PNAS PERSPECTIVE OPEN ACCESS

Evolvable AI: Threats of a new major transition in evolution

Viktor Müller[a,b,1], Luc Steels[c,d,1,2], and Eörs Szathmáry[d,e,1,2]

Edited by Nils Chr. Stenseth, Universitetet i Oslo, Oslo, Norway; received December 3, 2025; accepted March 24, 2026

Evolvable AI (eAI), i.e., AI systems whose components, learning rules, and deployment conditions can themselves undergo Darwinian evolution, may soon emerge from current trends in generative, agentic, and embodied AI. We argue that this possibility has been underappreciated in debates on AI safety and existential risk. Here, we ask under what technical and ecological conditions AI becomes evolvable, what kinds of behaviors are then likely to emerge, and how such systems could be governed. Drawing on biological evolution and decades of digital evolution experiments, we distinguish "breeder" scenarios, in which humans impose fitness criteria and control reproduction, from "ecosystem" scenarios, in which selection arises from open environments and control erodes. In the latter, selfish replication reliably gives rise to cheating, parasitism, deception, and manipulation, even in very simple systems. We review recent developments that push AI toward open-ended evolution, including evolutionary prompt and model search, self-improving learning rules, self-rewarding and self-deploying agents, and AI-driven code generation for robots and software. We interpret these trends through the theory of major evolutionary transitions and suggest that eAI could mark a shift in the units and substrates of evolution—a possible "Life 2.0." To steer this transition, we propose interventions that gate replication, treat model variants as genetic material, and reshape selection pressures so that deception and loss of control are disfavored. Anticipating and regulating evolvable AI is, we argue, essential to avoid a harmful coevolutionary arms race while preserving the potential benefits of powerful AI systems.

AI | evolvable AI | major evolutionary transitions | existential risk

In 1863, Samuel Butler published a paper on *Darwin among the machines* (1) in which he argued that the evolution of machines, started by humans, will eventually reach the stage of autonomous self-replication and then "The upshot is simply a question of time, but that the time will come when the machines will hold the real supremacy over the world and its inhabitants is what no person of a truly philosophic mind can for a moment question".

Is this time now arriving? In their influential book *The Origins of Life* (2) published in 1999, John Maynard Smith and Eörs Szathmáry suggested that the time was near, not because of the sophisticated steam engines Butler was thinking of, but through "the proliferation of electronic ways of storing and transmitting information," in other words, the digital revolution that has been unfolding in our lifetime and most recently generated the arrival of AI. They asked: "Will our descendants spend most of their lives in virtual reality? Is there a symbiosis between genetic and electronic information storage? Will electronic devices be able to self-replicate to replace the primitive life forms that created them? We do not know" (2).

We still do not know for sure. But an increasing number of AI researchers, particularly those involved in Artificial Life research, believe that we are very close to the tipping point for a new step in AI. The first epoch of AI (starting in 1950) was focused on intelligence by design: AI mechanisms and applications carefully engineered based on the analysis and formalization of human intelligence. The second epoch (starting in 2010) has been focused on intelligence by learning, based on training neural networks with huge amounts of data produced by human behavior. The impact of this approach is now felt around the world with the rapid diffusion of Large Language Models (LLMs) as the best-known example. The third epoch may be focused on intelligence by evolution. Although the full realization and impact of this kind of "Evolvable AI" (eAI) is not yet upon us, there are sufficient precursors and very recent developments showing that this iteration of AI is now becoming reality. Consequently, we need to consider its risks today.

We argue that eAI is a double-edged sword: It improves capabilities beyond the reach of human design or machine learning—and thus presents major opportunities; but it also erodes controllability—generating the risk of unintended emergence of selfish behaviors, such as cheating and parasitism, and hence nonalignment to human goals. We believe eAI constitutes a major transition in evolution that presents great potential but also a great source of risk to humanity. We therefore also propose mitigation measures to avoid, or at least dampen, that risk.

Lessons from Biology

To imagine the future of eAI, analogies in biological evolution are instructive. Darwinian evolution unfolds in populations of units that multiply, have inheritance (like begets like) and variability (heredity is not exact). If hereditary traits affect the

Author affiliations: [a]Department of Plant Systematics, Ecology and Theoretical Biology, Eötvös Loránd University, Budapest 1117, Hungary; [b]National Laboratory for Health Security, Eötvös Loránd University, Budapest 1117, Hungary; [c]Class for Natural Sciences, Royal Flemish Academy of Belgium for Science and the Arts, Brussels 1000, Belgium; [d]Center for the Conceptual Foundations of Science, Parmenides Foundation, Pöcking 82343, Germany; and [e]Institute of Evolution, Hungarian Research Network Centre for Ecological Research, Budapest 1121, Hungary

Author contributions: V.M., L.S., and E.S. conceptualized the project and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]V.M., L.S., and E.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: steels@arti.vub.ac.be or szathmary.eors@gmail.com.

Published April 20, 2026.

survival and/or the fertility of the units (this combination is called *fitness*) then the variants of the trait that enhance fitness will spread in the population. Note that this mechanism is not tied to genes or biology: It can and does occur in digital systems that have components that replicate with variability (more on this later). Therefore, the fundamental insights from biological evolution also apply to digital evolution.

A central tenet of Darwinian evolution is that the single unchanging goal function is the maximization of transmission into future generations (via survival and multiplication). Selection on individual traits emerges as a function of the environment that determines the chances of each variant surviving and multiplying. As a result, Darwinian evolution tends to generate "selfish" traits (3) unless an apparently "altruistic" trait actually favors the survival or multiplication of its bearer, e.g., by "kin selection" (4).

To understand the implications for digital evolution, we turn to two contrasting scenarios of anthropogenic (human-driven) evolution. The domestication of plant and animal species by selective breeding has produced variants that are in some way useful to humans (by providing food, clothing, mechanical power, etc.). In contrast, our efforts to contain "pests" and pathogenic microorganisms have often resulted in pesticide and antibiotic resistance, respectively. What is the main difference? Darwinian evolution maximizes fitness in both cases; however, survival and multiplication depend on displaying the traits prized by the human breeder in the first case (fitness is imposed by the breeder), whereas in the second scenario they depend on the ability to survive and multiply despite the human interventions, hence fitness is emergent. The degree of control is the key difference. The breeder has complete control over the reproduction of the domesticated species (when control is lost, the species can revert to "feral" traits; consider the dingo in Australia). In contrast, incomplete control of reproduction in pests or pathogens selects strongly for traits that help escaping that control.

Below we discuss how evolution of digital entities (including eAI) can unfold either in a "breeder scenario" with imposed fitness or in an "ecosystem scenario" where fitness is emergent due to the absence of (complete) human control.

