Special Issue, "What AI Can Learn from Biology"

Vol. 8, No. 1–2 (2025)
ISSN: 2532-5876

DOI: 10.13133/2532-5876/19210

Could Machines Develop Autonomous Agency?

Ana M. Soto[a*] *and Carlos Sonnenschein*[a*]

[a] *Tufts University School of Medicine, Boston, USA and Centre Cavaillès, École Normale Supérieure, Paris, France*

***Corresponding author:** Ana M. Soto, Email: ana.soto@tufts.edu

Abstract

"Could machines develop autonomous agency?" To address this question, we explored the recent return of the concept of agency in biological discourse. At the end of the 19th century, the successful development of physics and chemistry motivated some biologists to adopt a physicalist stance, positing that biology can be reduced to physics and chemistry. This theoretical approach became dominant during the 20th century with the advent of molecular biology while teleology, agency and normativity disappeared from the biological lexicon. The failure of molecular biology to explain complex biological organization probably led to the reintroduction of these concepts in the biological sciences and philosophy of biology. In addition to the historicity of organisms (they are the product of organismal reproduction throughout phylogenesis), the intrinsic properties of biological objects are linked to the precariousness of life as exemplified by the need to search for food and to avoid being eaten. Moreover, the continuous need to counteract entropy also involves the capacity of organisms to synthesize their own chemical components and reproduce. From this historical narrative, we conclude that it is unlikely that machines could develop minimal intrinsic agency. On the contrary, when they appear to express agency, it is of external origin, reflecting the agency of the humans that created such machines.

Keywords: teleology, historicity, goal-directedness, natural agency, artificial agency

Citation: Soto, AM, & Sonnenschein, C 2025, "Could Machines Develop Autonomous Agency?", *Organisms: Journal of Biological Sciences*, vol. 8, no. 1–2, pp. 33–38. DOI: 10.13133/2532-5876/19210

Introduction

We commend Ali Hossaini for having brought the issue of agency to the Artificial Intelligence (AI) agenda, and with it, the question: Could machines and artifacts created by humans, like AI, have true agency? Before answering this question, we should state that organisms are agents: that is to say, they have the capacity to generate action. The agency of organisms is a major distinction between the living and the inert. Organisms are also normative, that is to say, they have the capacity to generate their own rules. Different disciplines have different ways of conceptualizing agency. For example, in cognitive science, agency in humans is seen in the context of consciousness, beliefs and reason, while some philosophers and biologists study agency in the context of the purposiveness of unicellular organisms (Moreno 2018), in the context of the evolution of consciousness (Walsh 2015) and still other mental phenomena (Moreno 2023). Because we are examining whether machines could be agents, we will use definitions that apply to a minimal autonomous agent. According to Alvaro Moreno, "a system is autonomous if it actively maintains its identity: for example, by modulating its internal, constitutive organization..." However, maintaining its self-organization is not enough for considering such a system agential. An autonomous agent must also act upon the external environment, modifying the latter to the system's benefit. Thus, agency has an interactive dimension. Consequently, an autonomous system could be defined as "a system doing something by itself according to its own goals or norms within a specific environment" (Barandiaran *et al.* 2009). In this way, we bring together autonomy, agency and normativity because these are closely related terms. This definition of agent easily suggests that we are referring to living objects. In contrast, it is difficult to determine whether the apparent agency of artificial devices is just a mere extension of the agency of the people who created them. Thus, it is reasonable to inquire about the strong links between agency and the alive. In particular, how is minimal agency instantiated in biology, in order to best evaluate whether such minimal agency could also be instantiated by AI.

Before the 20th century, agency was considered a defining property of biological entities; during the 20th century, radical changes occurred regarding the conceptualization of biological phenomena. For example, the philosopher Lenny Moss described a radical change regarding the perception of the organism. In his own words, this represents a change

> ... between a theory of life which locates the agency for the acquisition of adapted form in ontogeny—that is, in some theory of epigenesis versus a view that expels all manner of adaptive agency from within the organism and relocates it in an external force—or as Daniel Dennett (1995) prefers to say, an algorithm called 'natural selection' (Moss 2003).

