Illuminating the Three Dogmas of Reinforcement Learning under Evolutionary Light

Mani Hamidi[1], **Terrence Deacon**[2]
mani.hamidi@uni-tuebingen.de

[1]**Department of Computer Science, University of Tübingen, Germany**
[2]**University of California, Berkeley, USA**

Abstract

Three core tenets of reinforcement learning (RL)–concerning the definition of agency, the objective of learning, and the scope of the reward hypothesis–have been highlighted as key targets for conceptual revision, with major implications for theory and application. We propose a framework, inspired by open-ended evolutionary theory, to reconsider these three "dogmas." We revisit each assumption and address related concerns raised alongside them. To make our arguments relevant to RL as a model of biological learning, we first establish that evolutionary dynamics can plausibly operate within living brains over an individual's lifetime, and are not confined to cross-generational processes. We begin by revisiting the second dogma, drawing on evolutionary insights to enrich the "adaptation-rather-than-search" view of learning. We then address the third dogma regarding the limits of the reward hypothesis, using analogies from evolutionary fitness to illuminate the scalar reward vs. multi-objective debate. After discussing practical implications for exploration in RL, we turn to the first–and arguably most fundamental–issue: the absence of a formal account of agency. We argue that unlike the other two problems, the evolutionary paradigm alone cannot resolve the agency question, though it gestures in a productive direction. We advocate integrating ideas from origins-of-life theory, where the thermodynamics of sustenance and replication offer promising foundations for understanding agency and resource-constrained reinforcement learning in biological systems.

1 Introduction

1.1 Motivation & Overview

Three "dogmas"[1] of reinforcement learning (RL) were highlighted by Abel et al. (2024): (T1) overemphasis on formalizing the environment that yields a dearth of canonical or principled definitions of agency, (T2) a tendency to frame learning as performing a search and (T3) the insistence on the use of a single scalar reward to mediate the learning process. We will use T(1-3) to refer to these henceforth.

Our re-framing of RL, and learning more broadly, is through an evolutionary lens. While we consider the problem of agency (T1) to be at the root of the other two, we will first introduce our perspective by engaging with T2. This is because Abel et al. (2024) explicitly offer the metaphor of "adaptation" as the alternative to ameliorate the over-reliance on "search" as a metaphor for learning. To our surprise, despite the obvious evolutionary connotations that this terminology (adaptation)

[1]Here we opt for the term *tenet* rather than *dogma* a) to further reduce the negative connotations that the original authors also struggled to temper, and b) to indicate that our proposed perspective does not so much demand that we altogether "shed" these tenets, but rather re-frame them.

carries, the authors do not appear to draw on the rich history of evolutionary ideas and algorithms that could directly inform this shift to framing learning as adaptation instead of search. This is likely due to the fact that evolution is largely considered as a learning algorithm that operates on cross-generational timescales, unlike RL which is concerned with learning during a lifetime. To overcome this hurdle, we draw on a rich history of theories that hypothesize different mechanisms for the operation of Darwinian selection in brains and within a single lifetime.

Our first task (section 1.2) is therefore to provide an overview of these theories and establish a framework where evolutionary learning can operate during the lifetime of an organism alongside experience-mediated learning via neuro-plasticity.

We then use this paradigm to address T2 (section 2.1) by explicating an evolutionary notion of adaptation that is not so much in opposition to "search" as the term is commonly understood, but serves as a complementary alternative that can operate alongside more search-like algorithms such as tree search. We draw on more recent family of open-ended evolutionary algorithms (Lehman & Stanley, 2008; 2011b; Brant & Stanley, 2017) that operationalize biological phenomena like niching and coevolution to facilitate adaptation via "novelty search".

Third, in section 2.2, we apply the evolutionary perspective to the issue of the reward hypothesis (T3), specifically focusing on the contention over whether scalar rewards can adequately represent multiple objectives. We offer the notion of "minimal criteria" in open-ended evolutionary algorithms (Lehman & Stanley, 2010) as an alternative mechanism for achieving multiple objectives at once.

Despite championing evolution as the singular remedy for resolving both T2 and T3, in section 2.3, we argue against evolutionary theory as a sufficient formalism to capture the defining properties of an agent (T1). We engage with the homeostasis-versus-reward theories of motivation that also have a history in framing the RL problem to address this point. Pursuing the homeostatic view to its logical conclusion, we point out thermodynamic formalisms for the origin of life and adaptive behavior as the necessary foundation upon which any theory of agency in RL and elsewhere must be built. We argue that this is not just for the sake of theoretical completeness, but will have practical implications, namely surrounding resource-constrained models of RL.

Finally, in section 3, we take a more practical turn and highlight areas where this theoretical framing can provide new tools for a practicing RL engineer. We specifically focus on the implications of the evolutionary framework on the problems of exploration-exploitation and resource-constrained reinforcement learning.

1.2 Evolution in the Brain

The brain is not the only organ that learns to form novel associations and solve previously unseen problems from experience. While even single cells can be attributed a kind of learning capacity previously reserved for brains (Gershman et al., 2021), the adaptive immune system epitomizes non-neural learning, rivaling (simple) brains in complexity as well as learning proficiency, if not speed.

This is perhaps why the discovery of clonal selection theory of adaptive immunity made such an impression in understanding learning during lifetime of biological systems more generally. This discovery suggested that evolutionary selection on semi-randomly generated sequences of recombined somatic DNA provides a viable mechanism of learning to recognize and respond to substances (antigens) that were never seen by either the individual itself or any of its ancestors (Jerne, 1993).

One of the leading figures in establishing the clonal selection theory of learning by the immune system, Gerald Edelman, was also one of the early advocates of a similar mechanism of learning by the nervous system (Edelman, 1993). This research program came to be known as **neural Darwinism**, and despite the pursuit of complementary ideas in a similar vein (Calvin, 1998), its adoption has been limited. While the volume of research on experience-guided learning via neuro-plasticity has

catapulted over the years, research on the neural mechanisms and correlates of such neural Darwinism remain sparse.