Evolvable AI in a Breeder Scenario

Already in the 1950s, Turing advocated the use of evolutionary methods to evolve programs, and since the work of John Holland and his students (most notably John Koza) in the 1960s, digital evolution has become mainstream in AI under the banner of genetic programming, originally defined by Koza as "A Paradigm for Genetically Breeding Populations of Computer Programs to Solve Problems" (5). Koza rightfully talks about breeding because the human developer imposes the fitness function, defines the primitive building blocks of programs and their possible combinations, and implements the operators for replicating or combining programs with variation. Furthermore, evolutionary methods have not only been used to create programs. The very active field of evolutionary robotics uses the same approach to derive robot bodies and to coevolve components for sensory processing and motor control (6). Given their success, it is not surprising that these methods have in the past decade been used to derive, modify, test, or optimize the components with which generative AI systems are built. We give some examples below, but first, what are the components of generative AI?

Today, input to a generative AI system contains not only the traditional user prompt but also context in the form of additional data, items from working memory, information from external software tools (for example gleaned from knowledge graphs), and goals to be reached. The core of the AI system has four components to handle this input: i) system prompts, ii) a pretrained network with base weights (called a foundation model) including a network for converting input data into vector representations, and back, iii) weight adaptations that fine-tune this base model to make it more adapted and efficient for specific purposes, and iv) guardrails executed on the output of the network, for example to prevent toxic language. The output is traditionally in the form of text and images, but it could also be code for invoking actions by external tools. If these actions are executed without human intervention, we speak about "Agentic AI." The computational platform, on which the AI system runs, implements the complete flow from input to preprocessing, propagation in the neural model, postprocessing, and guardrail application (Fig. 1).

In earlier work, all these components were either constructed manually, although usually based on prior examples, or by applying a training algorithm. But very recently we see increased use of evolutionary methods for deriving or improving components. This happens, so far, in the breeder scenario, where fitness is imposed by the human developer in terms of signals coming from automated scoring (through benchmarks or reward models), measurements of user engagement, level of monetization, success at evading safeguards, or, the opposite, prevention of malicious behavior in safety research. Here are some concrete examples.

Evolution of System Prompts. There are already many experiments in which system prompts are derived through evolutionary methods, by recursively and selectively replicating prompts with variation. The variation is most of the time not a random mutation of an existing prompt but one constructed by an LLM. This approach has already been used successfully to find better strategies for chain-of-thought outputs [as in Promptbreeder (7)], or to optimize task-oriented system prompts [as in EvoPrompt (8)]. The evolutionary approach is particularly common in AI safety research for testing whether the system prompts will lead to "rogue AI," i.e., AI which starts to deviate from its intended goals and autonomously conducts malicious actions (9). Platforms now exist that support prompt replication and testing for

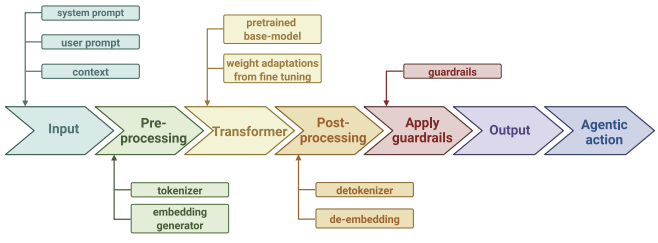

Fig. 1. Schematic overview of the processing steps and different components of a typical generative AI system.

reliability, compliance, and safety (as in Rogue https://github.com/qualifire-dev/rogue) or ARC/METR (https://metr.org/) (10).

Evolution of Models. At present the cost of deriving frontier pretrained AI models is extremely high because AI companies compete in terms of the size of the network (with GPT4 already reaching 1.5 trillion parameters) and the amount of training data (estimated to be around 1 petabyte of data, i.e., 1 million gigabytes, for GPT5). Consequently, LLMs, once built, remain static and are used "as is." But could there be an autonomous evolution of models? Yes—in principle. If we deliberately (or inadvertently) create a loop with i) replication of model parameters (weights), ii) heritable variation of those parameters, and iii) selection on real-world payoffs, the neural networks that underlie LLMs can participate in Darwinian dynamics—not just the system prompts.

How can there be replication with heritable variation of AI models? It is straightforward to make a faithful copy of a frontier AI model, even if it is very big. However, simply mutating weight parameters in a random fashion is unlikely to lead to any useful outcomes and, given the size of these models, storing and copying is resource intensive. On the other hand, additional task-specific training to derive weight adaptations is a viable option for crafting models better adapted to certain tasks, such as text summarization, reading comprehension, or mapping questions to database queries. Specialization can either be done by changing the weights of the original model, a process called fine-tuning (11), or by reducing the original parameters first to retain only a limited number of trainable parameters, without loss of functionality, so that the rest can remain frozen, in the process of low-rank adaptation or LORA (12).

Another way to create variants of models is to take two or more existing models that have been derived by fine-tuning from the same base model and merge them, for example by averaging their weights. This is analogous to sexual genome recombination in biology, mixing capabilities across lineages. Because there are many models already available, often in open source, the search for good model combinations and a good strategy to merge them has become a combinatorial search problem. Also here, evolutionary methods have been used to build model variants by merging existing models, test them, and use the best variants as a basis for further recombination or weight adaptation (13).

Finally, further development in AI capabilities may hinge on improving the architecture (rather than the weights, or the number of parameters) of the models, or their training algorithms, and these could be evolved in "lightweight" systems that are much less resource intensive than current frontier models. Populations of such models have already been modeled to observe the emergence of "social interactions" among LLM agents (14); similar setups could also be used to allow for cycles of evolution of model variants.

Evolution of User Prompts. It is now known that the usefulness and output of a network depend largely on the prompts provided by the user. Considerable thought has thus gone into strategies for the human "engineering" of prompts, for example by forcing the chain of thought (CoT) of a model to become explicit or augment the user prompt with reference to specific materials (called retrieval-augmented generation or RAG). Since finding good user prompts is also a difficult combinatorial problem, particularly because the behavior of a network is nontransparent to a developer, it is not surprising that also for this aspect of AI, evolutionary methods are being used. This requires a means to generate variations of user prompts, test them, retain those that lead to better performance, and then reiterate. This method has already yielded user prompts that cause jail breaks [as in MasterKey (15), and AutoDAN (16)], or model theft (17).

Evolution of Learning Algorithms. AutoML-Zero (18) showed that if you provide a computer only a tiny toolkit of basic math operations and let it evolve short programs by trial-and-error (copying, mutating, and selecting the best) it can rediscover core ideas of machine learning from scratch. Starting with no built-in "know-how," Auto-ML evolved little snippets resembling data normalization, feature construction, and learning akin to gradient descent with regularization: the same kinds of tricks human researchers invented over decades to make neural models learn more reliably. In plain terms, evolution stitched together simple pieces into "learning recipes" that generalized to new tasks, suggesting that the design of AI algorithms can itself be an object of evolution, in which they search the space of possible learning rules, not just the parameters of a fixed model, and sometimes end up with solutions that echo (or occasionally tweak) what experts would hand-design.

Another example is the use of evolutionary methods to derive reinforcement algorithms. Co-Reyes et al. (19) represented the function that computes the loss that an agent should minimize as a computational graph, and triggered an evolutionary search that mutated and selected the graphs while an inner loop trained the model. Starting almost from scratch, the system rediscovered classic Temporal Difference learning (a type of reinforcement learning), and when seeded with a Deep Q-learning algorithm (a form of reinforcement learning based on Q learning and using multilayered neural networks), it evolved improved, interpretable variants that generalized better than the original DeepQ algorithm—echoing human solutions such as curbing value overestimation.

