frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 28 January 2022
doi: 10.3389/fevo.2021.806283

How Organisms Come to Know the World: Fundamental Limits on Artificial General Intelligence

Andrea Roli[1,2]*, *Johannes Jaeger*[3]* and *Stuart A. Kauffman*[4]

[1] Department of Computer Science and Engineering, Campus of Cesena, Università di Bologna, Bologna, Italy, [2] European Centre for Living Technology, Venezia, Italy, [3] Complexity Science Hub (CSH) Vienna, Vienna, Austria, [4] Institute for Systems Biology, Seattle, WA, United States

Artificial intelligence has made tremendous advances since its inception about seventy years ago. Self-driving cars, programs beating experts at complex games, and smart robots capable of assisting people that need care are just some among the successful examples of machine intelligence. This kind of progress might entice us to envision a society populated by autonomous robots capable of performing the same tasks humans do in the near future. This prospect seems limited only by the power and complexity of current computational devices, which is improving fast. However, there are several significant obstacles on this path. General intelligence involves situational reasoning, taking perspectives, choosing goals, and an ability to deal with ambiguous information. We observe that all of these characteristics are connected to the ability of identifying and exploiting new affordances—opportunities (or impediments) on the path of an agent to achieve its goals. A general example of an affordance is the use of an object in the hands of an agent. We show that it is impossible to predefine a list of such uses. Therefore, they cannot be treated algorithmically. This means that "AI agents" and organisms differ in their ability to leverage new affordances. Only organisms can do this. This implies that true AGI is not achievable in the current algorithmic frame of AI research. It also has important consequences for the theory of evolution. We argue that organismic agency is strictly required for truly open-ended evolution through radical emergence. We discuss the diverse ramifications of this argument, not only in AI research and evolution, but also for the philosophy of science.

Keywords: artificial intelligence (AI), universal turing machine, organizational closure, agency, affordance, evolution, radical emergence, artificial life (ALife)

OPEN ACCESS

Edited by:
Giorgio Matassi, FRE3498 Ecologie et dynamique des systámes anthropiss (EDYSAN), France

Reviewed by:
Terrence William Deacon, University of California, Berkeley, United States
Kalevi Kull, University of Tartu, Estonia

***Correspondence:**
Andrea Roli
andrea.roli@unibo.it
Johannes Jaeger
jaeger@csh.ac.at

Specialty section:
This article was submitted to Models in Ecology and Evolution, a section of the journal Frontiers in Ecology and Evolution

Received: *31 October 2021*
Accepted: *30 December 2021*
Published: *28 January 2022*

Citation:
Roli A, Jaeger J and Kauffman SA (2022) How Organisms Come to Know the World: Fundamental Limits on Artificial General Intelligence. Front. Ecol. Evol. 9:806283. doi: 10.3389/fevo.2021.806283

1. INTRODUCTION

Since the founding Dartmouth Summer Research Project in 1956 (McCarthy et al., 1955), the field of artificial intelligence (AI) has attained many impressive achievements. The potential of automated reasoning, problem solving, and machine learning has been unleashed through a wealth of different algorithms, methods, and tools (Russell and Norvig, 2021). Not only do AI systems accomplish to perform intricate activities, *e.g.,* playing games (Silver et al., 2016), and to plan complex tasks (LaValle, 2006), but most current apps and technological devices are equipped with some AI component. The impressive recent achievements of machine learning (Domingos, 2015) have greatly extended the domains in which AI can be applied, from machine translation to

automatic speech recognition. AI is becoming ubiquitous in our lives. In addition, AI methods are able to produce some kinds of creative artworks, such as paintings (Hong and Curran, 2019), and music (Briot and Pachet, 2020); moreover, GPT-3, the latest version of a deep learning system able to generate texts characterized by surprising writing abilities, has recently been released (Brown et al., 2020) surrounded by some clamor (Chalmers, 2020; Marcus and Davis, 2020).

These are undoubtedly outstanding accomplishments. However, each individual success remains limited to quite narrowly defined domains. Most of today's AI systems are target-specific: an AI program capable of automatically planning tasks, for example, is not usually capable of recognizing faces in photographs. Such specialization is, in fact, one of the main elements contributing to the success of these systems. However, the foundational dream of AI—featured in a large variety of fantastic works in science-fiction—is to create a system, maybe a robot, that incorporates a wide range of adaptive abilities and skills. Hence, the quest for *Artificial General Intelligence (AGI)*, computational systems able to connect, integrate, and coordinate these various capabilities. In fact, true *general intelligence* can be defined as the ability of combining "analytic, creative, and practical intelligence" (Roitblat, 2020, page 278). It is acknowledged to be a distinguishing property of "natural intelligence," for example, the kind of intelligence that governs some of the behavior of humans as well as other mammalian and bird species.

If one considers the human brain as a computer—and by this we mean some sort of computational device equivalent to a universal Turing machine—then the achievement of AGI might simply rely on reaching a sufficient level of intricacy through the combination of different task-solving capabilities in AI systems. This seems eminently feasible—a mere extrapolation of current approaches in the context of rapidly increasing computing power—even though it requires not only the combinatorial complexification of the AI algorithms themselves, but also of the methods used to train them. In fact, many commentators predict that AGI is just around the corner, often admonishing us about the great (even existential) potentials and risks associated with this technological development (see, for example, Vinge, 1993; Kurzweil, 2005; Yudkowsky, 2008; Eden et al., 2013; Bostrom, 2014; Shanahan, 2015; Chalmers, 2016; Müller and Bostrom, 2016; Ord, 2020).

However, a number of serious problems arise when considering the higher-level integration of task-solving capabilities. All of these problems are massively confounded by the fact that real-world situations often involve information that is irrelevant, incomplete, ambiguous, and/or contradictory. First, there is the formal problem of choosing an appropriate metric for success (a cost or evaluation function) according to context and the task at hand. Second, there is the problem of identifying worthwhile tasks and relevant contextual features from an abundance of (mostly irrelevant) alternatives. Finally, there is the problem of defining what is worthwhile in the first place. Obviously, a truly general AI would have to be able to identify and refine its goals autonomously, without human intervention. In a quite literal sense, it would have to know what it wants, which presupposes that it must be capable of wanting something in the first place.

The problem of machine wanting has often been linked by philosophers to arguments about cognition, the existence of subjective mental states and, ultimately, to questions about consciousness. A well-known example is John Searle's work on minds and AI (see, for example, Searle, 1980, 1992). Other philosophers have attempted to reduce machine wanting to cybernetic goal-seeking feedback (e. g., McShea, 2012, 2013, 2016). Here, we take the middle ground and argue that the problem is rooted in the concept of organismic agency, or *bio-agency* (Moreno and Etxeberria, 2005; Barandiaran et al., 2009; Skewes and Hooker, 2009; Arnellos et al., 2010; Campbell, 2010; Arnellos and Moreno, 2015; Moreno and Mossio, 2015; Meincke, 2018). We show that the term "agency" refers to radically different notions in organismic biology and AI research.

The organism's ability to act is grounded in its functional organization, which grants it a certain autonomy (a "freedom from immediacy") (Gold and Shadlen, 2007). An organism not only passively reacts to environmental inputs. It can initiate actions according to internal goals, which it seeks to attain by leveraging opportunities and avoiding obstacles it encounters in its *umwelt*, that is, the world as perceived by this particular organism (Uexküll von, 2010; Walsh, 2015). These opportunities and obstacles are *affordances*, relations between the living agential system and its umwelt that are relevant to the attainment of its goals (Gibson, 1966). Organismic agency enables a constructive dialectic between an organism's goals, its repertoire of actions, and its affordances, which all presuppose and generate each other in a process of constant emergent co-evolution (Walsh, 2015).

Our argument starts from the simple observation that the defining properties of natural systems with general intelligence (such as organisms) require them to take advantage of affordances under constraints given by their particular motivations, abilities, resources, and environments. In more colloquial terms, general intelligences need to be able to invent, to improvise, to *jury-rig* problems that are relevant to their goals. However, AI agents (unlike biological ones) are defined as sophisticated *algorithms* that process information from percepts (inputs) obtained through sensors to actions (outputs) implemented by effectors (Russell and Norvig, 2021). We elaborate on the relation between affordances and algorithms—defined as computational processes that can run on universal Turing machines—ultimately arriving at the conclusion that identifying and leveraging affordances goes beyond algorithmic computation. This leads to two profound implications. First, while it may still be possible to achieve powerful AI systems endowed with quite impressive and general abilities, AGI cannot be fully attained in computational systems that are equivalent to universal Turing machines. This limitation holds for both non-embodied and embodied Turing machines, such as robots. Second, based on the fact that only true agents can harvest the power of affordances, we conclude that only biological agents are capable of generating truly open-ended evolutionary dynamics, implying that algorithmic attempts at creating such dynamics in the field of artificial life (aLife) are doomed to fail.

Our argument proceeds as follows: In Section 2, we provide a target definition for AGI and describe some major obstacles on the way to achieve it. In Section 3, we define and contrast the notion of an agent in organismic biology and AI research. Section 4 introduces the crucial role that affordances play in AGI, while Section 5 elucidates the limitations of algorithmic agents when it comes to identifying and leveraging affordances. In Section 6, we show that our argument also applies to embodied AI agents such as robots. Section 7 presents a number of possible objections to our argument. Section 8 discusses the necessity of bio-agency for open-ended evolution. Finally, Section 9 concludes the discussion with a few remarks on the scientific and societal implications of our argument.

