen atoms [25].
14. For a discussion on this subject, see [48].

References

[1] P. Maes, "Modeling Adaptive Autonomous Agents," *Artificial Life*, **1**(1–2), 1994 pp. 135–162. doi:10.1162/artl.1993.1.135.

[2] T. Smithers, "Autonomy in Robots and Other Agents," *Brain and Cognition*, **34**(1), 1997 pp. 88–106. doi:10.1006/brcg.1997.0908.

[3] S. Franklin and A. Graesser, "Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents," in *Intelligent Agents III Agent Theories, Architectures, and Languages (ATAL 1996)*, (J. P. Müller, M. J. Wooldridge and N. R. Jennings, eds.), Berlin, Heidelberg: Springer, 1997. doi:10.1007/BFb0013570.

[4] S. A. Kauffman, *Investigations*, New York: Oxford University Press, 2000.

[5] W. D. Christensen and C. A. Hooker, "Autonomy and the Emergence of Intelligence: Organised Interactive Construction," *Communication and Cognition-Artificial Intelligence*, **17**(3–4), 2000 pp. 133–157.

[6] X. E. Barandiaran, E. Di Paolo and M. Rohde, "Defining Agency: Individuality, Normativity, Asymmetry and Spatio-temporality in Action," *Adaptive Behavior*, **17**(5), 2009 pp. 367–386. doi:10.1177/1059712309343819.

[7] A. Moreno and A. Etxeberria, "Agency in Natural and Artificial Systems," *Artificial Life*, **11**(1–2), 2005 pp. 161–175. doi:10.1162/1064546053278919.

[8] A. Moreno, A. Etxeberria and J. Umerez, "The Autonomy of Biological Individuals and Artificial Models," *BioSystems*, **91**(2), 2008 pp. 309–319. doi:10.1016/j.biosystems.2007.05.009.

[9] E. A. Di Paolo, "Autopoiesis, Adaptivity, Teleology, Agency," *Phenomenology and the Cognitive Sciences*, **4**(4), 2005 pp. 429–452. doi:10.1007/s11097-005-9002-y.

[10] X. Barandiaran and A. Moreno, "Adaptivity: From Metabolism to Behavior," *Adaptive Behavior*, **16**(5), 2008 pp. 325–344. doi:10.1177/1059712308093868.

[11] L. Bich, M. Mossio, K. Ruiz-Mirazo and A. Moreno, "Biological Regulation: Controlling the System from Within," *Biology & Philosophy*, **31**(2), 2016 pp. 237–265. doi:10.1007/s10539-015-9497-8.

[12] L. Bich and A. Moreno, "The Role of Regulation in the Origin and Synthetic Modelling of Minimal Cognition," *BioSystems*, **148**, 2016 pp. 12–21. doi:10.1016/j.biosystems.2015.08.002.

[13] H. G. Frankfurt, "The Problem of Action," *American Philosophical Quarterly*, **15**(2), 1978 pp. 157–162. www.jstor.org/stable/20009708.

[14] T. Burge, "Primitive Agency and Natural Norms," *Philosophy and Phenomenological Research*, **79**(2), 2009 pp. 251–278. www.jstor.org/stable/40380466.

[15] S. Wolfram, *A New Kind of Science*, Champaign, IL: Wolfram Media, Inc., 2002.

[16] C. G. Langton, “Studying Artificial Life with Cellular Automata,” *Physica D: Nonlinear Phenomena*, **22**(1–3), 1986 pp. 120–149. doi:10.1016/0167-2789(86)90237-X.

[17] C. G. Langton, (ed.), *Artificial Life: An Overview*, Cambridge, MA: MIT Press, 1995.

[18] M. M. Hanczyc and T. Ikegami, “Chemical Basis for Minimal Cognition,” *Artificial Life*, **16**(3), 2010 pp. 233–243. doi:10.1162/artl_a_ 00002.

[19] R. Seemann, J.-B. Fleury and C. C. Maass, “Self-Propelled Droplets,” *The European Physical Journal Special Topics*, **225**(11–12), 2016 pp. 2227–2240. doi:10.1140/epjst/e2016-60061-7.