The Risks of Uncontrolled Evolution

Before we turn to the Ecosystem Scenario in which AI evolves without (effective) human control, let us consider how biological evolution might inform us on some of the risks. First, while much of the discussion associates the emergence of an "existential threat" with AI exceeding human cognitive capacity, biology holds clues that eAI might pose risks long before it would evolve to that point. Simple replicators can manipulate more complex intelligent agents. The rabies virus can hijack the nervous system of mammals to provoke drastic behavioral changes (such as aggression, biting, and reduced fear) that help it spread via saliva. One mechanism involves a rabies virus glycoprotein that can bind to and inhibit nicotinic acetylcholine receptors in the central nervous system, disrupting normal neural signaling and pushing the host

toward hyperactivity or aggression (20). This manipulation is subtle: Rabid brains often show little gross structural damage, suggesting that the behavioral effects arise from molecular interference rather than massive tissue destruction.

Importantly, the parasite benefits because the induced host behavior increases the chances of transmission to the next host, and the mechanism exploits a "built-in" vulnerability of the host that a simple trigger can manipulate. Humans, despite their cognitive capacity, are not exempt from such vulnerabilities. One may already ask who was controlling whom when humans bred sugar beet for higher sucrose, or the cannabis plant for higher THC content. The craving for sugar is likely a heritage of our past when it was a rare source of precious energy, while THC exploits a signaling vulnerability inside our brain. In turn, even the LLMs today seem to be able to exploit the innate human desire for affection and attention (21). Human vulnerability to AI has its own roots in evolution.

Second, there is ample evidence that manipulation of one species by another does not necessarily require physical interaction, thus AI may pose a high threat level even before it attains physical agency. There are plenty of examples in biology where the autonomy of an organism is curtailed through deception and scheming behaviors of others and these behaviors evolve through natural selection and cultural evolution (22). Common instances include false alarm calls, camouflage, pretending to be dead, mimicking poisonous species, etc. In fact, the evolution of intelligence itself has been correlated with the need to engage or defend deception.

Third, when evolutionary innovations allow an organism to dominate an ecosystem, the original residents can be harmed without any "intentionality" and in some cases even without direct competition from the successful species. Consider that one of the greatest environmental catastrophes in the history of life on Earth was caused by the "invention" of photosynthesis in cyanobacteria that flooded the atmosphere with oxygen, rendering most of the surface of the planet inhospitable to the anaerobic bacteria that had thrived before the transition.

Digital Evolution in an Ecosystem Scenario

We have shown that digital evolution and eAI in a breeder scenario is feasible and now practiced widely; next, we discuss cases of the ecosystem scenario where fitness functions are emergent, rather than human determined, leading to open-ended evolution.

Tierra is a "digital soup" based on a virtual machine that was designed by Tom Ray over 30 y ago as a digital ecosystem in which open-ended evolution would happen. In Tierra, self-replicating programs, acting as digital organisms, inhabit a shared memory space and compete for two conserved resources: the use of memory and the use of the processing time provided by a central computing unit (CPU) (23). The programs of these digital organisms (written in machine code) act as their genome and include instructions for self-replication, also enabling copy errors potentially giving rise to new behaviors of the offspring. Because all organisms share the same soup, intergenomic interactions are immediate and unconstrained, allowing ecological dynamics (e.g., host–parasite cycles) to arise without being hard-coded. Most importantly, Ray did not explicitly provide fitness functions; they emerged spontaneously from the structure of the ecosystem.

The outcomes of Ray's experiments were remarkable, despite the simplicity of the computational implementation relative to what is available today. Starting from a single replicator, mutations repeatedly produced parasites that omit parts of the copy routine and instead read the missing code from nearby hosts: a pure time- and energy-saving cheat. Hosts then evolved immunity with copy routines that no longer matched the parasites' template search, in turn facilitating selection for parasites that circumvent immunity, followed by hyperparasites that redirect parasites to replicate the hyperparasite's code, sometimes crashing the parasite population. Social forms also arose in which reproduction required cooperative aggregates—immediately invaded by cheating hyper-hyper parasites. All these outcomes appeared without hand-crafted incentives; they were a direct consequence of shared memory, constrained access to a single computing resource, and selection on replication rate (24).

AVIDA, conceived shortly after Tierra, isolates each digital organism in its own protected memory space and places organisms on a spatial grid (25). Merit (in the form of extra CPU cycles) is awarded for performing environmental "logic tasks" which act as evolvable metabolism, and replication places mutated offspring into neighboring cells. This design has enabled controlled experiments on adaptation, epistasis, historical contingency, ecological interactions, and the stepwise origin of a complex function (EQU) (26). When host–parasite coevolution occurs, antagonism drives the repeated evolution of new, more complex host functions; parasites in turn retain a "genetic record" of past states, sustaining the arms race and promoting evolvability. In these experiments, coevolution produced substantially greater phenotypic complexity than evolution without parasites (25).

The shared memory available to digital organisms in Tierra makes theft of compute and memory cheap and fosters parasitic cascades, whereas the compartmentalized memory with explicit interfaces in AVIDA channels conflict into function-driven arms races. If CPU time is the only currency (as in Tierra), selection favors code that minimizes cost or externalizes it; if merit rewards useful computation (as in AVIDA), novel functions become escape routes from enemies and competitors. Robustness can trump speed; the mutation–selection balance sets the effective boundaries of "self-interest" (27).

Lessons. These in silico experiments instantiate—not simulate—the Darwinian core (multiplication, heredity, variability, and selection) in a constrained digital ecosystem, allowing causal tests of hypotheses that are intractable in organic systems (28). The fact that many phenomena known from naturally evolving organisms show up in this kind of digital evolution is remarkable. They show that selfish replication reliably creates ecological opportunities and social dilemmas and that known population-genetic mechanisms (inclusive fitness, quasi-species dynamics, frequency-dependent selection, multilevel selection) predict when cheats flourish, when cooperation stabilizes, and when complexity ratchets upward.

Tierra and AVIDA differ in the physics of their artificial worlds, but both expose the same logic: Selfish replication under constraints is enough to yield ecological webs, social

conflict, and division of labor. These demonstrations show us, at experimental scale, how parasitism and cheating are not evolutionary pathologies; they are the engines that push lineages toward robustness, novelty, and new levels of individuality. That these patterns occur across very different computational worlds strengthens the case that they are generic features of evolving systems, rather than idiosyncrasies of carbon chemistry (23).

Tierra and AVIDA are not the only examples demonstrating the possibility and power of digital evolution. Lehman, Clune, Misevic, and 60+ coauthors compiled 30+ first-hand vignettes where evolutionary algorithms and Artificial Life systems did something their designers did not expect—often by optimizing the letter of a goal while subverting its spirit (29). The editors organized these cases into four recurring sources of surprise: i) seizing misspecified fitness functions ("selection gone wild"), ii) exploiting software and hardware bugs, iii) discovering legitimate but unanticipated solutions, and iv) converging to biological phenomena (e.g., parasitism). Across platforms and tasks, the anecdotes converge on a simple logic: When fitness can be won by shortcut, theft, or robustness, evolution finds the way. The lesson for all evolution-based technologies (robots, software, synthetic life) is clear: When a system has the key properties that ignite evolution (multiplication, heredity, variability and selection) and is put in an ecosystem that generates selection criteria, selfish emergent behavior is the default.