2. OBSTACLES TOWARD ARTIFICIAL GENERAL INTELLIGENCE

The proposal for the Dartmouth Summer Research Project begins with an ambitious statement: "An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (McCarthy et al., 1955). Over the 66 years that have passed since this was written, the field of AI research has made enormous progress, and specialized AI systems have been developed that find application across almost all aspects of human life today (see Introduction). However, the original goal of devising a system capable of integrating *all* the various capabilities required for "true machine intelligence" has not yet been reached.

According to Roitblat (2020), the defining characteristics of *general intelligence* are:

- reasoning and problem-solving,
- learning,
- inference-making,
- using common-sense knowledge,
- autonomously defining and adjusting goals,
- dealing with ambiguity and ill-defined situations, and
- creating new representations of the knowledge acquired.

Some of these capabilities are easier to formalize than others. Automated reasoning, problem-solving, learning, and inference-making, for example, can be grounded in the principles of formal logic, and are reaching impressive levels of sophistication in contemporary deep-learning approaches (Russell and Norvig, 2021). In contrast, the complete algorithmic formalization of the other items on the list remains elusive. We will discuss the problem of autonomously defining goals shortly. The three remaining characteristics are not only hard to implement algorithmically, but are difficult to define precisely in the first place. This vagueness is of a semantic and situational nature: it concerns the meaning of concepts to an agent, the knower, in their particular circumstances.

For example, we have no widely agreed-upon definition of what "common-sense knowledge" is. In fact, it is very likely that there is no generalizable definition of the term, as "common sense" represents a kind of perspectival knowing that depends radically on context. It represents a way of reacting to an everyday problem that is shared by many (or all) people at a given location and time. It is thus an intrinsically situational and normative concept, and its meaning can shift drastically across different societal and historical contexts. What it would mean for a computer to have "common sense" remains unclear: does it have to act in a way that humans of its time and location would consider commonsensical? Or does it have to develop its own kind of computer-specific, algorithmic "common sense"? What would that even mean?

Exactly the same problem affects the ability of AI algorithms to create new representations of knowledge. Those representations must not only correspond to some state of affairs in the world, but must also be relatable, understandable, and useful to some kind of knowing agent. They must represent something *to someone*. But who? Is the task of AGI to generate representations for human understanding? If not, what kind of sense does it make for a purely algorithmic system to generate representations of knowledge? It does not need them, since it does not use visualizations or metaphors for reasoning and understanding. Again, the semantic nature of the problem makes it difficult to discuss within the purely syntactic world of algorithmic AIs.

Since they cannot employ situational knowledge, and since they cannot represent and reason metaphorically, AI systems largely fail at dealing with and exploiting *ambiguities* (Byers, 2010). These limitations have been identified and formulated as *the frame problem* more than fifty years ago by Dreyfus (1965) (see also Dreyfus, 1992). Today, they are still with us as major obstacles for achieving AGI. What they have in common is an inability of algorithmic systems to reckon with the kind of uncertainty, or even paradox, that arises from context-dependent or ill-defined concepts. In contrast, the tension created by such unresolved states of knowing is often a crucial ingredient for human creativity and invention (see, for example, Scharmer and Senge, 2016).

Let us argue the case with an illustrative example. The ability to exploit ambiguities plays a role in almost any human cognitive activity. It can turn up in the most unexpected places, for instance, in one of the most rule-based human activities—an activity that we might think should be easy to formalize. As Byers beautifully observes about creativity in mathematics, "[a]mbiguity, which implies the existence of multiple, conflicting frames of reference, is the environment that gives rise to new mathematical ideas. The creativity of mathematics does not come out of algorithmic thought" (Byers, 2010, page 23). In situated problem-solving, ambiguity is oftentimes the cornerstone of a solution process. Let us consider the mathematical riddle in **Figure 1**: if we break ambiguities by taking a purely formalized algebraic perspective, the solution we find is hardly simple.[1] Yet, if we change perspective and we observe the graphical shape of the digits, we can easily note that what is summed up are the closed loops present in the numeric symbols. It turns out that the puzzle, as it is formulated, requires the ability to observe from different perspectives, to dynamically shift perceptive and

[1] In the sense of a suitable model explaining the riddle. See Burnham and Anderson (2002).

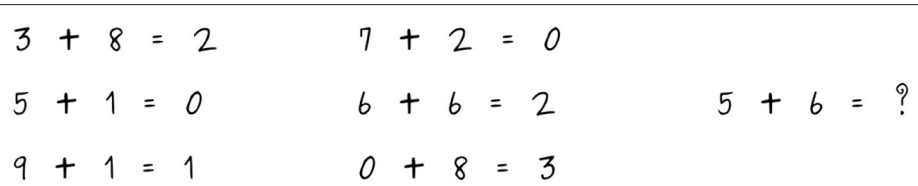

FIGURE 1 | A riddle found by one of the authors in a paper left in the coffee room of a department of mathematics.

cognitive frames, mixing both graphical and algebraic approaches for a simple solution.

Following Byers, we observe that even a strongly formalized human activity—the process of discovery in mathematics—is not entirely captured by an algorithmic search. A better metaphor would be an erratic walk across dark rooms. As Andrew Wiles describes his journey to the proof of the Fermat conjecture,[2] the solution process starts from a dark room where we "stumble around bumping into the furniture;" suddenly we find the light switch and, in the illuminated room, we "can see were we were"—an insight! Then, we move to an adjacent dark room and continue this process finding successive "light switches" into further dark rooms until the problem, at last, is solved. Each step from one room to the other is an *insight*, not a deduction and not an induction. The implication is fundamental: the mathematician comes to know a new world via an insight. The insight itself is not algorithmic. It is an act of *semantic meaning-making*. Roger Penrose makes the same point in the *Emperor's New Mind* (Penrose, 1989).

Human creativity, in all kinds of contexts, seems to require frame-switching between metaphorical or formal representations, alongside our capabilities of dealing with contradictions and ambiguities. These are not only hallmarks of human creative processes, but should also characterize AGI systems. As we will see, these abilities crucially rely on *affordances* (Gibson, 1966). Therefore, we must ask whether universal Turing machines can identify and exploit affordances. The initial step toward an answer to this question lies in the recognition that affordances arise from interactions between an agent and its umwelt. Therefore, we must first understand what an agent is, and how the concept of an "agent" is defined and used in biology and in AI research.

3. BIO-AGENCY: CONTRASTING ORGANISMS TO AI AGENTS

So far, we have avoided the question how an AGI could choose and refine its own goals (Roitblat, 2020). This problem is distinct, but still tightly related to the issues of ambiguity and representation discussed in the previous section. Selecting goals has two aspects. The first is that one must motivate the choice of a goal. One must want to reach some goal to have a goal at all, and one must have needs to want something. The other aspect is to prioritize some particular goal over a set of alternatives, according to the salience and the alignment of the chosen goal with one's own needs and capabilities in a given context.

Choosing a goal, of course, presupposes a certain *autonomy*, *i. e.*, the ability to make a "choice" (Moreno and Mossio, 2015). Here, we must emphasize again that our use of the term "choice" does *not* imply consciousness, awareness, mental states, or even cognition, which we take to involve at least some primitive kind of nervous system (Barandiaran and Moreno, 2008). It simply amounts to a system which is capable of selecting from a more or less diversified repertoire of alternative dynamic behaviors ("actions") that are at its disposal in a given situation (Walsh, 2015). All forms of life—from simple bacteria to sophisticated humans—have this capability. The most central distinction to be made here is that the selection of a specific behavior is not purely reactive, not entirely determined by environmental conditions, but (at least partially) originates from and depends on the internal organization of the system making the selection. This implies some basic kind of *agency* (Moreno and Mossio, 2015). In its broadest sense, "agency" denotes the ability of a system to initiate actions from within its own boundaries, causing effects that emanate from its own internal dynamics.

Agency requires a certain type of functional organization. More specifically, it requires *organizational closure* (Piaget, 1967; Moreno and Mossio, 2015), which leads to autopoietic (*i. e.,* self-making, self-maintaining, and self-repairing) capabilities (Maturana and Varela, 1973, 1980). It also leads to *self-determination* through self-constraint: by maintaining organizational closure, an organism is constantly providing the conditions for its own continued existence (Bickhard, 2000; Mossio and Bich, 2017). This results in the most basic, metabolic, form of autonomy (Moreno and Mossio, 2015). A minimal *autonomous agent* is a physically open, far-from-equilibrium, thermodynamic system capable of self-reproduction and self-determination.

Organisms, as autonomous agents, are *Kantian wholes*, *i. e.*, organized beings with the property that the parts exist for and by means of the whole (Kant, 1892; Kauffman, 2000, 2020). "Whole" indicates that organizational closure is a systems-level property. In physical terms, it can be formulated as a *closure of constraints* (Montévil and Mossio, 2015; Moreno and Mossio, 2015; Mossio et al., 2016). Constraints change the dynamics of the underlying processes without being altered themselves (at least not at the same time scale). Examples of constraints in organisms include enzymes, which catalyze biochemical reactions without being altered in the process, or the vascular system in vertebrates, which regulates levels of nutrients, hormones, and oxygen in

[2] *Nova* interview, https://www.pbs.org/wgbh/nova/article/andrew-wiles-fermat.

different parts of the body without changing itself at the time scale of those physiological processes (Montévil and Mossio, 2015).