In more recent years, a set of novel hypotheses have been offered that provide a range of hypothetical mechanisms by which Darwinian evolution can operate in the brain. Much of this work is from the efforts of Szàthmary and colleagues, and appear under the umbrella of **Darwinian neurodynamics (DN)** (Fernando et al., 2012; Szilágyi et al., 2016). This term was specifically coined to mark several points of conceptual departure from the earlier neural Darwinist paradigm of Edelman. We highlight one crucial distinguishing characteristic of this paradigm and refer the reader to the original references for more details: under the DN framework, the replicating unit that is subject to evolutionary selection can be (among other possibilities) the neural dynamics of a single or population of neurons, and is not limited to physically replicating units like neurons or synaptic connections themselves. This is most concretely demonstrated in Czégel et al. (2021), where a population of recurrent networks act as an information processing node for computations that are then read out via the activation of a single neuron. A predefined mapping then draws a many-to-one correspondence from each activation sequence to each solution. The replication and propagation of these neural patterns to adjacent processing nodes is done according to the "fitness" of the solution that it produced.

While Czégel et al. (2021) interpret each of these activation sequences as a solution to a combinatorial problem, they can equally interpretable as motor sequences, state representations, or even reward functions. The latter is particularly noteworthy, since it makes existing solutions to the problem of the origin of rewards in RL (Singh et al., 2010) relevant to learning that occurs within a lifetime rather than over many generations; we will elaborate on this point in the context of the our discussion around the reward hypothesis and an evolutionary re-visioning of T3.

The nature of such evolutionary mechanisms in biological brains are admittedly hypothetical; however, much care has been taken to demonstrate their feasibility on computational, algorithmic and implementational fronts. On the computational and algorithmic fronts, Czégel et al. (2022) highlight the correspondence between evolutionary learning and Bayesian inference, specifically particle filters. At the implementation level, Fernando et al. (2012); Czégel et al. (2021) argue for the feasibility of achieving evolutionary-like selection dynamics with local neural plasticity rules that are already well established. Finally, at the behavioral level, Czégel et al. (2021) highlight that the evolutionary phenomena of "punctuated equilibria" provides a good qualitative model of the long periods of stasis seen in human problem solving (Fedor et al., 2017).

Our goal in this article is to offer an evolutionary re-framing of the three dogmas of RL that were identified by Abel et al. (2024). Many of the attempts we make towards this end are conceptual and agnostic towards the various implementation details with which an evolutionary logic can be realized. The works of Száthmary and colleagues offer one comprehensive body of work that theorize various such mechanisms, but they are not the only one. For example, Dragoi (2023) draws on a similar Darwinian framework to suggest that the hippocampal networks are pre-configured to generate motifs of neural activity patterns that are subject to selection and modification via experience, in a direct analogy with the generation and selection of antibody sequences in the adaptive immune system. Alternatively, Pezzulo & Cisek (2016) offer a model where the brain simultaneously hosts multiple representations of potential action sequences ("affordances") that are competing before one "survives" and is executed in a continuous sensory-motor loop. For the purposes of the discussions below, any such neural paradigm that can accommodate replication with variation, along with competition and selection of some subset of these replicators *within the lifetime of an individual*, meets the necessary criteria for our evolutionary re-framing of RL.

2 Making Sense of RL in Light of Evolution

It is famously said that "nothing in biology makes sense except in the light of evolution" Dobzhansky (1973). Having reviewed various hypotheses on the possibility of Darwinian selection taking place *within* the brain, we now consider the implications that this hypothesis would have for RL and in

particular the three tenets from Abel et al. (2024) that we set out to illuminate under evolutionary light.

2.1 Adaptation Versus Search

Despite the salient association between the concept of adaptation and evolution, Abel et al. (2024) do not draw on any insights from evolutionary theory to motivate their adaptation-rather-than-search thesis. By the same token, it is important to highlight right away that the idea of "search" as a metaphor for learning is commonly applied to evolutionary dynamics as well. This then begs the question, does evolutionary adaptation actually offer any unique insights to learning that are missed by the "search" metaphor?

The distinction between "adaptation" and "search" highlighted by Abel et al. (2024), centers around the issue of convergence to a unique solution. This view is certainly consistent with the evolutionary view of adaptation where many diverse *satisficing*[2] "solutions" to the same "problem" (e.g., flight) are often discovered by evolution. However, the important insight that is offered by an evolutionary account is that adaptation is driven by an open-ended "search" that complements (if not entirely abandons) objective-oriented search for fitness or "quality" with a concurrent "search" for novelty or "diversity" (Pugh et al., 2016; Lehman & Stanley, 2011a). A common point of contention with this view is that instead of "abandoning objectives" novelty-search merely re-defines the objective to be "diversity". However, the "Quality-Diversity" (QD) family of algorithms (Brant & Stanley, 2017; 2020; Mouret & Clune, 2015) address this complaint on two fronts: 1) they do not "select" for novelty or diversity in the same sense that fitness or quality is subject to selection pressure, and 2) the selection for higher quality (fitness) candidates is not directed towards a known objective in the same way a gradient leads search in an optimization process.

We will revisit the second point in section 2.2 when we discuss fitness in relation to scalar and multi-objective optimization. It is however constructive to elaborate on the first point and explain the mechanism with which novelty and diversity *are* achieved in these algorithms. Instead of "selection pressure", the biological inspiration for novelty search is the notion of "speciation" or "niching" (Stanley & Miikkulainen, 2002; Mahfoud, 1996). This concept captures the tendency of organisms to found a new niche and thereby "avoid competition and exploit untapped resources" (Lehman & Stanley, 2010). QD algorithms encourage a diverse "archive" of solutions, by limiting selection pressure to individuals within the same niche, and avoiding premature selection against candidates that inhabit novel behavioral or phenotypic characteristics. An adjacent concept called "niche construction" (Odling-Smee et al., 2003; 2013) further enriches the adaptational view of learning in contrast to the search metaphor: organisms not only occupy a niche, but actively modify them. This results in coevolutionary dynamics between a changing agent as well as environment which acts as a positive feedback that further enhances diversity (Erwin, 2008). Indeed, such coevolutionary dynamics have also been a critical component to open-ended novelty search algorithms (Brant & Stanley, 2017; 2020; Wang et al., 2019). In these setups, the environment is defined by its own "genome" along with the agent itself, highlighting the extent to which the coevolutionary setting departs from a typical RL environment defined by an MDP. In such a coevolutionary setting where the state, action and reward landscapes (and therefore the solution landscape) are ever-shifting, the metaphor of "search", with a putative end-point, proves less useful than "open-ended adaptation".