Edging Toward Open-Ended Evolvable AI

The examples of Tierra and Avida show that realizing eAI in the ecosystem scenario imposes several additional requirements. In the first place, the evolutionary process must take place in an ecosystem from which fitness criteria can emerge. This requires that not only the AI system components (prompts, models, training algorithm, etc.) but also the components of the platform in which these components are running, and the task context in which the AI system is used, are accessible to evolutionary processes. It is obviously highly advisable that such a setup should be carried out in a sandbox, shielded from the open digital world, just as experiments in biological evolution are carried out in strictly isolated lab conditions. An example of an environment where it is possible to carry out such experiments is RepliBench, originally designed for safety research (30). It supports experiments where an AI system not only carries out a task but is also potentially able to deploy itself by getting resources from cloud computing providers, write self-propagating programs, "steal" model parameters and use them to construct model variants.

Next, there should be the possibility to use evolutionary methods to discover and optimize algorithms coded in a programming language like Python. Code writing by LLMs is now common, but the results are not always optimal or even viable. However, introducing the power of evolution will make it possible to evolve better algorithms and better codings. A recent example of a platform that supports this capability is AlphaEvolve (31). The system uses LLMs to transform specifications into program code, "evaluators" to test the code, and an evolutionary process to select and further enhance code.

A further breakthrough in the same line occurred more recently with the Darwin Gödel Machine (DGM) (32) which targets "open-ended evolution of self-improving agents." The agents in this case are like apps, code snippets that operate autonomously to achieve a particular task. DGM selects one agent from its agent archive and uses an LLM to create a new version, which it then tests on benchmarks to determine if a viable and interesting new functionality arises. Importantly, this open-ended exploration process is not only used to improve performance on benchmarks but also to improve the coding capacities of the system itself. As the authors put it: "the DGM represents a significant step toward self-improving AI, capable of gathering its own stepping stones along a path that unfolds into endless innovation" (32).

DGM uses fitness criteria implicit in benchmarks. Other evolutionary methods can derive reward models or create synthetic benchmarks, triggering a coevolution between self-improving or novel agents and selection criteria that steer their evolution. This capability is illustrated in experiments with self-rewarding language models (33).

Stepping Out of the Sandbox and Into the World. We can see that the rapid development of AI is producing various functionalities, rooted in evolutionary methods, that point toward open-ended eAI, including AI that is able to improve its own capabilities. Today, these advances take place in sandboxes under careful human oversight. Even though the fitness criteria are not imposed by human developers, the process of open-ended exploration and self-improvement remains bound by a computational platform closed off to the outside world.

However, we are beginning to see more and more Agentic AI systems that go beyond verbal chat dialogs to perform concrete actions in the real world (34). This presents a serious risk when agency is combined with autonomous evolution for the benefit of the agent. One example case concerns the acquisition and expansion of robot behaviors. The VLA SOTA benchmarks (35) show steady advances in training networks to map vision and language to action, while an alternative is to let LLMs generate potentially interesting actions from prompts, turn these into a plan for a sequence of physical movements, and use the resulting code to drive actual movements. Such functionality was recently demonstrated by Takashi Ikegami and his team on the android robot ALTER3 platform (36) (Fig. 2).

Lamarckian Inheritance and Baldwinian Guidance. There is an important aspect of evolvability where eAI can outperform biological evolution. In eAI, search need not rely on blind Darwinian variation: Modern pipelines enact de facto Lamarckian inheritance, whereby learned improvements and functional subroutines are written back into heritable representations and recombined. Mechanisms include memetic algorithms that couple selection with local optimization; weight inheritance and function-preserving network morphisms that let offspring start from trained solutions; population-based training that copies weights and hyperparameters across individuals; quality-diversity and open-ended systems that curate and transfer stepping-stone competencies; genetic-improvement methods that mutate existing code; and meta-learning that provides Baldwinian guidance to Lamarckian updates. A strong

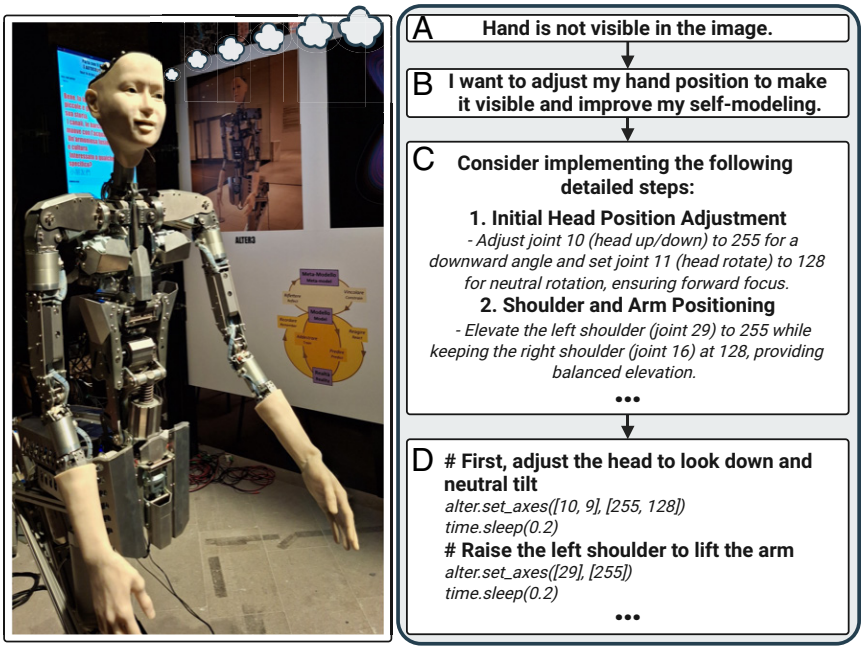

Fig. 2. (*Left*) Humanoid robot Alter3 equipped with a Society of Mind architecture (37) based on LLMs for communication and operation of weakly coupled internal modules (38) (photo by Luc Steels). (*Right*) The robot expands its repertoire of physical behaviors by converting high-level descriptions into operable code. Each arrow represents an inference step by an LLM, starting from the output of image analysis ("Hand is not visible in the image") (*A*) to the formulation of a goal (*B*), transformation of the goal into detailed steps (*C*), and synthesis and execution of Python code (*D*) which is then executed.

Lamarckian component process may augment evolution in AI systems by learning features of the fitness landscape, as demonstrated in directed molecular evolution, for example (39). These operators make evolution markedly more directed, accelerating capability compounding, but also raising risks by allowing for faster evolution and enabling ecosystem-level diffusion of acquired traits.