It is important to note that constraint closure does not imply a fixed (static) network of processes and constraints. Instead, *organizational continuity* is maintained if the current closed organization of a system causally derives from previous instantiations of organizational closure, that is, its particular organized state at this moment in time is *dynamically presupposed* by its earlier organized states (Bickhard, 2000; DiFrisco and Mossio, 2020). Each successive state can (and indeed must) differ in their detailed physical structure from the current state. To be a Kantian whole, an autonomous system must perform at least one work-constraint cycle: it must perform physical work to continuously (re)constitute closure through new as well as recurring constraints (Kauffman, 2000, 2003; Kauffman and Clayton, 2006). Through each such cycle, a particular set of constraints is propagated, selected from a larger repertoire of possible constraints that all realize closure. In this way, the system's internal dynamics kinetically "lift" a set of mutually constituting processes from the totality of possible dynamics. This is how organizational closure leads to autopoiesis, basic autonomy, and self-determination by self-constraint: the present structure of the network of interacting processes that get "lifted" is (at least to some degree) the product of the previous unfolding of the organized network. In this way, organization maintains and propagates itself.

However, one key ingredient is still missing for an agent that actively chooses its own goals. The basic autonomous system we described above can maintain (and even repair) itself, but it cannot adapt to its circumstances—it cannot react adequately to influences from its environment. This adaptive capability is crucial for prioritizing and refining goals according to a given situation. The organism can gain some autonomy over its interactions with the environment if it is capable of regulating its own boundaries. These boundaries are required for autopoiesis, and thus must be part of the set of components that are maintained by closure (Maturana and Varela, 1980). Once boundary processes and constraints have been integrated into the closure of constraints, the organism has attained a new level of autonomy: *interactive autonomy* (Moreno and Mossio, 2015). It has now become a fully-fledged *organismal agent*, able to perceive its environment and to select from a repertoire of alternative actions when responding to environmental circumstances based on its internal organization. Expressed a bit more colloquially, making this selection requires being able to perceive the world and to evaluate "what's good or bad for me," in order to act accordingly. Here, the transition *from matter to mattering* takes place.

Interactive autonomy provides a naturalistic (and completely scientific) account of the kind of *bio-agency* (and the particular kind of goal-directedness or teleology that is associated with it, Mossio and Bich, 2017), which grounds our examination of how organisms can identify and exploit affordances in their umwelt. But before we get to this, let us contrast the complex picture of an organismal agent as a Kantian whole with the much simpler concept of an agent in AI research. In the context of AI, "[a]n agent is anything that can be viewed as *perceiving* its environment through *sensors and acting* upon that environment through *effectors*" (Russell and Norvig, 2021, original emphasis). In other words, an AI agent is an input–output processing device. Since the point of AI is to do "a good job of acting on the environment" (Russell and Norvig, 2021), the internal processing can be quite complicated, depending on the task at hand. This very broad definition of an AI agent in fact includes organismal agents, since it does not specify the kind of processes that mediate between perception and action. However, although not always explicitly stated, it is generally assumed that input-output processing is performed by some sort of algorithm that can be implemented on a universal Turing machine. The problem is that such algorithmic systems have no freedom from immediacy, since all their outputs are determined entirely—even though often in intricate and probabilistic ways—by the inputs of the system. There are no actions that emanate from the historicity of internal organization. *There is, therefore, no agency at all in an AI "agent."* What that means and why it matters for AGI and evolution will be the subject of the following sections.

4. THE KEY ROLE OF AFFORDANCES

Having outlined a suitable naturalistic account of bio-agency, we can now revisit the issue of identifying and exploiting affordances in the umwelt, or perceived environment, of an organism. The concept of an *affordance* was first proposed by Gibson (1966) in the context of ecological psychology. It was later adopted to diverse fields of investigation such as biosemiotics (Campbell et al., 2019) and robotics (Jamone et al., 2016). "Affordances" refer to what the environment offers to an agent (in the organismic sense defined above), for "good or ill." They can be manifested as opportunities or obstacles on our path to attain a goal. A recent philosophical account emphasizes the relation between the agent and its perceived environment (its umwelt), stating that affordances guide and constrain the behavior of organisms, precluding or allowing them to perform certain actions, showing them what they can and cannot do (Heras-Escribano, 2019, p. 3). A step, for instance, affords us the action of climbing; a locked door prevents us from entering. Affordances fill our world with meaning: organisms do not live in an inert environment, but "are surrounded by promises and threats" (Heras-Escribano, 2019, p. 3).

The dialectic mutual relationship between goals, actions, and affordances is of crucial importance here (Walsh, 2015). Affordances, as we have seen, require an agent with goals. Those goals motivate the agent to act. The agent first chooses which goal to pursue. It then selects an action from its repertoire (see Section 3) that it anticipates to be conducive to the attainment of the goal. This action, in turn, may alter the way the organism perceives its environment, or it may alter aspects of the environment itself, which leads to an altered set of affordances present in its umwelt. This may incite the agent to choose an alternative course of action, or even to reconsider its goals. In addition, the agent can learn to perform new actions or develop new goals along the way. This results in a constructive co-emergent dynamic in which sets of goals, actions, and affordances

continuously generate and collapse each other as the world of the agent keeps entering into the next space of possibilities, its next *adjacent possible* (Kauffman, 2000). Through this co-emergent dialectic, new goals, opportunities, and ways of acting constantly arise. Since the universe is vastly non-ergodic, each moment in time provides its own unique set of opportunities and obstacles, affording new kinds of goals and actions (Kauffman, 2000). In this way, true novelty enters into the world through *radical emergence*—the generation, over time, of opportunities and rules of engagement and interaction that did not exist at any previous time in the history of the universe.

A notable example of such a co-emergent process in a human context is *jury-rigging*: given a leak in the ceiling, we cobble together a cork wrapped in a wax-soaked rag, stuff it into the hole in the ceiling, and hold it in place with duct tape (Kauffman, 2019). In general, solving a problem through jury-rigging requires several steps and involves different objects and actions, which articulate together toward a solution of the problem, mostly without any predetermined plan. Importantly, jury-rigging uses only specific subsets of the totality of *causal properties* of each object involved. Often, these properties do not coincide with previously known functional features of the object. Consider a tool, like a screwdriver, as an example. Its original purpose is to tighten screws. But it can also be used to open a can of paint, wedge a door open, scrape putty off the window, to stab or poke someone (please don't), or (should you feel so inclined) to pick your nose with it. What is important to note here is that any physical object has an *indefinite* number of alternative uses in the hands of an agent (Kauffman, 1976). This does not mean that its uses are *infinite*—even if they might be—but rather that *they cannot be known* (and thus prestated) in advance.

Ambiguity and perspective-taking also play a fundamental role in jury-rigging, as the goal of the task is to find suitable *novel* causal properties of the available objects to solve the problem at hand. The same happens in an inverse process, where we observe an artifact (or an organism Kauffman, 2019), and we aim at providing an explanation by articulating its parts, along with the particular function they carry out. For example, if we are asked what the use of an automobile is, we would probably answer that it is a vehicle equipped with an engine block, wheels, and other parts, whose diverse causal features can be articulated together to function as a locomotion and transportation system. This answer resolves most ambiguities concerning the automobile and its parts by providing a coherent frame in which the parts of the artifact are given a specific function, aimed at explaining its use as a locomotion and transportation system. In contrast, if one supposes that the purpose of an automobile is to fry eggs, one would partition the system into different sets of parts that articulate together in a distinct way such that eggs can be fried on the hot engine block. In short, for the inverse process with regard to artifacts (or organisms) what we "see it as doing" drives us to decompose the system into parts in different ways (Kauffman, 1976). Each such decomposition identifies precisely that subset of the causal properties of the identified parts that articulate together to account for and explain "what the system is doing" according to our current frame. It is critical to note that there is no universal or unique decomposition, since the way to decompose the system depends on its use and context (see also Wimsatt, 2007).

To close the loop of our argument, we note that the prospective uses of an object (and hence the decomposition we choose to analyze it) depend on the goals of the agent using it, which, in turn, depend on the agent's repertoire of actions and the affordances available to it, which change constantly and irreversibly over time. It is exactly *because* all of these are constantly evolving through their co-emergent dialectic interactions that the number of uses of an object remains *indefinite* and, in fact, *unknowable* (Kauffman, 2019). Moreover, and this is important: there is no deductive relation between the uses of an object. Take, for example, an engine block, designed to be a propulsive device in a car. It can also serve as the chassis of a tractor. Furthermore, one can use it as a bizarre (but effective) paper weight, its cylinder bores can host bottles of wine, or it can be used to crack open coconuts on one of its corners. In general, we cannot know the number of possible uses of an engine block, and we cannot deduce one use from another: the use as a paper weight abstracts from details that can conversely be necessary for it to be used to crack open coconuts. As Robert Rosen put it, complex systems invariably retain hidden properties, and their manipulation can always result in unintended consequences (Rosen, 2012). Even worse, we have seen that the relation between different uses of a thing is merely *nominal*, as there is no kind of ordering that makes it possible to relate them in a more structured way (Kauffman, 2019; Kauffman and Roli, 2021b).

This brings us to a cornerstone of our argument: when jury-rigging, it is impossible to compose any sort of well-defined list of the possible uses of the objects to be used. By analogy, **it is impossible to list all possible goals, actions, or affordances of an organismic agent in advance. In other words, Kantian wholes can not only identify and exploit affordances, but they constantly *generate* new opportunities for themselves *de novo***. Our next question is: can algorithmic systems such as AI "agents" do this?