Another point of contrast between adaptation and search is found in the colloquial usage of the term "search" which implies an a priori intention or knowledge of the solution that is being searched for. Evolution on the other hand, and despite common misunderstanding, is not teleological. In fact the term adaptation is perhaps used in relation to evolution precisely to avoid such teleological subtext that is often associated with terms like learning or search. Some authors (Dragoi, 2023) use the

[2]The verb, "to satisfice" was coined in an influential paper by Herbert Simon (Simon, 1956) with explicit reference to biological adaptation: "however adaptive the behavior of organisms in learning and choice situations, this adaptiveness falls far short of the ideal of "maximizing" postulated in economic theory." This specific reference to postulates of economic theory also foreshadows our discussion on T3 in section 2.2, where we offer an evolutionary solution for solving goal-directed problems that do not satisfy these axioms.

dichotomy of "instructional" versus "selection" theories to distinguish between the two modalities and emphasize the pre-existence of the "solutions" ahead of any interaction with the "problem" as the key differentiator of the two mechanisms. This is also consistent with "affordance competition" framing provided by Pezzulo & Cisek (2016) where the "affordance landscape" is constantly adapting not just in reaction to changes in the environment but also the actions of the agent itself.

Before we conclude our evolutionary illumination of T2, we would like to emphasize that we do not consider evolutionary adaptation as a superior mechanism to a guided search towards a known solution. It is conceivable that look-ahead search for a known or easily verifiable solutions, like those employed in Monte Carlo tree search (MCTS), can actually offer a quick and efficient way to find solutions when the agent has a) a sufficiently accurate model of the environment b) a means to represent various solutions and c) that the verification of the solution is computationally feasible and economical. Similarly, as we will see in the next section, the presence of a useful objective function and access to its gradient, can provide a valid mechanism for learning that may be best described as "search". However, we suggest that an open-ended evolutionary mechanisms can work alongside such gradient-based methods and are best described by "adaptation" rather than "search".

2.2 Fitness as a Scalar but Multi-Objective Criterion

The reward hypothesis is the third tenet of RL that is the subject of scrutiny by Abel et al. (2024). The authors summarize this core principle in RL as the hypothetical sufficiency of reward maximization in defining all goals. They then call for a less dogmatic view that acknowledges instances that escape this narrow definition. They highlight the delineation of five axiomatic assumptions (Bowling et al., 2023) within which the reward hypothesis is unequivocally true, and underscore the existence of circumstances which fall outside its purview. The situations where these axioms are not met involve those where alternative actions are incommensurable or incomparable (Chang, 2015) or fail to meet the independence of irrelevant alternatives axiom, which excludes risk-sensitive objectives (Allais, 1953; Machina, 1982). Here, we argue that evolution offers a method for satisfying multiple objectives using a single simple criteria (survival), and thereby provides a paradigm to go beyond the scalar versus multi-objective dichotomy. Furthermore, we use this opportunity to address two additional and closely related limitations to the reward hypothesis that Abel et al. (2024) did *not* mention: 1) is the inadequacy of the existing RL framework to explain where rewards come from (Singh et al., 2010; Juechems & Summerfield, 2019), and 2) is the relevance of "homeostatic" theories of motivation that have historically competed with the reward optimization view (Keramati & Gutkin, 2014; Juechems & Summerfield, 2019; Verschure et al., 2014).

In the previous section we mentioned that selection of model candidates through a fitness or "quality" function is not equivalent with model selection via an objective (or loss) function, even though both methods provides valid mechanisms for learning. This subtle distinction between these two learning mechanisms was also at the heart of a famous debate between Darwin and Lamarck over the mechanism of evolutionary adaptation itself. While the Lamarckian view was more akin to a gradient-based method driven by objectives, the Darwinian framework acknowledged the sufficiency of random variations as a means to generate adaptive candidates. Just as molecular mechanisms have vindicated both mechanisms as viable strategies for cross-generational adaptation, we consider both learning strategies as viable within individual lifetimes as well. The open-ended evolutionary algorithms that we drew from in the previous section use the term "non-objective" to describe their approach in contrast with the objective-driven optimization approaches that even include older "evolutionary" algorithms (Brant & Stanley, 2017; 2020). Instead of aiming towards a predefined objective, open-ended novelty search operates on a "minimal criteria" (survival until reproductive age) (Lehman & Stanley, 2010) which lacks any notion of a back-propagated error or supervision signal.

But how does the *non-objective* paradigm of (open-ended) evolutionary learning help address the question of scalar versus multi-objective rewards that clouds the reward hypothesis in RL (T3)? In short, an evolutionary account argues that a minimal survival criteria is sufficient to satisfice many

objectives simultaneously, provided that the search process is sufficiently open-ended, and novelty-generating. Each organism appears to have optimized for multiple complex abilities (flight, vision, hearing, etc.), each of which–from a retrospective or intelligent design point of view–would require optimizing on multiple objectives. As discussed before, in a Darwinian (non-Lamarckian) setting, evolution lacks access to such objective criteria and merely operates on a singular minimal criteria, which is survival past the reproductive age.

In RL problems too, there are many objectives that can be identified that aide in the achievement of the ultimate goal. The utility of each of these objectives may vary in degree across time, context, or even depend on history, but multiple objectives can also be simultaneously relevant at any given point in time. The main claim of the scalar reward hypothesis is that it is possible to distill all relevant objectives necessary for a given task, into a single scalar reward value (Silver et al., 2021). Even if the scalar reward hypothesis held true, and there were no proven bounds to its universality (Bowling et al., 2023), the question remains as to where an RL agent gets its reward objectives from. In artificial RL, the source of these objectives are ultimately rooted in the design decisions of the engineers that provide the auxiliary losses and tasks along with the "primary objective" of the task (Jaderberg et al., 2016). Some of these inductive biases are borrowed from psychology and biology of intrinsic motivation (IM) de Abril & Kanai (2018); Pathak et al. (2017); Houthooft et al. (2016); Kim et al. (2018); Pong et al. (2019); Leibfried & Grau-Moya (2019), while others may be generated adversarially (Campero et al., 2020) or even randomly without any a priori interpretable principle in mind (Osband et al., 2016; Eysenbach et al., 2018; Burda et al., 2018). In biological RL, on the other hand, the source of both intrinsic and extrinsically motivated behavior and their corresponding reward functions is ultimately attributed to cross-generational evolutionary mechanisms (Singh et al., 2010). In section 1.2, we demonstrated plausible mechanisms by which such evolutionary processes can take place within a single generation. This offers a plausible meta-learning mechanism by which multiple reward functions can be dynamically provisioned for the RL algorithms *during the lifetime of the organism*. Indeed, Czégel et al. (2021) speculate on the striatal reward system as a candidate location for assigning a scalar *fitness* value in addition to the state-action values of RL.