Additional conceptual changes imposed by the molecular biology revolution and the modern evolutionary synthesis hindered the study of agency and its companion, normativity, because teleology (goal-directedness) was incompatible with the dominant mechanicist view among biologists (Soto & Sonnenschein 2018). Teleology is defined as the explanation of phenomena in terms of the purpose they serve rather than of the cause by which they arise. Organisms exhibit goal-directed behaviors, for example, to maintain themselves alive. Biologists describe organs by their purpose (the heart to pump blood; the intestine to absorb nutrients).

After removing teleology from the biological lexicon, cells and organisms became passive recipients of a program (Longo *et al.* 2012). Because of these changes, agency, normativity and individuation, until then considered the main characteristics of the living, almost disappeared from biological language. This absence is now being contested by organicists; they favor reinstating agency where it belongs, into the organism (Walsh 2015; Soto & Sonnenschein 2023). This movement generated a renewed interest in agency and its practically non-dissociable companion, normativity (Moreno 2018).

In the natural world, only biological entities display agency, normativity and goal-directedness. This is why we need to delve into biological theory and philosophy to understand whether agency is inextricably linked exclusively to organisms or, alternatively, whether it can also be attributed to machines and other artifacts created by humans. In this regard, we need to look into some properties of biological objects (organisms) that make them different from physical objects and machines; these properties include intrinsic goal-directedness (which originates internally, like the

organism's goal of keeping itself alive), autonomy and historicity. Self-organizing systems like flames are 'a-historical' because they appear spontaneously and can be analyzed independently. In contrast, organisms are not spontaneous but historical. This means that they are a consequence of the reproductive activity of a pre-existing organism. Organisms are historical in two contexts, ontogeny, meaning their history as individuals from conception to death, and phylogeny, which is the history of a taxonomic group (for example, a species) throughout evolution.

Objectively, organisms are different from computers; whereas in the latter software is independent of the hardware, in the former, function is inseparable from the material specific to the biological object (Longo & Soto 2016).

1. The Organicist Tradition: From Intrinsic Teleology to Autopoiesis and Autonomy

Unlike inert objects in the classical mechanics tradition, biological objects are always active. Since Aristotle and Kant, biological objects are characterized by their goal-directedness (teleology). Kant stressed the inter-relatedness of the organism and its parts and the circular causality implied by this relationship. Since the late 18th century, following Kant's ideas, teleology has been an extremely useful concept for the development of several biological disciplines (Lenoir 1982, Gambarotto 2014). However, the conceptual clarity of causal mechanics and its successes inspired biologists to adopt a physicalist reductionist stance and thus deny any special state to biological entities. As a result of this change in consensus, during the last two centuries, physicalism, reductionism and organicism co-existed.

Organicism has its philosophical basis in Aristotle's and Kant's conceptions of the organism and is a materialistic philosophical stance contrary to reductionism. It asserts that properties that could not have been predicted from the analysis of the lower levels appear at each level of biological organization. Therefore, explanations should address biological phenomena at all pertinent levels of organization. Also, implicit in this view is the idea that organisms are not just 'things' but objects in relentless change. Central to organicism are four concepts, namely, organization, historicity, organisms as normative agents, and biological specificity (organisms are individuals). Closely related to organization is the notion of 'organisational closure', which is a "distinct level of causation, operating in addition to physical laws, generated by the action of material structures acting as constraints" (Mossio & Moreno 2010). Finally, while objects in physics are generic and thus interchangeable, like rocks and planets, biological objects are specific – that is, they are individuals that are permanently undergoing individuation (Soto & Sonnenschein 2006).