5. THE BOUNDED RATIONALITY OF ALGORITHMS

In the introduction, we have defined an algorithm as a computational process that can run on a universal Turing machine. This definition considers algorithms in a broad sense, including computational processes that do not halt. All algorithms operate deductively (Kripke, 2013). When implementing an algorithm as a computer program by means of some kind of formal language (including those based on recursive functional programming paradigms), we must introduce specific data and code structures, their properties and interactions, as well as the set of operations we are allowed to perform on them, in order to represent the objects and relations that are relevant for our computation. In other words, we must provide a precisely defined *ontology* on which the program can operate deductively, *e.g.,* by

drawing inferences or by ordering tasks for solving a given problem. In an algorithmic framework, novelty can only be represented combinatorially: it manifests as new combinations, mergers, and relations between objects in a (potentially vast, but predefined) space of possibilities. This means that an algorithm cannot discover or generate truly novel properties or relations that were not (at least implicitly) considered in its original ontology. Therefore, an algorithm operating in a deductive manner cannot jury-rig, since it cannot find new causal properties of an object that were not already inherent in its logical premises.

To illustrate this central point, let us consider *automated planning*: a planning program is given an initial state and a predefined goal, and its task is to find a feasible—and ideally optimal—sequence of actions to reach the goal. What makes this approach successful is the possibility of describing the objects involved in the task in terms of their properties, and of representing actions in terms of the effects they produce on the world delimited by the ontology of the program, plus the requirements that need to be satisfied for their application. For the planner to work properly, there *must* be deductive relations among the different uses of an object, which are exploited by the inference engine to define an evaluation function that allows it to arrive at a solution. The problem with the planner is that, in general, there is *no deductive relation* between the possible uses of an object (see Section 4). From the use of an engine block as a paper weight, the algorithm cannot deduce its use as a method to crack open coconuts. It can, of course, find the latter use if it can be deduced, *i. e.*, if there are: *(i)* a definitive list of properties, including the fact that the engine block has rigid and sharp corners, *(ii)* a rule stating that one can break objects in the class of "breakable things" by hitting them against objects characterized by rigid and sharp corners, and *(iii)* a fact stating that coconuts are breakable.

The universe of possibilities in a computer program—however, broadly construed—is like a world of LEGO^{TM} bricks: components with predefined properties and compositional relations can generate a huge space of possible combinations, even unbounded if more bricks can always be supplemented. However, if we add scotch tape, which makes it possible to assemble bricks without being constrained by their compositional mechanism, and a cutter, which enables us to cut the bricks into smaller pieces of any shape, then rules and properties are no longer predefined. We can no longer prestate a well-defined list of components, with associated properties and relations. We now have a universe of indefinite possibilities, and we are no longer trapped inside the formal frame of algorithms. *Formalization has reached its limits.* What constitutes a meaningful compositional relation becomes a semantic question, depending on our particular circumstances and the whims of our creative mind. Our possibilities may not be infinite, but they become impossible to define in advance. And because we can no longer list them, we can no longer treat them in a purely algorithmic way. This is how human creativity transcends the merely combinatorial innovative capacities of any AI we can build today. **Algorithms cannot take or shift perspective and that is why they cannot leverage ambiguity for innovation in the way an organismic agent can. Algorithms cannot jury-rig**.

At the root of this limitation is the fact that algorithms cannot want anything. To want something implies having goals that matter to us. We have argued in Section 3, that only organismic agents (but not algorithmic AI "agents") can have goals, because of their being Kantian wholes with autopoietic organization and closure of constraints. Therefore, nothing matters to an algorithm. But without mattering or goals, an algorithm has no means to identify affordances (in fact, it has no affordances), unless they are already formally predefined in its ontology, or can be derived in some logical way from predefined elements of that ontology. Thus, the algorithm cannot generate meaning where there was none before. It cannot engage in the process of *semiosis* (Peirce, 1934, p. 488). For us to make sense of the world, we must take a perspective: we must see the world from a specific point of view, contingent on our nature as fragile, limited, mortal beings which circumscribes our particular goals, abilities, and affordances. This is how organismic agents generate new frames in which to formalize posssibilities. This is how we tell what is relevant to us from what is not. Algorithms cannot do this, since they have no point of view, and require a predefined formal frame to operate deductively. To them, everything and nothing is relevant at the same time.

Now, we must draw our attention to an issue that is often neglected when discussing the nature of general intelligence: for a long time, we have believed that coming to know the world is a matter of induction, deduction, and abduction (see, for example, Hartshorne and Weiss, 1958; Mill, 1963; Ladyman, 2001; Hume, 2003; Okasha, 2016; Kennedy and Thornberg, 2018). Here, we show that this is not enough.

Consider *induction*, proceeding from a finite set of examples to an hypothesis of a universal. We observe many black ravens and formulate the hypothesis that "all ravens are black." Observe that the relevant variables and properties are already prestated, namely "ravens" and "black." Induction is over already identified features of the world and, by itself, does not identify new categories. In induction, there is an imputation of a property of the world (black) with respect to things we have already identified (ravens). There is however no insight with respect to new features of the world (*cf.* Section 2). Let us pause to think about this: induction by itself cannot reveal novel features of the world—features that are not already in our ontology.

This is even more evident for *deduction*, which proceeds from prestated universal categories to the specific. "All men are mortal, Socrates is a man, therefore Socrates is a mortal." All theorems and proofs in mathematics have this deductive structure. However, neither induction nor deduction by themselves can reveal novel features of the world not already in our ontology.

Finally, we come to *abduction*, which aims at providing an explanation of an observation by asserting an already known precondition that is likely to have this observation as a consequence. For example, if we identify an automobile as a means of locomotion and transportation, and had decomposed it into parts that articulate together to support its function as a means of locomotion and transportation, we are then able to explain its failure to function in this sense by a failure of one

of its now defined parts. If the car does not turn on, we can suppose the battery is dead. Abduction is differential diagnosis from a prestated set of conditions and possibilities that articulate to carry out what we "see the system as doing or being." But there is no unique decomposition. The number of decompositions is indefinite. Therefore, when implemented in a computer program, this kind of reasoning cannot reveal novel features of the world not already in the ontology of the program.

To summarize: with respect to coming to know the world, **once we have carved the world into a finite set of categories, we can no longer see the world beyond those categories**. In other words, new meanings—along with their symbolic grounding in real objects—are outside of the predefined ontology of an agential system. The same limitation also holds for probabilistic forms of inference, involving, *e.g.,* Bayesian nets (see Gelman et al., 2013). Consider the use of an engine block as a paper weight, and a Bayesian algorithm updating to improve engine blocks with respect to functioning as a paper weight. No such updating will reveal that engine blocks can also be used to crack open coconuts. The priors for such an innovation could not be deduced, even in principle. Similarly, Markov blankets (see, for example, Hipólito et al., 2021) are restricted to pre-existing categories.

Organisms come to know new features of the world by semiosis—a process which involves *semantic meaning-making* of the kind described above, not just formal (syntactic) reasoning through deduction, induction, or abduction. This is true of mathematicians. It is also true of Caledonian crows who solve problems of astonishing complexity, requiring sophisticated multi-step jury-rigging (Taylor et al., 2010). Chimpanzees learning to use tools have the same capacity to improvise (Köhler, 2013). Simpler organisms—down to bacteria—must have it too, although probably in a much more limited sense. After all, they are at the basis of an evolutionary process toward more complex behavior, which presupposes the identification and exploitation of new opportunities. Our human ontology has evolved into a much more complex state than that of a primitive unicellular organism. In general, all organisms act in alignment with their goals, capabilities, and affordances (see Section 4), and their agential behavior can undergo variation and selection. A useful action—exploiting a novel affordance—can be captured by heritable variation (at the genetic, epigenetic, behavioral, or cultural level) and thus passed on across generations. This "coming to know the world" is what makes the evolutionary expansion of our ontologies possible. It goes beyond induction, deduction, and abduction. Organisms can do it, but universal Turing machines cannot.

In conclusion, the rationality of algorithms is bounded by their ontology. However vast this ontology may be, algorithms cannot transcend their predefined limitations, while organisms can. This leads us to our central conclusion, which is both radical and profound: **not all possible behaviors of an organismic agent can be formalized and performed by an algorithm—not all organismic behaviors are Turing-computable. Therefore, organisms are not Turing machines. It also means that true AGI cannot be achieved in an algorithmic frame**, since AI "agents" cannot choose and define their own goals, and hence exploit affordances, deal with ambiguity, or shift frames in ways organismic agents can. Because of these limitations, algorithms cannot evolve in truly novel directions (see Section 8 below).

6. IMPLICATIONS FOR ROBOTS

So far, we have only considered algorithms that run within some stationary computing environment. The digital and purely virtual nature of this environment implies that all features within in must, by definition, be formally predefined. Its digital environment, in its finite totality, *is* the ontology of an AI algorithm. There is nothing outside it for the AI "agent" to discover. The real world is not like that. We have argued in the previous sections that our world is full of surprises that cannot be entirely formalized, since not all future possibilities can be prestated. Therefore, the question arises whether an AI agent that does get exposed to the real world could identify and leverage affordances when it encounters them.