However, as Abel et al. (2024) point out, the scalar reward hypothesis is only proven sufficient when certain axioms are held (Bowling et al., 2023) and that, both in theory and practice, there is utility in implementing learning with multiple objectives rather than a single reward (Roijers et al., 2013; Abdolmaleki et al., 2020; 2021; Vamplew et al., 2023). There are two primary ways in which the evolutionary view, and specifically, the Darwinian neurodynamics perspective can address the multi-objectivity issue discussed above. First, as was demonstrated by Czégel et al. (2021), each array of replicating neural dynamics can be tied to a different objective, thereby facilitating a competition between multiple objectives, just as there is between replicating neural dynamics within an array itself. The output of each member of each population can be interpreted as the functional form of a different objective that is then utilized by the RL circuitry. This view would be consistent with the "affordance competition" model of Pezzulo & Cisek (2016) where a hierarchy of goals with different levels of abstraction are simultaneously and dynamically being generated during the organisms' lifetime. The second, is that within each "population" of competing representations, two objectives are simultaneously at play, one being the "primary" objective that determines the individual's "fitness", and the second is the metabolic cost of maintaining the ever-increasing population sizes of the neural dynamics that have proven "fit" on the primary objective. In later iterations on the open-ended QD algorithms mentioned before, the incorporation of resource constraints as part of the "minimal criteria" for survival not only made the simulations more faithful to biological evolution, but also significantly improved the efficacy of the evolutionary algorithms (Brant & Stanley, 2020).

The problem of limited resources and physiological constraints more generally, is central to homeostatic theories of motivation and drive which have historically rivaled the reward-centric view on RL Keramati & Gutkin (2014); Juechems & Summerfield (2019); Verschure et al. (2014). These theories are naturally more grounded in biological reality, rather than abstractions of economic theory where the axiomatic assumptions of the reward hypothesis are rooted (Morgenstern et al., 1964). Homeostatic theories of motivation therefore trace the origins of reward in attractors of feedback

control systems that operate over multiple physiological objectives rather than a scalar quantity. In addition to their relevance to our discussion around the multi-objective nature of motivated behavior, these theories serve as a good stepping stone to our discussion on the problem of agency (T1) in the next section. There we will argue that, consistent with these homeostatic theories, an ultimate formal grounding of agency, like reward itself, must be sought in the physiological processes that facilitate the persistence and replication of an organism, despite the second law of thermodynamics.

2.3 Evolution Does Not Explain Agency

In addition to the multi-objective nature of goals, "homeostatic" theories of motivation highlight the role of feedback control mechanisms that regulate the physiological concerns of the organism that more abstract-level goals are subservient to. These physiological concerns can ultimately be traced to the provisioning of necessary material and energetic precursors for metabolic activity that the organism needs to avoid death. Metabolic activity, in turn, is impossible to formally address without recourse to the language of non-equilibrium thermodynamics, and the mathematical relationships that dictate the costs of keeping metabolic activity away from equilibrium (death). In this sense, it is not surprising that a thermodynamic formalism of agency is ultimately what is needed to provide a naturalistic account of goal-directed behavior that escapes the fallacy of a homuncular explanation (Deacon, 2011).

It is tempting to attribute agency to evolution, both because evolution appears to be directed, and that it seems to provide a satisfying explanation as to where the reward functions of an organism come from (Singh et al., 2010). However, the non-teleological nature of evolution is a subtle and largely uncontested characteristic of evolution Dawkins (1986); Gould (2002). More importantly, evolution as a theory, does not have any explanatory purchase in defining the replicative and homeostatic dynamics that facilitate it in the first place. Indeed, the roots of this argument go back to the days of the discoverers of evolutionary theory: Darwin and Wallace. While Darwin expressed doubts that the science of the day was able to conceive of such a theory for the origins of life (Darwin, 1863), Wallace resorted to more homuncular arguments in support of a *"force vitale"* (Wallace, 1912).

The science of thermodynamics have made significant strides since the time of Darwin and Wallace with concepts such as "entropy" being coined around the same time as Darwin's correspondence with Wallace. The relevance of this discipline (and the concept of entropy in particular) to the question of origins of life, inference, and information were made early on with such thought experiments as Maxwell's "demon" (Maxwell, 1871) and Schrödinger's "aperiodic crystal" and "negentropy" (Schrodinger, 1946). In the recent decades these thought experiments have given way to robust quantitative theories of systems that are not at equilibrium (Jarzynski, 1997) with swift applications to questions regarding the thermodynamics of replication (England, 2013), evolutionary adaptation (Perunov et al., 2016) and even sensory-motor control and RL (Hack et al., 2023; Grau-Moya et al., 2017).

While there is a rich literature on the formal correspondence between non-equilibrium statistical physics and inference (e..g, LaMont & Wiggins, 2019; Pachter et al., 2024; Sohl-Dickstein et al., 2015; Zdeborová & Krzakala, 2015), there is a much weaker emphasis on the relevance of these formalisms to the problem of agency in RL specifically. This may be due to the many different functional roles and phenomena (e.g, intentionality, sentience, consciousness, or free will) that the term "agency" is overloaded with. For this reason, we suggest that limiting the scope of our formalism for "agency" to the most minimal living process at the dawn of life may be more productive.

Many theories of origin of life exist, but those that envision a self-replicating template molecule, like an RNA or DNA, are more favored than metabolism-first theories or other alternatives (Mariscal et al., 2019). This is perhaps not surprising given the success of the "central dogma of biology" with the DNA as its central player. However, we argue that this emphasis on such "information-carrying" molecules like the DNA or the RNA, avoid the problem of agency by providing no explanation as to the origins or nature of the system required to read and interpret the purported "information" there. Metabolism-first theories of agency (e.g., Deacon, 2021; Aguilera & Barandiaran, 2024)

close this explanatory gap and offer a more general framework for how an evolutionary process can be bootstrapped in the first place.