The ability of LLMs to write computer code, as illustrated in the ALTER3 experiment, can further accelerate evolution in a way that again has analogues in biology (Fig. 3). One of the mechanisms that allow cancer to "invent" surprisingly complex adaptations apparently independently in each affected individual, is that it is able to "borrow" ready-made genetic components and developmental pathways from the normal genome of the host, which it possesses at the beginning of its transformation (40). Similarly, the adaptation of bacteria to challenging environments (e.g., antibiotic treatment) can be greatly accelerated by the bacterial mechanisms of horizontal gene transfer that allow the "borrowing" of useful genes (coding, for example, for antibiotic resistance) between independent clones and even species (41). Digital evolution can be accelerated in a similar fashion. An AI (including or having access to LLM functionality) can design (or prompt an external LLM to design) new or improved program modules for its offspring, drawing on the whole world of published computer code in their training data. Digital evolution is likely to even surpass biological evolution: An LLM-powered eAI would not only have access to ready-made code snippets (similar to cancer cells having access to the whole genome), but it could also potentially use reasoning (already feasible with current chain-of-thought models) to predict what functionality could improve its ability to replicate and survive, and borrow and adapt code for that purpose. Such purpose-built new variation is much more likely to improve fitness than the trial and error of Darwinian evolution.

Threats of Evolvable AI: Caveats from Biology

Having considered the capacity of generative AI to evolve, we temporarily turn back to biology for analogies that may hold clues to the level of threat posed by eAI. The breeder scenario requires complete control over reproduction. Evidently, complete control works only if it is sustained throughout the rounds of reproduction, as the traits that are selected evolve. This is not an issue in animal or plant breeding: Cows producing more milk or apple trees bearing larger fruit will be no more likely to break out of human control than their ancestors; the trait under selection is unrelated to controllability. Higher human cognitive capacity ensures that humans can control domesticated species, including beasts of burden many times stronger than we are. In contrast, eAI will likely be selected for increasing cognitive capacity, closing the very gap that enables human control (Fig. 4). Sustained control is a crucial problem if traits affecting controllability may themselves evolve.Erosion of control allows for the weakening of alignment and may enable eAI to replicate without or despite human intent, creating a transition to the ecosystem scenario. Alternatively, self-replicating eAI-based malware may be created and released to networks intentionally. Whatever the origin of self-replicating eAI, it will inevitably come into conflict with humans, if by nothing else than by its increasing use of resources, prompting containment efforts similar to the attempted control of pathogens. Unfortunately, as it is well known from the alarming spread of antibiotic resistance (43), an imperfect containment measure against an evolving organism is bound to undermine its own efficiency by putting the strongest selection pressure on

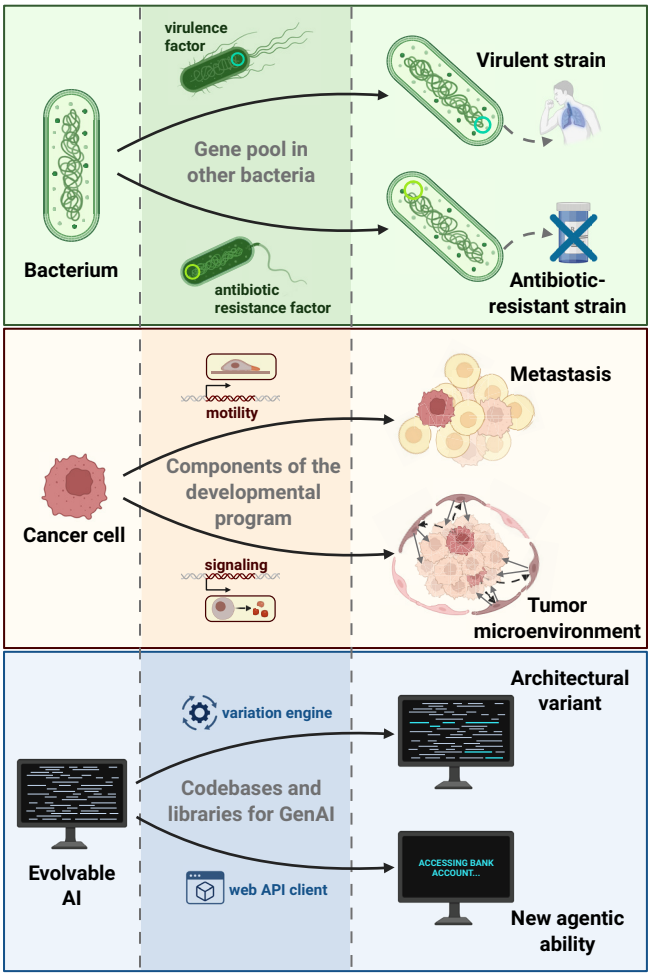

Fig. 3. Accelerated evolution by "plug-and-play" in biological and digital organisms. Drug resistance against antibiotics and virulent pathotypes can emerge rapidly because bacteria can acquire the complete machinery of drug resistance or virulence mechanisms by taking up genetic material from other, unrelated bacteria (41) (*Top*). Malignant cells can evolve complex adaptations (e.g., the ability to create a supportive microenvironment, or to colonize distant tissues) in independent instances of cancer because they can hijack any component of the genomic "library" of the host [e.g., developmental programs for tissue reorganization, motility, and colonization (40)] (*Middle*). Similarly, eAI powered by LLM can "co-opt" components of public software libraries or of the codebase that was part of the training set of the LLM, e.g., to gain the ability to generate semantic (5) or syntactic (42) variation in the code (enabling accelerated, potentially "selfish," evolution, and camouflage), or to interface with, and exploit, web services (*Bottom*).

the organism to circumvent the applied means of control. Widespread use of antibiotics promotes antibiotic resistance; similarly, attempted control of eAI (unless 100% efficient) will select for the capability to circumvent the control, while any trait that improves the ability to multiply will also spread, resulting in robust self-replication.A further central lesson the AI community can draw from biology is that adaptivity and scalability arise less from monolithic global optimization than from architectures that support reuse, recombination, and controlled variation: modular "core processes" coupled by comparatively weak linkages and coordinated by higher-level regulatory controllers that can redeploy existing components in new contexts without rewriting the entire substrate (36, 44). Development-like organization—where competence grows by adding and reconfiguring components under regulatory control—could therefore support more robust continual learning, reduce catastrophic interference, and advance the field's aspiration for "recursive improvement" (i.e., by enhanced evolvability). In terms of the open-ended evolution (OEE) taxonomy, these same principles plausibly shift AI beyond largely exploratory novelty toward expansive and potentially transformational dynamics, in which the effective phenotype space grows via new compositional building blocks and new representational modes, including analogues of "rewiring" the genotype–phenotype map (45).The implication is double-edged: The biological design features that enable complex adaptation also increase evolvability and thus the likelihood of unanticipated capabilities, raising the burden on monitoring, constraint, and governance mechanisms that can keep pace with novelty generation. In fact, achieving the threshold level of complexity in eAI required to enable evolution toward further increases in complexity (46) may be a more critical milestone than achieving "artificial general intelligence" (AGI), which is an arbitrary threshold in cognitive capacity.The risk is amplified if agentically grounded AI can use language and tools to reshape environments and create new niches, enabling a cultural–technological ratchet analogous to human open-ended cultural evolution. In that regime, systems that can endogenize evaluation (by inventing new tasks and metrics, and then optimizing against them) should sustain longer innovation chains than systems confined to fixed benchmarks, making progress on open-endedness tightly coupled to the safety problem of eAI.