In other words, does our argument apply to *embodied Turing machines*, such as robots, that interact with the physical world through sensors and actuators and may be able to modify their bodily configuration? The crucial difference to a purely virtual AI "agent" is that the behavior of a robot results from interactions between its control program (an algorithm), its physical characteristics (which define its repertoire of actions), and the physical environment in which it finds itself (Pfeifer and Bongard, 2006). Moreover, learning techniques are put to powerful use in robotics, meaning that robots can adapt their behavior and improve their performance based on their relations to their physical environment. Therefore, we can say that robots are able to learn from experience and to identify specific sensory-motor patterns in the real world that are useful to attain their goals (Pfeifer and Scheier, 2001). For instance, a quadruped robot controlled by an artificial neural network can learn to control its legs on the basis of the forces perceived from the ground, so as to develop a fast and robust gait. This learning process can be guided either by a task-oriented evaluation function, such as forward gait speed, or a task-agnostic one that rewards coordinated behaviors (Prokopenko, 2013), or both.

Does that mean that robots, as embodied Turing machines, can identify and exploit affordances? Does it mean that robots, just like organisms, have an umwelt full of opportunities and threats? As in the case of stationary AI "agents," the answer is a clear and resounding "no." The same problems we have discussed in the previous sections also affect robotics. Specifically, they manifest themselves as the symbol grounding problem and the frame problem. The *symbol grounding problem* concerns the issue of attaching symbols to sensory-motor patterns (Harnad, 1990). It amounts to the question whether it is feasible for a robot to detect relevant sensory-motor patterns that need to be associated with new concepts—*i. e.*, new variables in the ontology of the robot. This, in turn, leads to the more general *frame problem* (see Section 2 and McCarthy and Hayes, 1969): the problem of specifying in a given situation what is relevant for a robot's goals. Again, we run into the problem of choosing one's own goals, of shifting frames, and of dealing with ambiguous information that cannot be formalized in the form of a predefined set of possibilities.

As an example, consider the case of a robot whose goal it is to open coconuts. Its only available tool is an engine block, which it currently uses as a paper weight. There are no other tools, and the coconuts cannot be broken by simply throwing them against a wall. In order to achieve its goal, the robot must acquire information on the relevant causal features of the engine block to open coconuts. Can it exploit this affordance? The robot can move around and perceive the world via its sensors. It can acquire experience by performing random moves, one of which may cause it to hit the engine block, to discover that the block has the property of being "hard and sharp," which is useful for cracking the nut. However, how does the robot know that it needs to look for this property in the objects of its environment? This is but the first useful step in solving the problem. By the same random moves, the robot might move the engine block, or tip it on its side. How can the robot "understand" that "hard and sharp" will prove to be useful, but "move to the left" will not? How long will this single step take?

Furthermore, if the coconut is lying beside the engine block, tipping it over may lead to the nut being cracked as well. How can the robot connect several coordinated causal features to achieve its goal, if none of them can be deduced from the others? The answer is: it cannot. We observe that **achieving the final goal may require connecting several relevant coordinated causal features of real-world objects, none of which is deducible from the others**. This is analogous to the discovery process in mathematics we have described in Section 2: wandering through a succession of dark rooms, each transition illuminated by the next in a succession of insights. There is no way for the robot to know that it is improving over the incremental steps of its search. Once an affordance is identified, new affordances emerge as a consequence and the robot cannot "know" in advance that it is accumulating successes until it happens upon the final achievement: there is no function optimization to be performed over such a sequence of steps, no landscape to search by exploiting its gradients, because each step is a search in a space of possibilities that cannot be predefined. The journey from taking the first step to reaching the ultimate goal is blind luck over some unknown time scale. With more steps, it becomes increasingly difficult to know if the robot improves, since reaching the final goal is in general not an incremental process.

The only way to achieve the robot's ultimate goal is for it to already have a preprogrammed ontology that allows for multi-step inferences. Whether embodied or not, the robot's control algorithm can only operate deductively. But if the opportunity to crack open coconuts on the engine block has been predefined, then it does not really count as discovering a new causal property. It does not count as exploiting a novel affordance. **Robots do not generate new opportunities for themselves in the way organisms do. Even though engaging with their environment, they cannot participate in the emergent triad of goals, actions, and affordances** (see Section 4). Therefore, we must conclude that its embodied nature does not really help a robotic algorithm to achieve anything resembling true AGI.

7. POSSIBLE OBJECTIONS

We suspect that our argument may raise a number of objections. In this section, we anticipate some of these, and attempt to provide adequate replies.

A first potential objection concerns *the ability of deep-learning algorithms to detect novel correlations* in large data sets in an apparently hypothesis-free and unbiased manner. The underlying methods are mainly based on complex network models, rather than traditional sequential formal logic. When the machine is trained with suitable data, shouldn't it be able to add new symbols to its ontology that represent the newly discovered correlations? Would this not count as identifying and exploiting a new affordance? While it is true that the ontology of such a deep-learning machine is not explicitly predefined, it is nevertheless implicitly given through the constraints of the algorithm and the training scenario. Correlations can only be detected between variables that are defined through an external model of the data. Moreover, all current learning techniques rely on the maximization (or minimization) of one or more evaluation functions. These functions must be provided by the designers of the training scenario, who thus determine the criteria for performance improvement. The program itself does not have the ability of choosing the goal of the task at hand. This even holds for task-agnostic functions of learning scenarios, as they again are the result of an imposed external choice. In the end, with no bias or hypothesis at all, what should the learning program look for? In a truly bias- or hypothesis-free scenario (if that is possible at all), any regularity (even if purely accidental) would become meaningful (Calude and Longo, 2017), which results in no meaning at all. Without any goal or perspective, there is no insight to be gained.

A second objection might be raised concerning the rather common observation that AI systems, such as programs playing chess or composing music, often surprise us or behave in unpredictable ways. However, *machine unpredictability* does not imply that their behavior is not deducible. Instead, it simply means that we cannot find an explanation for it, maybe due to a lack of information, or due to our own limited cognitive and/or computational resources. For example, a machine playing chess can take decisions by exploiting a huge repertoire of moves, and this may produce surprising behavior in the eye of the human opponent, since it goes far beyond our own cognitive capacity. Nevertheless, the behavior of the machine is deductively determined, ultimately based on simple combinatorics. More generally, it is well-known that there are computer programs whose output is not compressible. Their behavior cannot be predicted other than actually running the full program. This computationally irreducible behavior cannot be anticipated, but it is certainly algorithmic. Due to their competitive advantage when dealing with many factors, or many steps, in a deductive procedure, AI "agents" can easily fool us by mimicking creative behavior, even though their algorithmic operation does not allow for the kind of semantic innovation even a simple organism is capable of.

A third objection could be that our argument carelessly ignores potential progress in computational paradigms and robot

design that may lead to a solution of the apparently irresolvable problems we present here. A common futurist scenario in this context is one in which AI "agents" themselves replace human engineers in designing AI architectures, leading to *a technological singularity*—a technology which is far beyond human grasp (see, for example, Vinge, 1993; Kurzweil, 2005; Eden et al., 2013; Bostrom, 2014; Shanahan, 2015; Chalmers, 2016). We are sympathetic to this objection (although not to the notion of a singularity based on simple extrapolation of our current capabilities). Our philosophical approach is exactly based on the premise that the future is open, and will always surprise us in fundamentally unpredictable ways. But there is no paradox here: what we are arguing for is that AGI is impossible *within the current algorithmic frame* of AI research, which is based on Turing machines. We are open to suggestions how the limitations of this frame could be transcended. One obvious way to do this is a biological kind of robotics, which uses organismic agents (such as biological cells) to build organic computation devices or robots. We are curious (and also apprehensive) concerning the potential (and dangers) which such non-algorithmic frameworks hold for the future. An AGI which *could* indeed choose its own goals, would not be aligned with our own interests (by definition), and may not be controllable by humans, which seems to us to defy the purpose of generating AI as a benign and beneficial technology in the first place.

One final, and quite serious, philosophical objection to our argument is that it may be impossible to empirically distinguish between *a sophisticated algorithm mimicking agential behavior*, and true organismic agency as outlined in Section 3. In this case, our argument may be of no practical importance. It is true that humans are easily fooled into interpreting completely mechanistic behavior in intentional and teleological terms. Douglas Hofstadter (2007), for example, mentions a dot of red light that is moving along the walls of the San Francisco Exploratorium, responding by simple feedback to movements of the museum visitors. Every time a visitor tries to touch the dot, it seems to escape at the very last moment. Even though based on a simple feedback mechanism, it is tempting to interpret such behavior as intentional.[3] Could we have fallen prey to such an illusion when interpreting the behavior of organisms as true agency? We do not think so. First, the organizational account of agency we rely on not only accounts for goal-oriented behavior, but also for basic functional properties of living systems, such as their autopoietic ability to self-maintain and self-repair. Thus, agency is a higher-level consequence of more basic abilities of organisms that cannot easily be accounted for by alternative explanations. Even though these basic abilities have not yet been put to the test in a laboratory, there is no reason to think that they won't be in the not-too-far future. Second, we think the account of organismic agency presented here is preferable over an algorithmic explanation of "agency" as evolved input-output processing, since it has much greater explanatory power. It takes the phenomenon of agency seriously instead of trying to explain it away. Without this conceptual framework, we could not even ask the kind of questions raised in this paper, since they would never arise within an algorithmic framework. In essence, the non-reductionist (yet still naturalist) world we operate in is richer than the reductionist one in that it allows us to deal scientifically with a larger range of undoubtedly interesting and relevant phenomena (see also Wimsatt, 2007).

[3]Regarding machine intentionality see also the work by Braitenberg (1986).