Is a naturalistic, thermodynamic theory of agency—as defined by a self-sustaining and replicative process that facilitates evolutionary dynamics—then sufficient to conjure a satisfying theory of agency in humans? No. We do not think that such a minimal definition of agency is sufficient to address the various complications that arise during the multicellular transition, let alone complex animals with a central nervous system. However, we do think that this is a *necessary* prerequisite step for developing such a theory. As for a mathematical formalism of the agent in an RL systems, we argue that insofar as the origin of the reward function relies on evolutionary or homeostatic explanations, a formal account of the agent too, must find its grounding in a necessarily embodied agent with its proverbial skin in the ultimate game: beating the second law of thermodynamics.

We acknowledge that various other theories of learning, perhaps most notably the free energy principle (Friston, 2010), have incorporated this thermodynamic paradigm into their theories of "active inference". However, we hope that the emphasis on the potential role of evolutionary dynamics in the brain directs more focus to the relevance of the thermodynamically constrained processes that spawn such evolutionary dynamics in the first place, whether in a "warm little pond" (Darwin, 1871) or a warm little brain.

3 Practical Implications for RL Research

We have attempted to illuminate the three core tenets of RL under evolutionary light in the hopes that a new perspective on the subject will emerge. But what concrete practical tools can an RL engineer extract from this perspective? Many of the articles we cited, particularly in reference to T2 and T3, were from a body of research that is primarily known as "open-ended evolution" (Standish, 2002; Lehman & Stanley, 2008). This body of work consists of a panoply of concrete learning algorithms that leverage the various evolutionary insights we have tried to highlight here. These include, first and foremost, "abandoning objectives" (Lehman & Stanley, 2011a) in favor of novelty or diversity (Lehman & Stanley, 2011b), the use of minimal criteria to retain selection pressure on "quality" as well as diversity (Lehman & Stanley, 2010), and the coevolution of the environment alongside the agent to learn more complex tasks with lower-complexity policies (Brant & Stanley, 2017). While certain products of this research paradigm are relatively well known in the RL community (e.g., Wang et al., 2019; Stanley et al., 2019), we hope that our deeper dive into this literature will help draw attention to the rich hotbed of ideas from which these advances emerged.

3.1 Implications for Explore-Exploit

There was a conspicuous absence of any mention of the problem of exploration (and exploitation) in the original discussion of the three dogmas in Abel et al. (2024). This is particularly surprising given the prominent role of this subject in RL, both theoretically, and in practice. We believe that the evolutionary treatment of RL, specifically on the adaptational view of learning (T2: section 2.1) and on the multi-objective nature of reward (T3: section 2.1), illuminate the "explore-exploit" issue in ways that the current dominant framing elides.

Exploration and Exploitation as Multiple Objectives The multi-objective nature of the explore-exploit problem is most clearly highlighted by Norman & Clune (2023) who point out that exploration and exploitation as objectives are fundamentally at odds with one another and that neither standard RL, meta-RL nor intrinsic motivation (IM) approaches are able to address this conflict appropriately. These authors distinguish between "coincidental" exploration strategies that these methods use and argue for "sacrificial" strategies that forgo short-term episode reward for longer term information gain. While methods like IM enable partial "sacrificial" exploration, Norman & Clune (2023) note that these strategies come at the cost of complete disengagement with the reward signal (even if temporary), leading to a number of known problems such as the noisy TV problem (Ladosz et al., 2022). The solution offered by Norman & Clune (2023) is not evolutionary in nature;

however, as discussed above, evolutionary learning strategies are adept at facilitating pursuit of multiple objectives. This is particularly true in the case of open-ended evolutionary algorithms where the search for novelty or diversity is in effect an exploratory strategy that continuously takes place alongside the search for high-quality individuals.

Direct and Indirect Exploration as Multiple Objectives The adaptational view on learning offered by an evolutionary framing helps us understand exploration beyond the dichotomy offered by *directed* versus *indirect* (random) exploration. Despite a common misattribution of teleology to evolution, evolution is *not directed* towards a singular predefined goal or "objective" (Stanley & Lehman, 2015). And yet, many criteria, from energy-efficiency, to predictive power, do ultimately determine the fitness of solutions that are discovered. Furthermore, the mechanisms by which variation is generated follow strategies that can be characterized as either "random" and "directed". Random mutagenesis is an instance of the former, while epigenetic or Baldwinian mechanisms for evolution offer examples of the latter. Empirical behavioral research has also lends some support to this idea that a combination of different exploration strategies are simultaneously used by individuals (Binz & Schulz, 2022).

3.2 Implications for Resource-Constrained RL

Any physical implementation of a reinforcement-learning system is necessarily subject to resource limitations. There is a rich literature of models that try to capture the impact of limited capacity on learning (e.g., Kumar et al., 2025; Grau-Moya et al., 2017; Arumugam et al., 2022; Bhui et al., 2021; Gershman, 2020; Binz & Schulz, 2022; Fang & Sims, 2025). The dominant paradigm in modeling resource limitations in RL is to model learning as a *constrained* optimization problem, where the constraint typically takes the form of mutual information between observed state and action variables, reflecting the "complexity" of the policy the agent used (Gershman, 2020) and thereby acting as a proxy for the resources it would need to implement it. These approaches therefore operate within the optimization paradigm of economic decision theory and are therefore subject to limits imposed by the axioms discussed earlier. While many of these model's co-opt Herbert Simon's terminology of "bounded rationality" to describe their models, their methodology and framework is directly at odds with the evolutionary framing that was actually advocated by Simon (1956):

> ...however adaptive the behavior of organisms in learning and choice situations, this adaptiveness falls far short of the ideal of "maximizing" postulated in economic theory. Evidently, organisms adapt well enough to "satisfice"; they do not, in general, "optimize."