[20] N. Horibe, M. M. Hanczyc and T. Ikegami, “Mode Switching and Collective Behavior in Chemical Oil Droplets,” *Entropy*, **13**(3), 2011 pp. 709–719. doi:10.3390/e13030709.

[21] C. Jin, C. Krüger and C. C. Maass, “Chemotaxis and Autochemotaxis of Self-Propelling Droplet Swimmers,” *Proceedings of the National Academy of Sciences of the United States of America*, **114**(20), 2017 pp. 5089–5094. doi:10.1073/pnas.1619783114.

[22] T. Ikegami, N. Horibe and M. M. Hanczyc, “Potential Memory Effects in Self-Moving Oil Droplets,” *International Journal of Unconventional Computing*, **11**(5–6), 2016, pp. 345–355.

[23] S. Seangupta, M. E. Ibele and A. Sen, “Fantastic Voyage: Designing Self-Powered Nanorobots,” *Angewandte Chemie International Edition*, **51**(34), 2012 pp. 8434–8445. doi:10.1002/anie.201202044.

[24] I. Lagzi, “Chemical Robotics-Chemotactic Drug Carriers,” *Central European Journal of Medicine*, **8**(4), 2013 pp. 377–382. doi:10.2478/s11536-012-0130-9.

[25] S. Kassem, A. T. L. Lee, D. A. Leigh, V. Marcos, L. I. Palmer and S. Pisano, “Stereodivergent Synthesis with a Programmable Molecular Machine,” *Nature*, **549**, 2017 pp. 374–378. doi:10.1038/nature23677.

[26] J. Valero, F. Lohmann and M. Famulok, “Interlocked DNA Topologies for Nanotechnology,” *Current Opinion in Biotechnology*, **48**, 2017 pp. 159–167. doi:10.1016/j.copbio.2017.04.002.

[27] B. Lewandowski, G. De Boi, J. W. Ward, M. Papmeyer, S. Kuschel, M. J. Aldegunde, P. M. E. Gramlich, et al., “Sequence-Specific Peptide Synthesis by an Artificial Small-Molecule Machine,” *Science*, **339**(6116), 2013 pp. 189–193. doi:10.1126/science.1229753.

[28] T. Kudernac, N. Ruangsupapichat, M. Parschau, B. Maciá, N. Katsonis, S. R. Harutyunyan, K.-H. Ernst and B. L. Feringa, “Electrically Driven Directional Motion of a Four-Wheeled Molecule on a Metal Surface,” *Nature* **479**, 2011 pp. 208–211. doi:10.1038/nature10587.

[29] B. Wang and P. Král, “Chemically Tunable Nanoscale Propellers of Liquids,” *Physical Review Letters*, **98**(26), 2007 266102. doi:10.1103/PhysRevLett.98.266102.

[30] R. Dawkins, *The Extended Phenotype: The Long Reach of the Gene*, New York: Oxford University Press, 1983.

[31] A. Pross, "The Driving Force for Life's Emergence: Kinetic and Thermodynamic Considerations," *Journal of Theoretical Biology*, **220**(3), 2003 pp. 393–406. doi:10.1006/jtbi.2003.3178.

[32] D. C. Dennett, *Kinds of Minds: Toward an Understanding of Consciousness*, New York: Basic Books, 1996 Chapter 2.

[33] A. Moreno and M. Mossio, *Biological Autonomy: A Philosophical and Theoretical Enquiry*, Dordrecht: Springer, 2015 Chapter 5.

[34] D. L. Kacian, D. R. Mills, F. R. Kramer and S. Spiegelman, "A Replicating RNA Molecule Suitable for a Detailed Analysis of Extracellular Evolution and Replication," *Proceedings of the National Academy of Sciences of the United States of America*, **69**(10), 1972 pp. 3038–3042. www.ncbi.nlm.nih.gov/pmc/articles/PMC389702.