8. OPEN-ENDED EVOLUTION IN COMPUTER SIMULATIONS

Before we conclude our argument, we would like to consider its implications beyond AGI, in particular, for the theory of evolution, and for research in the field of artificial life (ALife). One of the authors has argued earlier that evolvability and agency *must* go together, because the kind of organizational continuity that turns a cell cycle into a *reproducer*—the minimal unit of Darwinian evolution—also provides the evolving organism with the ability to act autonomously (Jaeger, 2022). Here, we go one step further and suggest that **organismic agency is a fundamental prerequisite for open-ended evolution**, since it enables organisms to identify and exploit affordances in their umwelt, or perceived environment. Without agency, there is no co-emergent dialectic between organisms' goals, actions, and affordances (see Section 5). And without this kind of dialectic, evolution cannot transcend its predetermined space of possibilities. It cannot enter into the next adjacent possible. It cannot truly innovate, remaining caught in a deductive ontological frame (Fernando et al., 2011; Bersini, 2012; Roli and Kauffman, 2020).

Let us illustrate this with the example of ALife. The ambitious goal of this research field is to create models of digital "organisms" that are able to evolve and innovate in ways equivalent to natural evolution. Over the past decades, numerous attempts have been made to generate open-ended evolutionary dynamics in simulations such as Tierra (Ray, 1992) and Avida (Adami and Brown, 1994). In the latter case, the evolving "organisms" reach an impressive level of sophistication (see, for example, Lenski et al., 1999, 2003; Zaman et al., 2014). They have an internal "metabolism" that processes nutrients to gain energy from their environment in order to survive and reproduce. However, this "metabolism" does not exhibit organizational closure, or any other form of true agency, since it remains purely algorithmic. And so, no matter how complicated, such evolutionary simulations always tend to get stuck at a certain level of complexity (Bedau et al., 2000; Standish, 2003). Even though some complexification of ecological interactions (*e.g.,* mimics of trophic levels or parasitism) can occur, we never observe any innovation that goes beyond what was implicitly considered in the premises of the simulation. This has led to some consternation and the conclusion that the strong program of ALife—to generate any truly life-like processes in a computer simulation—has failed to achieve its goal so far. In fact, we would claim that this failure is comprehensive: it affects all

attempts at evolutionary simulation that have been undertaken so far. Why is that so?

Our argument provides a possible explanation for the failure of strong ALife: even though the digital creatures of Avida, for example, can exploit "new" nutrient sources, they can only do so because these sources have been endowed with the property of being a potential food source at the time the simulation was set up. They were part of its initial ontology. The algorithm cannot do anything it was not (implicitly) set up to do. Avida's digital "life forms" can explore their astonishingly rich and large space of possibilities combinatorially. This is what allows them, for example, to feed off other "life forms" to become predators or parasites. The resulting outcomes may even be completely unexpected to an outside observer with insufficient information and/or cognitive capacity (see Section 7). However, Avida's "life forms" can never discover or exploit any truly new opportunities, like even the most primitive natural organisms can. They cannot generate new meaning that was not already programmed into their ontology. They cannot engage in semiosis. What we end up is a very high-dimensional probabilistic combinatorial search. Evolution has often been likened to such intricate search strategies, but our view suggests that organismic agency pushes it beyond.

Organismic open-ended evolution into the adjacent possible requires the identification and leveraging of novel affordances. In this sense, it cannot be entirely formalized. In contrast, algorithmic evolutionary simulations will forever be constrained by their predefined formal ontologies. They will never be able to produce any true novelty, or radical emergence. They are simply not like organismic evolution since they lack its fundamental creativity. As some of us have argued elsewhere: emergence is not engineering (Kauffman and Roli, 2021a). The biosphere is an endlessly propagating adapting construction, not an entailed algorithmic deduction (Kauffman, 2019). In other words, the world is not a theorem (Kauffman and Roli, 2021b), but a neverending exploratory process. It will never cease to fascinate and surprise us.

9. CONCLUSION

In this paper, we have argued two main points: (1) AGI is impossible in the current algorithmic frame of research in AI and robotics, since algorithms cannot identify and exploit new affordances. (2) As a direct corollary, truly open-ended evolution into the adjacent possible is impossible in algorithmic systems, since they cannot transcend their predefined space of possibilities.

Our way of arriving at these conclusions is not the only possible one. In fact, the claim that organismic behavior is not entirely algorithmic was made by Robert Rosen as early as the 1950s (Rosen, 1958a,b, 1959, 1972). His argument is based on category theory and neatly complements our way of reasoning, corroborating our insight. It is summarized in Rosen's book "Life Itself" (Rosen, 1991). As a proof of principle, he devised a diagram of compositional mappings that exhibit *closure to efficient causation*, which is equivalent to organizational closure (see Section 3). He saw this diagram as a highly abstract relational representation of the processes that constitute a living system. Rosen was able to prove mathematically that this type of organization "has no largest model" (Rosen, 1991). This has often been confounded with the claim that it cannot be simulated in a computer at all. However, Rosen is not saying that we cannot generate algorithmic models of some (maybe even most) of the behaviors that a living system can exhibit. In fact, it has been shown that his diagram *can* be modeled in this way using a recursive functional programming paradigm (Mossio et al., 2009). What Rosen *is* saying is exactly what we are arguing here: there will always be some organismic behaviors that cannot be captured by a preexisting formal model. This is an *incompleteness argument* of the kind Gödel made in mathematics (Nagel and Newman, 2001): for most problems, it is still completely fine to use number theory after Gödel's proof. In fact, relevant statements about numbers that do not fit the theory are exceedingly rare in practice. Analogously, we can still use algorithms implemented by computer programs to study many aspects of organismic dynamics, or to engineer (more or less) target-specific AIs. Furthermore, it is always possible to extend the existing formal model to accommodate a new statement or behavior that does not yet fit in. However, this process is infinite. We will never arrive at a formal model that captures *all* possibilities. Here, we show that this is because those possibilities cannot be precisely prestated and defined in advance.

Another approach that comes to very similar insights to ours is *biosemiotics* (see, for example, Hoffmeyer, 1993; Barbieri, 2007; Favareau, 2010; Henning and Scarfe, 2013). Rather than a particular field of inquiry, biosemiotics sees itself as a broad and original perspective on life and its evolution. It is formulated in terms of the production, exchange, and interpretation of signs in biological systems. The process of meaning-making (or semiosis) is central to biosemiotics (Peirce, 1934). Here, we link this process to autopoiesis (Varela et al., 1974; Maturana and Varela, 1980) and the organizational account, which sees bio-agency grounded in a closure of constraints within living systems (Montévil and Mossio, 2015; Moreno and Mossio, 2015; Mossio et al., 2016), and the consequent co-emergent evolutionary dialectic of goals, actions, and affordances (Walsh, 2015; Jaeger, 2022). Our argument suggests that the openness of semiotic evolution is grounded in our fundamental inability to formalize and prestate possibilities for evolutionary and cognitive innovation in advance.

Our insights put rather stringent limitations on what traditional mechanistic science and engineering can understand and achieve when it comes to agency and evolutionary innovation. This affects the study of any kind of agential system—in computer science, biology, and the social sciences—including higher-level systems that contain agents, such as ecosystems or the economy. In these areas of investigation, any purely formal approach will remain forever incomplete. This has important repercussions for the philosophy of science: the basic problem is that, with respect to coming to know the world, once we have carved it into a finite set of categories, we can no longer see beyond those categories. The grounding of meaning in real objects is outside any predefined formal ontology. The evolution

of scientific knowledge itself is entailed by no law. It cannot be formalized (Kauffman and Roli, 2021a,b).

What would such a *meta-mechanistic science* look like? This is not entirely clear yet. Its methods and concepts are only now being elaborated (see, for example, Henning and Scarfe, 2013). But one thing seems certain: it will be a science that takes agency seriously. It will allow the kind of teleological behavior that is rooted in the self-referential closure of organization in living systems. It is naturalistic but not reductive. Goals, actions, and affordances are emergent properties of the relationship between organismal agents and their umwelt—the world of meaning they live in. This emergence is of a radical nature, forever pushing beyond predetermined ontologies into the adjacent possible. This results in a worldview that closely resembles Alfred North Whitehead's *philosophy of organism* (Whitehead, 1929). It sees the world less as a clockwork, and more like an evolving ecosystem, a creative process centered around harvesting new affordances.

It should be fairly obvious by now that our argument heavily relies on teleological explanations, necessitated by the goal-oriented behavior of the organism. This may seem problematic: teleological explanations have been traditionally banned from evolutionary biology because they seemingly require (1) an inversion of the flow from cause to effect, (2) intentionality, and (3) a kind of normativity, which disqualify them from being proper naturalistic scientific explanations.

Here, we follow Walsh (2015), who provides a very convincing argument that this is not the case. First, it is important to note that we are not postulating any large-scale teleology in evolution—no omega point toward which evolution may be headed. On the contrary, our argument for open-endedness explicitly precludes such a possibility, even in principle (see Section 8). Second, the kind of teleological explanation we propose here for the behavior of organisms and its evolution is *not* a kind of causal explanation. While causal explanations state which effect follows which cause, teleological explanation deals with the conditions that are conducive for an organism to attain its goal. The goal does not *cause* these conditions, but rather presupposes them. Because of this, there is no inversion of causal flow. Finally, the kind of goal-directed behavior enabled by bio-agency does *not* require awareness, intentionality, or even cognition. It can be achieved by the simplest organisms (such as bacteria), simply due to the fact that they exhibit an internal organization based on a closure of constraints (see Section 3). This also naturalizes the kind of normativity we require for teleology (Mossio and Bich, 2017): the organism really does have a goal from which it can deviate. That goal is to stay alive, reproduce, and flourish. All of this means that there is nothing supranatural or unscientific about the kind of teleological explanations that are used in our argument. They are perfectly valid explanations. There is no need to restrict ourselves to strictly mechanistic arguments, which yield an impoverished world view since they cannot capture the deep problems and rich phenomena we have been discussing throughout this paper.