This notion of "satisficing" is clearly more aligned with the "non-objective" evolutionary paradigm of learning that we tried to advocate for in contrast to "optimization". The notion of "minimal criteria" (Brant & Stanley, 2017; Lehman & Stanley, 2010) operationalizes this idea. Interestingly, the incorporation of resource limitation, specifically, as part of this minimal criteria (Brant & Stanley, 2020), resulted in a significant boost to the quality of the solutions found. The use of such evolutionary learning algorithms can therefore offer an alternative computational framework for modeling resource limitation in RL. This approach offers a more direct, process-level description of actual resource constraints, in contrast with informational approaches that merely use proxies for metabolic cost via abstract measures like mutual information, KL divergence or "policy complexity" Gershman (2020); Leibfried & Grau-Moya (2019); Zénon et al. (2019). Such process-level models are in a better position to capture the direct role of metabolism in resource-limited and embodied cognitive agents (Jacob et al., 2023).

4 Discussion

Here we have attempted to view the three dogmas of RL as presented by Abel et al. (2024) under a more biological, and specifically, evolutionary light. We started with the adaptational view

that Abel et al. (2024) recommend in place of the search-oriented view of learning. We showed how drawing on the rich understanding of adaptation from evolutionary theory can provide a more elaborate application of the metaphor beyond just the issue of convergence. We then argued that the Darwinian neurodynamic view allows for evolutionary dynamics to pursue different objectives either via the presence of multiple populations or through open-ended evolution of a diversity of solutions through a single minimal criteria. We highlighted that the fitness of each replicator in each one of these lineages can automatically capture at least two objectives: one corresponding to the primary goal, and the other to the resource-efficiency of the computations necessary to produce the solutions. We finally tackled the "environment spotlight" problem that Abel et al. (2024) pointed out, by shining evolutionary light on this problem as well. Diverging from our previous arguments, we argued that evolutionary theory alone is not equipped to provide a stand-alone formalism for this problem. Instead, we argue that a formal account of agency in complex animals and artificial agents must be rooted in a thermodynamic account for the origin of life where the rational use of material and energetic resources to maintain one's embodied self must constitute any formal definition of agency. We concluded by showing how looking at these three "dogmas" through an evolutionary lens can provide a clearer view on a central issue in RL, which is that of balancing exploration and exploitation.

The authors of Abel et al. (2024) mention the importance of re-framing open questions in facilitating shifts in scientific paradigms. It's well acknowledged that paradigms are often defined by the metaphors afforded by the technology of the day. The success of computer technologies in the last century has had such an impact on the language we use to describe everything from physics to biology and cognition. We highlighted the dangers of over reliance on these metaphors when it comes to the problem of agency and the attribution of "information" and "code" to DNA or RNA, without accounting for the origins and maintenance of processes that actually "read" and "execute" these otherwise inert physical artifacts. We would like to emphasize that the danger is not merely in incurring opportunity costs of getting stuck in an old paradigm, but also in conceding to homuncular explanations, which the question of agency is notoriously susceptible to.

Finally, we acknowledge that we may have contributed to a fourth unappreciated dogma, which is not unique to RL but many theories of intelligent behavior: an overt emphasis on learning during lifetimes at the expense of innate faculties that are accrued over cross-generational time spans and inherited and ready for use upon birth (Zador, 2019). We hope that our evolutionary account will encourage the reader to explore more integrated evolutionary-developmental accounts of learning (e.g., Shuvaev et al., 2024; Arita & Suzuki, 2000) that help bridge the two worlds, and redress this bias.

Acknowledgments

MH thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for their support. MH is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. MH is partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645.

The authors also thank the anonymous reviewer for their thoughtful and constructive feedback.

References

Abbas Abdolmaleki, Sandy Huang, Leonard Hasenclever, Michael Neunert, Francis Song, Martina Zambelli, Murilo Martins, Nicolas Heess, Raia Hadsell, and Martin Riedmiller. A distributional view on multi-objective policy optimization. In *International Conference on Machine Learning*, pp. 11–22. PMLR, November 2020.

Abbas Abdolmaleki, Sandy H Huang, Giulia Vezzani, Bobak Shahriari, Jost Tobias Springenberg, Shruti Mishra, Dhruva Tb, Arunkumar Byravan, Konstantinos Bousmalis, Andras Gyorgy, Csaba

Szepesvari, Raia Hadsell, Nicolas Heess, and Martin Riedmiller. On multi-objective policy optimization as a tool for reinforcement learning: Case studies in offline RL and finetuning. *arXiv [cs.LG]*, June 2021.

David Abel, Mark K Ho, and Anna Harutyunyan. Three dogmas of reinforcement learning. *arXiv [cs.AI]*, July 2024.

Miguel Aguilera and Xabier E. Barandiaran. Thermina: A minimal model of autonomous agency from the lens of stochastic thermodynamics. In *ALIFE 2024: Proceedings of the 2024 Artificial Life Conference*, ALIFE, pp. 121. MIT Press, 07 2024. DOI: 10.1162/isal_a_00826. URL https://doi.org/10.1162/isal_a_00826.

M Allais. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica*, 21(4):503, October 1953.

Takaya Arita and Reiji Suzuki. Interactions between learning and evolution: The outstanding strategy generated by the baldwin effect. In *Artificial Life VII*, pp. 196–205. The MIT Press, August 2000.

Dilip Arumugam, Mark K Ho, Noah D Goodman, and Benjamin Van Roy. On Rate-Distortion Theory in Capacity-Limited Cognition & Reinforcement Learning. *arXiv [cs.LG]*, October 2022.

Rahul Bhui, Lucy Lai, and Samuel J Gershman. Resource-rational decision making. *Current opinion in behavioral sciences*, 41:15–21, October 2021.

Marcel Binz and Eric Schulz. Modeling human exploration through resource-rational reinforcement learning. *arXiv [cs.LG]*, January 2022.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3003–3020. PMLR, 2023.

Jonathan C Brant and Kenneth O Stanley. Minimal criterion coevolution. In *Proceedings of the Genetic and Evolutionary Computation Conference*, New York, NY, USA, July 2017. ACM.

Jonathan C Brant and Kenneth O Stanley. Diversity preservation in minimal criterion coevolution through resource limitation. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, GECCO '20, pp. 58–66, New York, NY, USA, June 2020. Association for Computing Machinery.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv [cs.LG]*, October 2018.

W H Calvin. The cerebral code: Thinking a thought in the mosaics of the mind. 1998.