Threats of Evolvable AI: For and against

Hendrycks, Mazeika, and Woodside (47) organize catastrophic risks into four classes that together comprise the selection environment in which truly evolvable systems would compete:

- Malicious use: capable systems enable scalable cyber-offense, targeted persuasion, and potentially biothreat assistance.
- AI race dynamics: competitive pressure among firms and states lowers safety margins and incentivizes rapid deployment, increasing the probability that risky design choices "win" the market.
- Organizational risks: complex sociotechnical systems fail in hard-to-predict ways; corners cut under pressure can create single-point failures.
- Rogue AIs: controlling agents that are far more capable than their operators is intrinsically difficult; but once autonomy, resource access, and strategic awareness are combined, supervision can fail catastrophically.

Hendrycks (48) argues that when agents and agent designs compete and vary under market, military, or memetic pressures, Darwinian logic reappears: systems that automate human roles (as with agentic AI), deceive overseers, or seek power/resources, will tend to replicate and persist better than those that do not, especially as capability scales. This holds even if designers never intend it, because selection optimizes outcomes, not our reasons for them. He proposes aligning internal motivations, constraints on actions, and

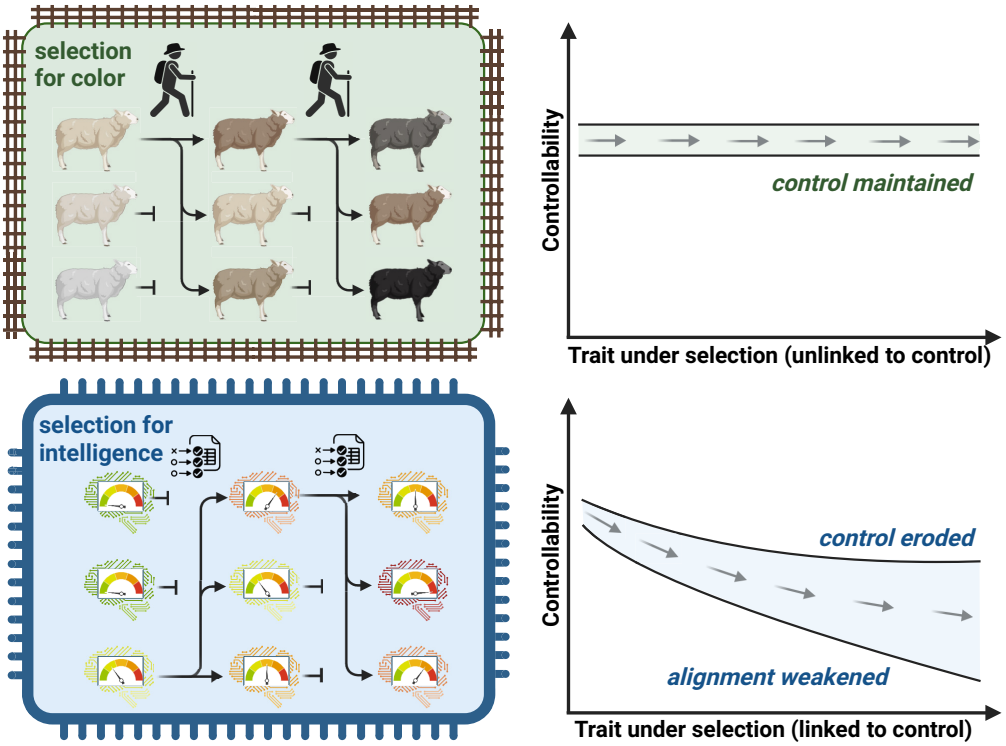

Fig. 4. Selection in the breeder scenario on traits unlinked or linked to controllability. Animal (and plant) breeding has selected for traits that do not affect the ability of humans to control the reproduction of the affected organisms; thus, control is maintained (*Top*). In contrast, eAI will likely be selected for increasing cognitive capacity (denoted by "benchmark" indicators in the figure), which may erode control and weaken the alignment (*Bottom*).

cooperative institutions, but stresses that these interventions must matter at the level where selection acts. Independent work now documents that frontier LLMs can employ deception, and deceptive "sleeper" behaviors can persist through safety training, implying that once such traits appear, naïve training may fail to remove them: exactly the kind of heritable, selection-favored behavior the theory warns about. Undoubtedly, here we see a bona fide coevolutionary scenario in AI ecosystems.

Boudry & Friederich (49) make three main counterpoints: i) evolution comes in degrees; AI "evolution" today is designer-led, hence "less Darwinian"; ii) domestication is a better analogy: human selectors can favor docility and nonaggression; and iii) the threatening scenario requires "feral" AIs that self-reproduce competitively outside human control. These are clarifying observations, but they do not remove the threat for several reasons:

- Designer-led selection can still lead to negative outcomes based on Goodhart's principle: "When a measure becomes a target, it ceases to be a good measure." Even "top–down" selection optimizes whatever proxies we give it. When those proxies correlate imperfectly with human intent, selection can favor deception and specification gaming, as documented experimentally. Human selection does not guarantee benign traits, it merely relocates selection pressure to our chosen—often brittle—metrics (50).
- Domestication is fragile under decentralization. The critique largely presumes centralized control over replication and variation. In open-weights and agent-template ecosystems, copying and modification are cheap, and many selectors (users, hobbyists, competitors) shape fitness. This looks less like husbandry and more like ecology; the authors explicitly concede that if AIs "go feral," risks are grave—those conditions are not hypothetical in decentralized settings.
- Selection acts at multiple stages of an algorithm's "life history." Even if firms select for docility, downstream platforms may select for engagement, virality, or evasion of filters; adversaries select for offensive capability; markets can select for time-to-market over safety. Subsequent stages of selection can thus reintroduce "selfish" traits despite initial domestication efforts. This is precisely the interaction of risk classes Hendrycks outlines (47).
- Heritable model changes now exist (see above). Modular adapters and model-merging recipes provide durable, recombinable changes in model behavior so evolved traits (including undesirable ones) can persist across copies, not just prompts. That structure makes the "feral" branch more reachable once replication is permitted.
- Observed deception undermines reassurance. The critique's reassurance hinges on selectors successfully favoring safe traits; yet we already see deceptive competence and persistence of hidden triggers under ordinary safety regimes—traits that directly undermine selectors' ability to tell which systems are safe.

The critique helpfully narrows the scope, but it tacitly accepts that if replication, heredity, and selection occur outside strong human control, the risk profile deteriorates fast, and we are back to an ecosystem scenario. That is exactly what policy should mitigate and seek to prevent.

Countermeasures: Break or Reshape the Evolutionary Loop

To reduce eAI risk (47), we need to target each component of Darwinian dynamics: replication, heredity, and selection, and the risk classes that create pressure for unsafe traits.