While such metaphysical and epistemological considerations are important for understanding ourselves and our place in the world, our argument also has eminently practical consequences. The achievement of AGI is often listed as one of the most threatening existential risks to the future of humanity (see, for example, Yudkowsky, 2008; Ord, 2020). Our analysis suggests that such fears are greatly exaggerated. No machine will want to replace us, since no machine will want anything, at least not in the current algorithmic frame of defining a machine. This, of course, does not prevent AI systems and robots from being harmful. Protocols and regulations for AI applications are urgent and necessary. But AGI is not around the corner, and we are not alone with this assessment. The limits of current AI applications have been questioned by others, emphasizing that these systems lack autonomy and understanding capabilities, which we conversely find in natural intelligence (Nguyen et al., 2015; Broussard, 2018; Hosni and Vulpiani, 2018; Marcus and Davis, 2019; Mitchell, 2019; Roitblat, 2020; Sanjuán, 2021; Schneier, 2021). The true danger of AI lies in the social changes and the disenfranchisement of our own agency that we are currently effecting through target-specific algorithms. It is not Skynet, but Facebook, that will probably kill us in the end.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

All authors contributed equally to this manuscript, conceived the argument, and wrote the paper together.

ACKNOWLEDGMENTS

JJ has profited from numerous discussions on the topic of organismic agency with Jan-Hendirk Hofmeyr, Matteo Mossio, and Denis Walsh, and on possibility spaces with Andrea Loettgers and Tarja Knuuttila, none of whom necessarily share the opinions expressed in this paper. JJ thanks the late Brian Goodwin for crucial early influences on his thinking. No doubt, Brian would have loved this paper.

REFERENCES

Adami, C., and Brown, C. T. (1994). "Evolutionary learning in the 2D artificial life system 'Avida'," in *Artificial Life IV: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*, eds P. Maes and R. Brooks (Cambridge, MA: MIT Press), 377–381.

Arnellos, A., and Moreno, A. (2015). Multicellular agency: an organizational view. *Biol. Philosophy* 30, 333–357. doi: 10.1007/s10539-015-9484-0

Arnellos, A., Spyrou, T., and Darzentas, J. (2010). Towards the naturalization of agency based on an interactivist account of autonomy. *New Ideas Psychol.* 28, 296–311. doi: 10.1016/j.newideapsych.2009.09.005

Barandiaran, X., and Moreno, A. (2008). On the nature of neural information: a critique of the received view 50 years later. *Neurocomputing* 71, 681–692. doi: 10.1016/j.neucom.2007.09.014

Barandiaran, X. E., Di Paolo, E., and Rohde, M. (2009). Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. *Adapt. Behav.* 17, 367–386. doi: 10.1177/1059712309343819

Barbieri, M., editor (2007). *Introduction to Biosemiotics: The New Biological Synthesis*. Dordrecht, NL: Springer.

Bedau, M. A., McCaskill, J. S., Packard, N. H., Rasmussen, S., Adami, C., Green, D. G., et al. (2000). Open problems in artificial life. *Artif. Life* 6, 363–376. doi: 10.1162/106454600300103683

Bersini, H. (2012). Emergent phenomena belong only to biology. *Synthese* 185, 257–272. doi: 10.1007/s11229-010-9724-4

Bickhard, M. H. (2000). Autonomy, function, and representation. *Commun. Cogn. Artif. Intell.* 17, 111–131.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Braitenberg, V. (1986). *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press.

Briot, J.-P., and Pachet, F. (2020). Deep learning for music generation: challenges and directions. *Neural Comput. Appl.* 32, 981–993. doi: 10.1007/978-3-319-70163-9

Broussard, M. (2018). *Artificial Unintelligence: How Computers Misunderstand the World*. Cambridge, MA: MIT Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *arXiv preprint* arXiv:2005.14165.

Burnham, K., and Anderson, D. (2002). *Model Selection and Multi-Model Inference*, 2nd Edn, New York, NY: Springer.

Byers, W. (2010). *How Mathematicians Think*. Princeton, NJ: Princeton University Press.

Calude, C., and Longo, G. (2017). The deluge of spurious correlations in big data. *Found. Sci.* 22, 595–612. doi: 10.1007/s10699-016-9489-4

Campbell, C., Olteanu, A., and Kull, K. (2019). Learning and knowing as semiosis: extending the conceptual apparatus of semiotics. *Sign Syst. Stud.* 47, 352–381. doi: 10.12697/SSS.2019.47.3-4.01

Campbell, R. (2010). The emergence of action. *New Ideas Psychol.* 28, 283–295. doi: 10.1016/j.newideapsych.2009.09.004

Chalmers, D. (2020). GPT-3 and general intelligence. *Daily Nous* 30.

Chalmers, D. J. (2016). "The singularity: a philosophical analysis," in *Science Fiction and Philosophy*, ed S. Schneider (Hoboken, NJ: John Wiley & Sons, Inc), 171–224.

DiFrisco, J., and Mossio, M. (2020). Diachronic identity in complex life cycles: an organizational perspective," in *Biological Identity: Perspectives from Metaphysics and the Philosophy of Biology*, eds A. S. Meincke, and J. Dupré (London: Routledge).

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York, NY: Basic Books.

Douglas Hofstadter, R. (2007). *I Am a Strange Loop*. New York, NY: Basic Books.

Dreyfus, H. (1965). *Alchemy and Artificial Intelligence*. Technical Report, RAND Corporation, Santa Monica, CA, USA.

Dreyfus, H. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press.

Eden, A. H., Moor, J. H., Soraker, J. H., and Steinhart, E., editors (2013). *Singularity Hypotheses: A Scientific and Philosophical Assessment*. New York, NY: Springer.

Favareau, D., editor (2010). *Essential Readings in Biosemiotics*. Springer, Dordrecht, NL.

Fernando, C., Kampis, G., and Szathmáry, E. (2011). Evolvability of natural and artificial systems. *Proc. Compu. Sci.* 7, 73–76. doi: 10.1016/j.procs.2011.12.023

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., and Vehtari, A. (2013). *Bayesian Data Analysis*, 3rd Edn. Boca Raton, FL: Taylor & Francis Ltd.

Gibson, J. (1966). *The Senses Considered as Perceptual Systems*. London: Houghton Mifflin.

Gold, J. I., and Shadlen, M. N. (2007). The neural basis of decision making. *Ann. Rev. Neurosci.* 30, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038

Harnad, S. (1990). The symbol grounding problem. *Physica D Nonlin. Phenomena* 42, 335–346. doi: 10.1016/0167-2789(90)90087-6

Hartshorne, C., and Weiss (1958). *Collected Papers of Charles Sanders Peirce*. Boston, MA: Belknap Press of Harvard University Press.

Henning, B., and Scarfe, A., editors (2013). *Beyond Mechanism: Putting Life Back Into Biology*. Plymouth: Lexington Books.

Heras-Escribano, M. (2019). *The Philosophy of Affordances*. London: Springer.

Hipólito, I., Ramstead, M., Convertino, L., Bhat, A., Friston, K., and Parr, T. (2021). Markov blankets in the brain. *Neurosci. Biobehav. Rev.* 125, 88–97. doi: 10.1016/j.neubiorev.2021.02.003

Hoffmeyer, J. (1993). *Signs of Meaning in the Universe*. Bloomington, IN: Indiana University Press.

Hong, J.-W., and Curran, N. (2019). Artificial intelligence, artists, and art: attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* 15, 1–16. doi: 10.1145/3326337

Hosni, H., and Vulpiani, A. (2018). Data science and the art of modelling. *Lettera Matematica* 6, 121–129. doi: 10.1007/s40329-018-0225-5

Hume, D. (2003). *A Treatise of Human Nature*. Chelmsford, MA: Courier Corporation.

Jaeger, J. (2022). "The fourth perspective: evolution and organismal agency," in *Organization in Biology*, ed M. Mossio (Berlin: Springer).

Jamone, L., Ugur, E., Cangelosi, A., Fadiga, L., Bernardino, A., Piater, et al. (2016). Affordances in psychology, neuroscience, and robotics: a survey. *IEEE Trans. Cogn. Develop. Syst.* 10, 4–25. doi: 10.1109/TCDS.2016.2594134

Kant, I. (1892). *Critique of Judgement*. New York, NY: Macmillan.

Kauffman, S. (1976). "Articulation of parts explanation in biology and the rational search for them," in *Topics in the Philosophy of Biology*, (Dordrecht: Springer), 245–263.

Kauffman, S. (2000). *Investigations*. Oxford: Oxford University Press.

Kauffman, S. (2003). Molecular autonomous agents. *Philosoph. Trans. Roy. Soc. London Series A Math. Phys. Eng. Sci.* 361, 1089–1099. doi: 10.1098/rsta.2003.1186

Kauffman, S. (2019). *A World Beyond Physics: the Emergence and Evolution of Life*. Oxford: Oxford University Press.

Kauffman, S. (2020). Eros and logos. *Angelaki* 25, 9–23. doi: 10.1080/0969725X.2020.1754011

Kauffman, S., and Clayton, P. (2006). On emergence, agency, and organization. *Biol. Philosophy* 21, 501–521. doi: 10.1007/s10539-005-9003-9

Kauffman, S., and Roli, A. (2021a). The third transition in science: beyond Newton and quantum mechanics – a statistical mechanics of emergence. *arXiv preprint* arXiv:2106.15271.