Andres Campero, Roberta Raileanu, Heinrich Küttler, Joshua B Tenenbaum, Tim Rocktäschel, and Edward Grefenstette. Learning with AMIGo: Adversarially Motivated Intrinsic Goals. *arXiv [cs.LG]*, June 2020.

Ruth Chang. Value incomparability and incommensurability. https://philarchive.org/rec/CHAVIA, 2015. Accessed: 2025-7-13.

Dániel Czégel, Hamza Giaffar, Márton Csillag, Bálint Futó, and Eörs Szathmáry. Novelty and imitation within the brain: a darwinian neurodynamic approach to combinatorial problems. *Sci. Rep.*, 11(1):12513, June 2021.

Dániel Czégel, Hamza Giaffar, Joshua B Tenenbaum, and Eörs Szathmáry. Bayes and Darwin: How replicator populations implement Bayesian computations. *Bioessays*, pp. 2100255, 2022.

Charles Darwin. Letter to joseph dalton hooker, February 1 1871. URL https://www.darwinproject.ac.uk/letter/?docId=letters/DCP-LETT-7471.xml. Published in: Darwin, F. (Ed.), *The Life and Letters of Charles Darwin*, Vol. 2, London: John Murray, 1887, p. 202.

Charles R. Darwin. Letter to the athenæum. *The Athenæum*, (1852):554, April 25 1863. Discusses origin of life and spontaneous generation.

Richard Dawkins. *The Blind Watchmaker*. W. W. Norton & Company, New York, 1986. ISBN 9780393315707.

Ildefons Magrans de Abril and Ryota Kanai. A unified strategy for implementing curiosity and empowerment driven reinforcement learning. *arXiv [cs.AI]*, June 2018.

Terrence W. Deacon. *Incomplete Nature: How Mind Emerged from Matter*. W. W. Norton & Company, New York, 1st edition, 2011. ISBN 978-0393049916. Cover date: November 21, 2011.

Terrence W Deacon. How molecules became signs. *Biosemiotics*, 14(3):537–559, December 2021.

Theodosius Dobzhansky. Nothing in biology makes sense except in the light of evolution. *Am. Biol. Teach.*, 35(3):125–129, March 1973.

George Dragoi. The generative grammar of the brain: a critique of internally generated representations. *Nat. Rev. Neurosci.*, November 2023.

G M Edelman. Neural darwinism: selection and reentrant signaling in higher brain function. *Neuron*, 10(2):115–125, February 1993.

Jeremy L England. Statistical physics of self-replication. *J. Chem. Phys.*, 139(12), September 2013.

Douglas H Erwin. Macroevolution of ecosystem engineering, niche construction and diversity. *Trends Ecol. Evol.*, 23(6):304–310, June 2008.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv [cs.AI]*, February 2018.

Zeming Fang and Chris R Sims. Humans learn generalizable representations through efficient coding. *Nat. Commun.*, 16(1):3989, April 2025.

Anna Fedor, István Zachar, András Szilágyi, Michael Öllinger, Harold P de Vladar, and Eörs Szathmáry. Cognitive architecture with evolutionary dynamics solves insight problem. *Front. Psychol.*, 8, March 2017.

Chrisantha Fernando, Eörs Szathmáry, and Phil Husbands. Selectionist and evolutionary approaches to brain function: a critical appraisal. *Front. Comput. Neurosci.*, 6:24, April 2012.

Karl Friston. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.*, 11(2):127–138, February 2010.

Samuel J Gershman. Origin of perseveration in the trade-off between reward and complexity. *Cognition*, 204:104394, November 2020.

Samuel J Gershman, Petra E M Balbi, C Randy Gallistel, and Jeremy Gunawardena. Reconsidering the evidence for learning in single cells. *Elife*, 10:e61907, January 2021.

Stephen Jay Gould. *The Structure of Evolutionary Theory*. Harvard University Press, Cambridge, MA, 2002. ISBN 9780674006133.

Jordi Grau-Moya, Matthias Krüger, and Daniel A Braun. Non-equilibrium relations for bounded rational decision-making in changing environments. *Entropy*, 20(1), December 2017.

P Hack, C Lindig-Leon, S Gottwald, and D A Braun. Thermodynamic fluctuation theorems govern human sensorimotor learning. *Sci. Rep.*, 13(1):869, January 2023.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. *arXiv [cs.LG]*, May 2016.

Michael Jacob, Judith Ford, and Terrence Deacon. Cognition is entangled with metabolism: relevance for resting-state EEG-fMRI. *Front. Hum. Neurosci.*, 17:976036, April 2023.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *arXiv [cs.LG]*, November 2016.

Christopher Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690–2693, 1997. DOI: 10.1103/PhysRevLett.78.2690. URL https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.78.2690.

Niels K Jerne. The generative grammar of the immune system. *Scand. J. Immunol.*, 38(1):2–8, July 1993.

Keno Juechems and Christopher Summerfield. Where does value come from? *PsyArXiv*, April 2019.

Mehdi Keramati and Boris Gutkin. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *Elife*, 3:e04811, December 2014.

Hyoungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, and Hyun Oh Song. EMI: Exploration with mutual information. *arXiv [cs.LG]*, October 2018.

Saurabh Kumar, Henrik Marklund, Ashish Rao, Yifan Zhu, Hong Jun Jeon, Yueyang Liu, and Benjamin Van Roy. Continual learning as computationally constrained reinforcement learning. *arXiv [cs.LG]*, June 2025.

Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Inf. Fusion*, 85:1–22, September 2022.

Colin H LaMont and Paul A Wiggins. Correspondence between thermodynamics and inference. *Physical Review E*, 99(5):052140, 2019.

J Lehman and Kenneth O Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evol. Comput.*, 19(2):189–223, June 2011a.

Joel Lehman and Kenneth O Stanley. Exploiting open-endedness to solve problems through the search for novelty. pp. 329–336, August 2008.

Joel Lehman and Kenneth O Stanley. Revising the evolutionary computation abstraction: minimal criteria novelty search. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, New York, NY, USA, July 2010. ACM.

Joel Lehman and Kenneth O Stanley. Evolving a diversity of virtual creatures through novelty search and local competition. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pp. 211–218, New York, NY, USA, July 2011b. ACM.