Due to the increase in prestige of biochemistry in the mid-19th century and of molecular biology in the 20th, the idea that biology could be reduced to chemistry became dominant (Soto & Sonnenschein 2018). However, the advent of cybernetics in the 1940's stressing feedback systems and their circular causality produced tools that were applied both to artifacts and organisms. Additionally, the introduction of thermodynamics of dissipative systems provided an opportunity to examine the relevance of self-organizing physical systems to the understanding of biological systems. Both developments contributed to studies about the emergence of life, as exemplified by the pioneering work of Prigogine and his school (Nicolis & Prigogine 1977), of Kauffman's (Kauffman 1993), and that of Maturana and Varela (Maturana & Varela 1980) with their autopoiesis theory, to name just a few. These developments brought purposiveness back to biology and contributed to the revival of organicism.

Autopoiesis characterizes most of the fundamental features of biological objects. In particular, an autopoietic entity produces a physical boundary, which ensures a certain stability for the maintenance of the metabolic processes that generate the system's components, including their boundaries (Maturana & Varela 1980; Moreno & Mossio 2015). Such an autopoietic system is *autonomous* because it actively maintains its identity; i.e., it generates its own "law". In other words, it will respond to environmental fluctuations by regulating its constitutive organization; these actions safeguard the viability of the system. For a system to be alive, however, in addition to purposiveness, there is another component that differentiates it from the self-organization of physical systems which occur spontaneously such as flames and micelles. This notion is historicity (Cottrell 1979; Longo *et al.* 2015). Unlike flames and micelles,

organisms are produced by pre-existing organisms and they themselves produce a history.

2. Historicity

Stephen J. Gould was keenly aware of the contingency of evolutionary history as witnessed by his proposed metaphorical experiment of "replaying life's tape." In his own words,

> You press the rewind button and, making sure you thoroughly erase everything that actually happened, go back to any time and place in the past... Then let the tape run again and see if the repetition looks at all like the original (Gould 1990).

He anticipated that, "any replay of the tape would lead evolution down a pathway radically different from the road actually taken" (Gould 1990). This history and the contingency it implies also point to another important difference between physical (inert) objects and living objects, which is about the phase space. Physical objects are studied within a pre-given phase space. The phase space is the space of all possible states of a physical system. In classical mechanics, the phase space contains all possible positions of all the objects in the system and their momenta in order to determine the future behavior of that system. In contrast to physics, there is no pre-given phase space in biology. The phase space is created as novelty is being produced. For example, a swimming bladder provided an entirely new "phase space" for the bacteria that inhabit it (Longo, Montévil, & Kauffman 2012).

3. The Radical Materiality of the Living

Molecular biology brought the ideas of information, program and signal into biology. These ideas were borrowed from the rigorous mathematical theories of information (Longo *et al.* 2012, Soto & Sonnenschein 2020). This appropriation was metaphorical at best, rather than properly theoretical. In fact, these metaphors were interpreted as being real entities (Longo *et al.* 2012). Another consequence of this unfortunate development was that together with these ideas borrowed from mathematics and computer sciences came a duality, namely, the independence of software from hardware. However, life is based on the actual materials organisms are made from, from macromolecules such as DNA and proteins to membranes. There is no way to disassociate these materials from the functions organisms fulfill. In contrast, inert objects such as hammers could be made from different materials as long as the material does not prevent the intended function. This radical materiality of life rules out distinctions such as 'software vs. hardware', and thus is incompatible with theoretical transplants that do not take into consideration this material specificity (Longo & Soto 2016). Moreover, it also suggests that concepts such as agency, which are naturally instantiated in biological entities, are inevitably inseparable from their natural material substrate.