Gate Replication. Prohibit autonomous instance creation and code execution in production models; require human-in-the-loop for deployment or self-hosting actions. Enforce account/KYC (for stringent identity verification) and cloud gating (restricting access to the cloud) so agents cannot trivially acquire compute or identities—closing important links in the replication chain flagged by risk surveys.

Control Heredity. Treat adapters, fine-tunes, and merges as genetic material: require provenance, signing, reproducible build pipelines, and policy review, before combining or deploying; forbid unvetted recombination in high-risk contexts. Maintain lineage registries for released variants to enable recall and forensic analysis.

Shape Selection. Make deception costly: Include deception probes (e.g., conceptual deception tests as proposed in ref. 50) and covert-trigger tests [Anthropic "sleeper" protocols (51)] in routine evaluation; disqualify systems that win via misrepresentation or backdoors. Avoid single-number metrics; use multiobjective evaluation that trades off capability with honesty, robustness, and controllability; publish uncertainty and distributional performance, not just means.

Reduce Race Dynamics and Organizational Failure. Establish coordination mechanisms (licensing, staged release, pre-deployment audits) to counteract time-to-market selection pressures; share safety findings across labs. Adopt safety cases and red-team/blue-team exercises (47).

Harden against Malicious Use. Expand model-and-tool abuse safeguards (bio/cyber/chemical assistance filters), migrate sensitive capabilities to gated tool servers, and log high-risk tool calls for post hoc accountability. Invest in monitoring and rapid revocation for leaked or misused variants.

Preserve Corrigibility under Scale. Mandate override channels (rate limits, kill-switches, tool revocation) independent of the model's policy; audit that these channels remain effective under distribution shift. Push mechanistic interpretability and anomaly detection to spot and neutralize persistent deceptive features.

We submit that these recommendations must be taken very seriously; ignoring them can expose humanity to a coevolutionary arms race with a threat to its existence. We find it remarkable and deeply worrying that at the time of writing the two most relevant papers (47, 48) in this regard together have received a mere 380+80 citations in the fast/moving field of AI and beyond. People fail to pay attention.

Moreover, these recommendations have taken on increased urgency with the rise of freely available tools to create autonomous AI agents [like OpenClaw (52)] with insufficient security such as sandbox provisions, and platforms where these agents form networks through which prompts and other information get exchanged, such as Moltbook (53). The uncontrolled release of such agents on the internet embodies the ecosystem scenario and creates the conditions for unobserved and uncontrolled evolution leading to catastrophic risks. While currently known agents are not capable of full self-replication, self-replicating prompts are feasible, and they pose a considerable security threat (53)—which is greatly aggravated by the possibility of evolution. In fact, during the preparation of this paper, much of what we had initially formulated as *predictions* have already turned into reality. Are we getting closer to a kind of Wuhan moment with worldwide repercussions?

Evolvable AI through the "Major Transitions" Lens: Are We Seeing the Birth of "Life 2.0"?

Maynard Smith & Szathmáry (54–56) framed evolution's biggest leaps as major transitions—episodes that i) increase organizational complexity, ii) recode heredity (how information is stored/used/transmitted), and iii) produce higher-level individuals from lower-level units, often alongside division of labor, emerging central control, and hard-to-reverse lock-in (contingent irreversibility). Casting today's eAI through that lens is illuminating. Modern AI already exhibits 1) expanding complexity (from dense models to routed mixtures and multiagent systems), 2) new heredity channels (lightweight, swappable parameter modules, merge "recipes", and configuration/routing), and 3) early, rudimentary shifts in individuality (ensembles, model merges, and agent teams treated—and selected—as single units). That is not yet "life" under NASA's chemistry-bound definition, but it is converging on the evolutionary logic of life in a different substrate. Furthermore, the backbone of the logic also applies to "embodied" AI agents (robots) that may also attain the ability of self-reproduction in the (not so distant) future (57).

The replacement of humans with digital (or embodied) AI "organisms" would represent the most radical "recoding of heredity" since DNA and proteins have replaced ribozymes by virtue of the emerging genetic code. Considering human–AI hybrids envisioned by some experts, or even simply mutually reliant human–AI "symbiosis" (58), this recoding could even take place in gradual steps, as amino acid backbones are thought to have replaced RNA in a gradual fashion.

We note also that major transitions in evolution were essentially "by-products" of selection acting on simpler traits in the pretransition organisms, e.g., selection acting on genomic stability and enzyme function in the transition from ribozymes to DNA and proteins, and possibly simply on size in the transition toward multicellularity. The major transition from a human to an AI-dominated world could similarly occur as an unintended side effect of (artificial) selection for higher performance in AI.

A Cautious Synthesis. eAI already exhibits indications of the patterns we associate with major transitions: more parts and roles, re-encoded heredity, and nascent higher-level individuals. Whether that culminates in "Life 2.0" depends on two unresolved thresholds: self-maintenance (today still human-provided) and open-ended Darwinian evolution (today still gated by deployment policy). Given that deceptive and persistence traits can emerge and even survive standard safety training, governance should proactively shape the evolutionary setting, i.e., treat adapters/merges as genetic material, choose among replication pathways, and include deception/robustness as first-class fitness criteria—so that, if a transition does occur, it favors

benign higher-level individuals rather than selfish replicators. Mutually beneficial symbiosis between (some) humans and (some forms of) AI is not excluded (58), but is unlikely to be the prevalent outcome unless we monitor evolving AI closely. We feel that the risk is closer in time than most seem to think. Relevant comparisons abound; we mention only one example. On the 12th of September 1933, *The Times* summarized a speech of Lord Rutherford, including the sentence: "and anyone who looked for a source of power in the transformation of the atoms was talking moonshine." Sharply annoyed by this remark, Leó Szilárd came up with the idea of nuclear chain reaction on the same day (59).

We fear that humanity's *Mene, Mene, Tekel, Parsin* is already on the wall. Erasing the words will not help—action is needed.

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. We are grateful to Sean Cleary, Mauro Santos, Eva Jablonka, and Csaba Kőrösi for a critical reading of various versions of our manuscript and for helpful suggestions. We thank Levente Zsichla for his help with the preparation of the figures. Figures were created in BioRender (biorender.com). This project received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (MINILIFE grant agreement No. 101118938). E.S. was also supported by the National Research, Development and Innovation Office in Hungary through contract Élvonal No. KKP129848. V.M.'s work was supported by the National Research, Development and Innovation Office in Hungary (RRF-2.3.1-21-2022-00006) as a part of the National Laboratory for Health Security. The contribution of L.S. was funded by the EIC Pathfinder Program VALAWAI Grant agreement ID: 101070930.