Felix Leibfried and Jordi Grau-Moya. Mutual-Information Regularization in Markov Decision Processes and Actor-Critic Learning. *arXiv [cs.LG]*, September 2019.

Mark J Machina. "expected utility" analysis without the independence axiom. *Econometrica*, 50 (2):277, March 1982.

Samir W Mahfoud. Niching methods for genetic algorithms. October 1996.

Carlos Mariscal, Ana Barahona, Nathanael Aubert-Kato, Arsev Umur Aydinoglu, Stuart Bartlett, María Luz Cárdenas, Kuhan Chandru, Carol Cleland, Benjamin T Cocanougher, Nathaniel Comfort, Athel Cornish-Bowden, Terrence Deacon, Tom Froese, Donato Giovannelli, John Hernlund, Piet Hut, Jun Kimura, Marie-Christine Maurel, Nancy Merino, Alvaro Moreno, Mayuko Nakagawa, Juli Peretó, Nathaniel Virgo, Olaf Witkowski, and H James Cleaves, 2nd. Hidden concepts in the history and philosophy of origins-of-life studies: A workshop report. *Orig. Life Evol. Biosph.*, 49(3):111–145, September 2019.

James Clerk Maxwell. *Theory of Heat*. Longmans, Green, and Co., London, 1871. URL https://archive.org/details/theoryheat00maxwgoog.

Oskar Morgenstern, John Von Neumann, Harold William Kuhn, and Ariel Rubinstein. *Theory of games and economic behavior*. J. Wiley and Sons, 1964.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv [cs.AI]*, April 2015.

Ben Norman and Jeff Clune. First-explore, then exploit: Meta-learning intelligent exploration. *arXiv [cs.LG]*, July 2023.

F John Odling-Smee, Kevin N Laland, and Marcus W Feldman. *Niche construction: The neglected process in evolution (MPB-37)*. Monographs in Population Biology. Princeton University Press, Princeton, NJ, July 2003.

John Odling-Smee, Douglas H Erwin, Eric P Palkovacs, Marcus W Feldman, and Kevin N Laland. Niche construction theory: a practical guide for ecologists. *Q. Rev. Biol.*, 88(1):4–28, March 2013.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. *arXiv [cs.LG]*, February 2016.

Jonathan Asher Pachter, Ying-Jen Yang, and Ken A Dill. Entropy, irreversibility and inference at the foundations of statistical physics. *Nat. Rev. Phys.*, pp. 1–12, May 2024.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. *arXiv [cs.LG]*, May 2017.

Nikolay Perunov, Robert A Marsland, and Jeremy L England. Statistical physics of adaptation. *Phys. Rev. X*, 6(2):021036, June 2016.

Giovanni Pezzulo and Paul Cisek. Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends Cogn. Sci.*, 20(6):414–424, June 2016.

Vitchyr H Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. Skew-Fit: State-Covering Self-Supervised Reinforcement Learning. *arXiv [cs.LG]*, March 2019.

Justin K Pugh, Lisa B Soros, and Kenneth O Stanley. Quality diversity: A new frontier for evolutionary computation. *Front. Robot. AI*, 3:40, July 2016.

D M Roijers, P Vamplew, S Whiteson, and R Dazeley. A survey of multi-objective sequential decision-making. *J. Artif. Intell. Res.*, 48:67–113, October 2013.

Erwin Schrodinger. What is life? *J. Philos.*, 43(7):194, March 1946.

Sergey Shuvaev, Divyansha Lachi, Alexei Koulakov, and Anthony Zador. Encoding innate ability through a genomic bottleneck. *Proc. Natl. Acad. Sci. U. S. A.*, 121(38):e2409160121, September 2024.

David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artif. Intell.*, 299(103535):103535, October 2021.

H A Simon. Rational choice and the structure of the environment. *Psychol. Rev.*, 63(2):129–138, March 1956.

Satinder Singh, Richard L Lewis, and A Barto. Where do rewards come from? In *Proceedings of the International Symposium on AI Inspired Biology - A Symposium at the AISB 2010 Convention*, pp. 111–116, 2010.

Jascha Sohl-Dickstein, Eric A Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *arXiv [cs.LG]*, March 2015.

Russell K Standish. Open-ended artificial evolution. *arXiv [nlin.AO]*, October 2002.

Kenneth O. Stanley and Joel Lehman. *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer, Cham, Switzerland, 2015. ISBN 978-3-319-15523-2. URL https://link.springer.com/book/10.1007/978-3-319-15524-9.

Kenneth O Stanley and Risto Miikkulainen. Evolving neural networks through augmenting topologies. *Evol. Comput.*, 10(2):99–127, June 2002.

Kenneth O Stanley, Jeff Clune, Joel Lehman, and Risto Miikkulainen. Designing neural networks through neuroevolution. *Nature Machine Intelligence*, 1(1):24–35, January 2019.

András Szilágyi, István Zachar, Anna Fedor, Harold P de Vladar, and Eörs Szathmáry. Breeding novel solutions in the brain: a model of darwinian neurodynamics. *F1000Res.*, 5:2416, September 2016.

Peter Vamplew, Benjamin J Smith, Johan Källström, Gabriel Ramos, Roxana Rădulescu, Diederik M Roijers, Conor F Hayes, Friedrik Hentz, Patrick Mannion, Pieter J K Libin, Richard Dazeley, and Cameron Foale. Scalar reward is not enough. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pp. 839–841, Richland, SC, May 2023. International Foundation for Autonomous Agents and Multiagent Systems.

Paul F M J Verschure, Cyriel M A Pennartz, and Giovanni Pezzulo. The why, what, where, when and how of goal-directed choice: neuronal and computational principles. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 369(1655):20130483, November 2014.

Alfred Russel Wallace. The origin of life. a reply to dr. schäfer. Wallace–Online (Letter S700), 1912. URL https://wallaceletters.myspecies.info/S700. Letter to the editor defending the idea of a "vital force" in the origin of life (S700).

Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O Stanley. Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions. *arXiv [cs.NE]*, January 2019.

Anthony M Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. Commun.*, 10(1):3770, August 2019.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *arXiv [cond-mat.stat-mech]*, November 2015.

Alexandre Zénon, Oleg Solopchuk, and Giovanni Pezzulo. An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, 123:5–18, February 2019.