4.Minimal Biological Agency

In the organicist tradition, we recognize organisms as normative agents. This way of thinking was already implicit in the 18th and 19th century. For example, the biologist Xavier Bichat noticed that physical objects such as rocks or planets, do not get ill. He also remarked that "Whereas monsters are still living beings, there is no distinction between normal and pathological in physics and mechanics". "The distinction between the normal and the pathological holds for living beings alone" [cited by Canguilhem (Canguilhem 2008)]. And this remark about the normal and the pathological brings us specifically into normativity. According to Canguilhem, "life is not indifferent to the conditions in which it is possible, that life is polarity and thereby even an unconscious position of value; in short, life is in fact a normative activity." And, "...we do ask ourselves how normativity essential to human consciousness would be explained if it did not in some way exist in embryo in life." Furthermore,

> ...therapeutic need is a vital need, which, even in lower living organisms (with respect to vertebrate structure) arouses reactions of hedonic value or self-healing or self-restoring behaviors. The dynamic polarity of life and the normativity it expresses account for an epistemological fact of whose important significance Bichat was fully aware. Biological pathology exists but there is no physical or chemical or mechanical pathology. (Canguilhem, 1991).

The normativity of organisms is closely linked to their goal of actively keeping themselves alive (teleology). This function is accomplished by the mutual dependence among the different organs and between them and the whole organism. For example, the lung enables the organism to exchange gases by sending carbon dioxide to the external environment and taking in oxygen. The heart pumps blood transporting oxygen and nutrients to all cells of the organism. According to an organicist perspective, this interdependence is due to a causal regime technically referred to as the closure of constraints (Mossio *et al.* 2016, Montévil & Mossio 2020).

For a system to be an agent it needs to exert a causal effect on the environmental conditions of the system; this is an asymmetrical relationship because the organism imposes its norms on external entities. For example, an organism feeds on another organism in order to keep itself alive. This interactive dimension is the *sine-qua-non* of agency. Moreover, the agent needs to anticipate outcomes while choosing among options when reacting to changes in its environment. Furthermore, this ability to act towards a goal also includes the possibility of failing.

From what we discussed above, we posit that only cells, be they prokaryotes or eukaryotes, are able to express minimal agency. Viruses do not have a constitutive organization capable of generating a functionally active behavior by themselves even if in the end, by using a host cell, they can replicate (i.e., exhibiting a self-preserving goal). Overall, evolution has increased organismal complexity, but has also generated some adaptive simplifications and specializations; for example, ice fish without erythrocytes. Regarding agency, evolution has produced some counterintuitive cases; on the one hand, systems of great complexity, like ecosystems which are devoid of agency but contain agential organisms, and on the other hand, viruses, which deceptively show agency (although not a *bona-fide* one as explained above) but are not generally considered organisms.

Conclusions

Systems that instantiate biological agency are characterized by their organization, their autonomy, their historicity, their full dependency on the singularity and specificity of the materials they are made of, and on their complex and asymmetrical relationship with their environment to which they impose their norms. A salient characteristic of organisms is their sentience and precariousness; organisms must search for nutrients and avoid being eaten by other organisms that also need food for survival. Based on these characteristics, we argue against the likelihood that AI could develop artifacts endowed with veritable agency, belonging to the artifact and not the engineer who created it initially. Moreover, a purported AI agent would be unable to self-maintain and/or self-reproduce and generate its own material substrate (i.e., the hardware which is clearly designed by humans) as a *bona-fide* agent would. Additionally, as we mentioned above, it would be problematic to decide who is going to 'evaluate' the success of the AI's 'actions'. Would it be the purported agent (intrinsic agency) or its creator (extrinsic agency)? We conclude that the pressing problem with AI is not the creation of minimal artificial agents or truly agentive intelligence, but rather the possibility that AI constructs might generate nefarious consequences totally attributable to human agency, human intelligence and the human ethical standards of their designers and users. We concur with Noble and Noble (this issue) on the need to regulate the design and use of AI, regardless of whether it or any other artifacts created by humans will ever be able to generate true agency.

Acknowledgments

The authors are grateful to Matteo Mossio and Cheryl Schaeberle for their critical input. The authors have no competing financial interests to declare.

References

Barandiaran, X, Di Paolo, E, & Rohde, M 2009 "Defining agency. Individuality, normativity, asymmetry and spatio-temporality in action", *Journal of Adaptive Behavior*, vol. 17, article 367e386.