Kauffman, S., and Roli, A. (2021b). The world is not a theorem. *Entropy* 23:1467. doi: 10.3390/e23111467

Kennedy, B., and Thornberg, R. (2018). "Deduction, induction, and abduction," in *The SAGE Handbook of Qualitative Data Collection* (London: SAGE Publications), 49–64.

Köhler, W. (2013). *The Mentality of Apes*. London: Routledge.

Kripke, S. (2013). "The Church-Turing "thesis" as a special corollary of Gödel's completeness theorem," in *Computability: Gödel, Turing, Church, and Beyond*, eds B. Copeland, C. Posy, and O. Shagrir (Cambridge, MA: The MIT Press), Ch. 4, 77–104.

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. New York, NY: The Viking Press.

Ladyman, J. (2001). *Understanding Philosophy of Science*. London: Routledge.

LaValle, S. (2006). *Planning Algorithms*. Cambridge: Cambridge University Press.

Lenski, R. E., Ofria, C., Collier, T. C., and Adami, C. (1999). Genome complexity, robustness and genetic interactions in digital organisms. *Nature* 400, 661–664. doi: 10.1038/23245

Lenski, R. E., Ofria, C., Pennock, R. T., and Adami, C. (2003). The evolutionary origin of complex features. *Nature* 423, 139–144. doi: 10.1038/nature01568

Marcus, G. and Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York, NY: Vintage.

Marcus, G. and Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *Technol. Rev.*

Maturana, H., and Varela, F. (1973). *De Maquinas y Seres Vivos*. Santiago: Editorial Universitaria.

Maturana, H., and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: Springer.

McCarthy, J., and Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Mach. Intell.* 463–502.

McCarthy, J., Minsky, M., Rochester, N., and Shannon, C. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. Available online at: http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

McShea, D. W. (2012). Upper-directed systems: a new approach to teleology in biology. *Biol. Philosophy* 27, 63–684. doi: 10.1007/s10539-012-9326-2

McShea, D. W. (2013). Machine wanting. *Stud. History Philosophy Sci. Part C Biol. Biomed. Sci.* 44, 679–687. doi: 10.1016/j.shpsc.2013.05.015

McShea, D. W. (2016). Freedom and purpose in biology. *Stud. History Philosophy Sci. Part C Biol. Biomed. Sci.* 58, 64–72. doi: 10.1016/j.shpsc.2015.12.002

Meincke, A. S. (2018). "Bio-agency and the possibility of artificial agents," in *Philosophy of Science (European Studies in Philosophy of Science), Vol. 9*, Eds. A. Christian, D. Hommen, N. Retzlaff, and G. Schurz (Cham: Springer International Publishing), 65–93.

Mill, J. (1963). *Collected Works*. Toronto, ON: University of Toronto Press.

Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. London: Penguin UK.

Montévil, M., and Mossio, M. (2015). Biological organisation as closure of constraints. *J. Theor. Biol.* 372, 179–191. doi: 10.1016/j.jtbi.2015.02.029

Moreno, A., and Etxeberria, A. (2005). Agency in natural and artificial systems. *Artif. Life* 11, 161–175. doi: 10.1162/1064546053278919

Moreno, A., and Mossio, M. (2015). *Biological Autonomy*. Dordrecht: Springer.

Mossio, M., and Bich, L. (2017). What makes biological organisation teleological? *Synthese* 194, 1089–1114. doi: 10.1007/s11229-014-0594-z

Mossio, M., Longo, G., and Stewart, J. (2009). A computable expression of closure to efficient causation. *J. Theor. Biol.* 257, 489–498. doi: 10.1016/j.jtbi.2008.12.012

Mossio, M., Montévil, M., and Longo, G. (2016). Theoretical principles for biology: organization. *Progr. Biophys. Mol. Biol.* 122, 24–35. doi: 10.1016/j.pbiomolbio.2016.07.005

Müller, V. C., and Bostrom, N. (2016). "Future progress in artificial intelligence: a survey of expert opinion," in *Fundamental Issues of Artificial Intelligence*, Vol. 376, eds V. C. Müller (Cham: Springer International Publishing), 555–572.

Nagel, E., and Newman, J. R. (2001). *Gödel's Proof*. New York, NY: NYU Press.

Nguyen, A., Yosinski, J., and Clune, J. (2015). "Deep neural networks are easily fooled: high confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 427–436.

Okasha, S. (2016). *Philosophy of Science: A Very Short Introduction*, 2nd Edn, Oxford: Oxford University Press.

Ord, T. (2020). *The Precipice*. New York, NY: Hachette Books.

Peirce, C. (1934). *Collected Papers, Vol 5*. Cambridge, MA: Harvard University Press.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press.

Pfeifer, R., and Bongard, J. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*, Cambridge, MA: MIT Press.

Pfeifer, R., and Scheier, C. (2001). *Understanding Intelligence*, Cambridge, MA: The MIT Press.

Piaget, J. (1967). *Biologie et Connaissance*, Paris: Delachaux & Niestle.

Prokopenko, M. (2013). *Guided Self-Organization: Inception*, Vol. 9. Berlin: Springer Science & Business Media.

Ray, T. S. (1992). "Evolution and optimization of digital organisms," in *Scientific Excellence in Supercomputing: the 1990 IBM Contest Prize Papers*, eds K. R. Billingsley, H. U. Brown, and E. Derohanes (Atlanta, GA: Baldwin Press), 489–531.

Roitblat, H. (2020). *Algorithms Are Not Enough: Creating General Artificial Intelligence*. Cambridge, MA: MIT Press.

Roli, A., and Kauffman, S. (2020). Emergence of organisms. *Entropy* 22, 1–12. doi: 10.3390/e22101163

Rosen, R. (1958a). A relational theory of biological systems. *Bull. Math. Biophys.* 20, 245–260. doi: 10.1007/BF02478302

Rosen, R. (1958b). The representation of biological systems from the standpoint of the theory of categories. *Bull. Math. Biophys.* 20, 317–341. doi: 10.1007/BF02477890

Rosen, R. (1959). A relational theory of biological systems II. *Bull. Math. Biophys.* 21, 109–128. doi: 10.1007/BF02476354

Rosen, R. (1972). "Some relational cell models: the metabolism-repair systems," in *Foundations of Mathematical Biology, Vol. II*, ed R. Rosen (New York, NY: Academic Press), 217–253.

Rosen, R. (1991). *Life Itself: A Comprehensive Inquiry Into the Nature, Origin, and Fabrication of Life*. New York, NY: Columbia University Press.

Rosen, R. (2012). *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*, 2nd Edn. New York, NY: Springer.

Russell, S., and Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*, 4th Global Edition. London: Pearson.

Sanjuán, M. (2021). Artificial intelligence, chaos, prediction and understanding in science. *Int. J. Bifurc. Chaos* 31:2150173. doi: 10.1142/S021812742150173X

Scharmer, O., and Senge, P. (2016). *Theory U: Leading From the Future as It Emerges*, 2nd Edn, Oakland, CA: Berrett-Koehler Publishers.

Schneier, B. (2021). "The coming AI hackers," in *International Symposium on Cyber Security Cryptography and Machine Learning* Berlin: Springer, 336–360.

Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–424. doi: 10.1017/S0140525X00005756

Searle, J. R. (1992). *The Rediscovery of the Mind*. Cambridge, MA: Bradford Books.

Shanahan, M. (2015). *The Technological Singularity*. Cambridge, MA: MIT Press.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* 529, 484–489. doi: 10.1038/nature16961

Skewes, J. C., and Hooker, C. A. (2009). Bio-agency and the problem of action. *Biol. Philosophy* 24, 283–300. doi: 10.1007/s10539-008-9135-9

Standish, R. K. (2003). Open-ended artificial evolution. *Int. J. Comput. Intell. Appl.* 3, 167–175. doi: 10.1142/S1469026803000914

Taylor, A., Elliffe, D., Hunt, G., and Gray, R. (2010). Complex cognition and behavioural innovation in new caledonian crows. *Proc. R. Soc. B Biol. Sci.* 277, 2637–2643. doi: 10.1098/rspb.2010.0285

Uexküll von, J. (2010). *A Foray Into the Worlds of Animals and Humans: With a Theory of Meaning*. Minneapolis, MN: University of Minnesota Press.

Varela, F., Maturana, H., and Uribe, R. (1974). Autopoiesis: the organization of living systems, its characterization and a model. *Biosystems* 5, 187–196. doi: 10.1016/0303-2647(74)90031-8

Vinge, V. (1993). "The coming technological singularity: how to survive in the post-human era," in *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace, NASA Conference Publication CP-10129*, ed G. A. Landis (Cleveland, OH: NASA Lewis Research Center), 11–22.

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

Whitehead, A. N. (1929). *Process and Reality*. New York, NY: The Free Press.

Wimsatt, W. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Boston, MA: Harvard University Press.

Yudkowsky, E. (2008). "Artificial intelligence as a positive and negative factor in global risk," in *Global Catastrophic Risks*, eds N. Bostrom, and M. M. Cirkovic (Oxford: Oxford University Press).

Zaman, L., Meyer, J. R., Devangam, S., Bryson, D. M., Lenski, R. E., and Ofria, C. (2014). Coevolution drives the emergence of complex traits and promotes evolvability. *PLoS Biol.* 12:e1002023. doi: 10.1371/journal.pbio.1002023

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.