[35] S. Rasmussen, M. A. Bedau, L. Chen, D. Deamer, D. C. Krakauer, N. H. Packard and P. F. Stadler, eds., *Protocells: Bridging Nonliving and Living Matter*, Cambridge MA: MIT Press, 2009.

[36] K. Ruiz-Mirazo, "Protocell," *Encyclopedia of Astrobiology*, Vol. 3 (M. Gargaud, R. Amils, J. Cernicharo Quintanilla, H. J. Cleaves, W. M. Irvine, D. Pinti and M. Viso, eds.), Berlin, Germany: Springer, 2011 pp. 1353–1354.

[37] K. Ruiz-Mirazo and F. Mavelli, "Simulation Model for Functionalized Vesicles: Lipid-Peptide Integration in Minimal Protocells," in *European Conference on Artificial Life (ECAL 2007)*, Lisbon, Portugal, September 10–14, 2007 (F. Almeida e Costa, L. M. Rocha, E. Costa, I. Harvey and A. Coutinho, eds.), Berlin, Heidelberg: Springer, 2007 pp. 32–41. doi:10.1007/978-3-540-74913-4_ 4.

[38] D. W. Deamer, "On the Origin of Systems: Systems Biology, Synthetic Biology and the Origin of Life," *EMBO Reports*, **10**(1S), 2009 pp. S1–S4. doi:10.1038/embor.2009.117.

[39] W. R. Ashby, *Design for a Brain: The Origin of Adaptive Behavior*, London: Chapman & Hall, 1952.

[40] X. E. Barandiaran and M. D. Egbert, "Norm-Establishing and Norm-Following in Autonomous Agency," *Artificial Life*, **20**(1), 2014 pp. 5–28. doi:10.1162/ARTL_a_00094.

[41] M. van Duijn, F. A. Keijzer and D. Franken, "Principles of Minimal Cognition: Casting Cognition as Sensorimotor Coordination," *Adaptive Behavior*, **14**(2), 2006 pp. 157–170. doi:10.1177/105971230601400207.

[42] D. Ackerman, T. L. Schmidt, J. S. Hannam, C. S. Purohit, A. Heckel and M. Famulok, "A Double-Stranded DNA Rotaxane," *Nature Nanotechnology*, **5**, 2010 pp. 436–442. doi:10.1038/nnano.2010.65.

[43] S.-S. Jester and M. Famulok, “Mechanically Interlocked DNA Nanostructures for Functional Devices,” *Accounts of Chemical Research*, **47**(6), 2014 pp. 1700–1709. doi:10.1021/ar400321h.

[44] T. Gánti, *The Principles of Life*, New York: Oxford University Press, 2003.

[45] K. Ruiz-Mirazo and F. Mavelli, “On the Way towards ‘Basic Autonomous Agents’: Stochastic Simulations of Minimal Lipid-Peptide Cells,” *BioSystems*, **91**(2), 2008 pp. 374–387. doi:10.1016/j.biosystems.2007.05.013.

[46] W. Bechtel, “Biological Mechanisms: Organized to Maintain Autonomy,” *Systems Biology: Philosophical Foundations* (F. C. Boogerd, F. J. Bruggeman, J.-H. S. Hofmeyr and H. V. Westerhoff, eds.), Amsterdam: Elsevier, 2007 pp. 269–302. doi:10.1016/B978-044452085-2/50014-0.

[47] J. Griesemer and E. Szathmáry, “Ganti’s Chemoton Model and Life Criteria,” *Protocells: Bridging Nonliving and Living Matter* (S. Rasmussen, M. A. Bedau, L. Chen, D. Deamer, D. C. Krakauer, N. H. Packard and P. F. Stadler, eds.), Cambridge: MIT Press, 2009. doi:10.7551/mitpress/9780262182683.001.0001.

[48] A. Arnellos and A. Moreno, “Multicellular Agency: An Organizational View,” *Biology & Philosophy*, **30**(3), 2015 pp. 333–357. doi:10.1007/s10539-015-9484-0.