Canguilhem, G 1991 *The Normal and the Pathological*. New York: Zone Books.

Canguilhem, G 2008 *Knowledge of Life*. New York: Fordham University Press.

Cottrell, A 1979 "The natural philosophy of engines". *Contemporary Physics*, vol. 20, pp. 1–10.

Gambarotto, A 2014 "Vital forces and organization: Philosophy of nature and biology in Karl Friedrich Kielmeyer", *Studies in History and Philosophy of Science 4B Part A*, pp. 12–20. DOI: 10.1016/j.shpsc.2014.07.007

Gould, SJ 1990 *Wonderful Life: The Burgess Shale and the Nature of History*. New York: WW Norton and Company.

Kauffman, SA 1993 The Origins of Order. Oxford: Oxford University Press.

Lenoir, T 1982 *The Strategy of Life: Teleology and Mechanics in Nineteenth-Century Biology*. Dordrecht, Holland: D Reidel Publishing.

Longo, G, Miquel, PA, Sonnenschein, C, & Soto, AM 2012. "Is information a proper observable for biological organization?", *Progress in Biophysics and Molecular Biology*, vol. 109, pp. 108–114.

Longo, G, Montévil, M, & Kauffman, S 2012 "No entailing laws, but enablement in the evolution of the biosphere". Genetic and Evolutionary Computation Conference. New York: Association for Computing Machinery.

Longo, G, Montévil, M, Sonnenschein, C, & Soto, AM 2015 "In search of principles for a theory of organisms", *Journal of Biosciences*, vol. 40, pp, 955–968.

Longo, G, & Soto, AM 2016 "Why do we need theories?" *Progress in Biophysics and Molecular Biology*, vol. 122, pp. 4–10.

Maturana, HR, & Varela, FG 1980 *Autopoiesis and Cognition. The Realization of the Living*. Dordrecht: Reidel Publishing.

Montévil, M, & Mossio, M 2020 "The identity of organisms in scientific practice: Integrating historical and relational conceptions", *Frontiers in Physiology*, vol. 11, article 611.

Moreno, A 2018 "On minimal autonomous agency: Natural and artificial", *Complex Systems*, vol. 27, pp. 289–313.

Moreno, A 2023 "Some reflections on the evolution of conscious agents: The relevance of body plans", *Biosemiotics*, vol. 16, pp. 35–43.

Moreno, A, & Mossio, M 2015. *Biological Autonomy: A Philosophical and Theoretical Enquiry*. New York: Springer.

Moss, L 2003. *What Genes Can't Do*. Cambridge, MA: MIT Press.

Mossio, M, Montévil, M, & Longo, G 2016 "Theoretical principles for biology: Organization", *Progress in Biophysics and Molecular Biology*, vol. 122, pp. 24–35.

Mossio, M, & Moreno, A 2010 "Organisational closure in biological organisms", *History and Philosophy of Life Sciences*, vol. 32, pp. 269–288.

Nicolis, G & Prigogine, I 1977. *Self-organization in Non-equilibrium Systems*. New York: Wiley.

Soto, AM & Sonnenschein, C 2006. "Emergentism by default: A view from the bench. New perspectives on reduction and emergence in physics, biology and psychology", *Synthese*, vol. 151, pp. 361–376.

Soto, AM, & Sonnenschein, C 2018 "Reductionism, organicism, and causality in the biomedical sciences: A critique", *Perspectives in Biology and Medicine*, vol. 61, pp. 489–502.

Soto, AM, & Sonnenschein, C 2020 "Information, programme, signal: Dead metaphors that negate the agency of organisms", *Interdisciplinary Science Reviews*, vol. 45, pp. 331–343.

Soto, AM, & Sonnenschein, C 2023 "Georges Canguilhem, the health-disease transition and the return of organicism", *Organisms. Journal of Biological Sciences*, vol. 6, no. 1, pp. 41–48.

Walsh, D. 2015. *Organisms, Agency, and Evolution*. Cambridge: Cambridge University Press.