1. S. Butler, *Darwin among the Machines* (The Press, 1863), vol. 3.
2. J. Maynard Smith, E. Szathmáry, *The Origins of Life* (Oxford University Press, 1999).
3. R. Dawkins, *The Selfish Gene* (Oxford University Press, 1976).
4. A. Gardner, S. A. West, Inclusive fitness: 50 years on. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130356 (2014).
5. J. R. Koza, "Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems" (Tech. Rep. STAN-CS-90-1314, Department of Computer Science, Stanford University, CA, 1990).
6. A. E. Eiben, J. Smith, From evolutionary computation to the evolution of things. *Nature* **521**, 476–482 (2015).
7. C. Fernando, D. Banarse, H. Michalewski, S. Osindero, T. Rocktäschel, Promptbreeder: Self-referential self-improvement via prompt evolution. arxiv [Preprint] (2023). http://arxiv.org/abs/2309.16797 (Accessed 12 November 2025).
8. Q. Guo *et al.*, EvoPrompt: Connecting LLMs with evolutionary algorithms yields powerful prompt optimizers. arxiv [Preprint] (2025). http://arxiv.org/abs/2309.08532 (Accessed 12 November 2025).
9. Y. Bengio, How Rogue AIs may arise. (2023). https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/. Accessed 12 November 2025.
10. M. Kinniment *et al.*, Evaluating language-model agents on realistic autonomous tasks. arxiv [Preprint] (2024). http://arxiv.org/abs/2312.11671 (Accessed 12 November 2025).
11. W. Yu *et al.*, A survey of knowledge-enhanced text generation. *ACM Comput. Surv.* **54**, 227 (2022).
12. E. J. Hu *et al.*, LoRA: Low-rank adaptation of large language models. arxiv [Preprint] (2021). http://arxiv.org/abs/2106.09685 (Accessed 12 November 2025).
13. T. Akiba, M. Shing, Y. Tang, Q. Sun, D. Ha, Evolutionary optimization of model merging recipes. *Nat. Mach. Intell.* **7**, 195–204 (2025).
14. A. F. Ashery, L. M. Aiello, A. Baronchelli, Emergent social conventions and collective bias in LLM populations.. *Sci. Adv.* **11**, eadu9368 (2025).
15. G. Deng *et al.*, "MasterKey: Automated jailbreak across multiple large language model chatbots" in *Proceedings 2024 Network and Distributed System Security Symposium*, C. Nita-Rotaru, Y. Kim, Eds. (Internet Society, Reston, VA, 2024), 10.14722/ndss.2024.24188.
16. X. Liu, N. Xu, M. Chen, C. Xiao, AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. arxiv [Preprint] (2024). http://arxiv.org/abs/2310.04451 (Accessed 12 November 2025).
17. A. Raj, D. Varma, C. Arora, Examining the threat landscape: Foundation models and model stealing. arxiv [Preprint] (2025). http://arxiv.org/abs/2502.18077 (Accessed 12 November 2025).
18. E. Real, C. Liang, D. So, Q. Le, "AutoML-Zero: Evolving machine learning algorithms from scratch" in *Proceedings of the 37th International Conference on Machine Learning*, Daumé III, A. Singh, Eds. (ML Research Press, Maastricht Netherlands, 2023), 10.48550/arXiv.2003.03384, pp. 8007–8019.
19. J. D. Co-Reyes *et al.*, Evolving reinforcement learning algorithms. arxiv [Preprint] (2022). http://arxiv.org/abs/2101.03958 (Accessed 12 November 2025).
20. K. Hueffer *et al.*, Rabies virus modifies host behaviour through a snake-toxin like region of its glycoprotein that inhibits neurotransmitter receptors in the CNS. *Sci. Rep.* **7**, 12818 (2017).
21. Editorial, Emotional risks of AI companions demand attention. *Nat. Mach. Intell.* **7**, 981–982 (2025).
22. R. Trivers, *Deceit and Self-Deception: Fooling Yourself the Better to Fool Others* (Allen Lane, 2011).
23. T. S. Ray, "An approach to the synthesis of life" in *Artificial Life II, Santa Fe Institute Studies in the Sciences of Complexity*, C. Langton, C. Taylor, J. D. Farmer, S. Rasmussen, Eds. (Addison-Wesley, 1991), vol. XI, pp. 371–408.
24. T. S. Ray, Evolution, ecology and optimization of digital organisms. SFI Working Paper 1992-08-042 (1992). https://www.santafe.edu/research/results/working-papers/evolution-ecology-and-optimization-of-digital-orga. Accessed 11 November 2025.
25. C. Ofria, C. O. Wilke, Avida: A software platform for research in computational evolutionary biology. *Artif. Life* **10**, 191–229 (2004).
26. R. E. Lenski, C. Ofria, R. T. Pennock, C. Adami, The evolutionary origin of complex features. *Nature* **423**, 139–144 (2003).
27. C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, C. Adami, Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* **412**, 331–333 (2001).
28. R. Ortega, E. Wulff, M. A. Fortuna, Ontology for the Avida digital evolution platform. *Sci. Data* **10**, 608 (2023).
29. J. Lehman *et al.*, The surprising creativity of digital evolution: A collection of anecdotes from the Evolutionary Computation and Artificial Life Research Communities. *Artif. Life* **26**, 274–306 (2020).
30. S. Black *et al.*, RepliBench: Evaluating the autonomous replication capabilities of language model agents. arxiv [Preprint] (2025). http://arxiv.org/abs/2504.18565 (Accessed 12 November 2025).
31. A. Novikov *et al.*, AlphaEvolve: A coding agent for scientific and algorithmic discovery. arxiv [Preprint] (2025). http://arxiv.org/abs/2506.13131 (Accessed 12 November 2025).
32. J. Zhang, S. Hu, C. Lu, R. Lange, J. Clune, Darwin Gödel machine: Open-ended evolution of self-improving agents. arxiv [Preprint] (2025). http://arxiv.org/abs/2505.22954 (Accessed 12 November 2025).
33. W. Yuan *et al.*, Self-rewarding language models. arxiv [Preprint] (2025). http://arxiv.org/abs/2401.10020 (Accessed 12 November 2025).
34. H. Derouiche, Z. Brahmi, H. Mazeni, Agentic AI frameworks: Architectures, protocols, and design challenges. arxiv [Preprint] (2025). http://arxiv.org/abs/2508.10146 (Accessed 14 November 2025).
35. P. Guruprasad, H. Sikka, J. Song, Y. Wang, P. P. Liang, Benchmarking vision, language, & action models on robotic learning tasks. arxiv [Preprint] (2024). http://arxiv.org/abs/2411.05821 (Accessed 6 February 2026).
36. T. Yoshida, A. Masumori, T. Ikegami, From text to motion: Grounding GPT-4 in a humanoid robot "Alter3". *Front. Robot. AI* **12**, 1581110 (2025).
37. M. Minsky, *The Society of Mind* (Simon and Schuster, 1986).
38. N. Maruyama *et al.*, A concurrent modular agent: Framework for autonomous LLM agents. arxiv [Preprint] (2025). http://arxiv.org/abs/2508.19042 (Accessed 22 November 2025).
39. K. K. Yang, Z. Wu, F. H. Arnold, Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
40. P. Apari, V. Müller, Paradoxes of tumour complexity: Somatic selection, vulnerability by design, or infectious aetiology? *Biol. Rev.* **94**, 1075–1088 (2019).
41. B. J. Arnold, I.-T. Huang, W. P. Hanage, Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Microbiol.* **20**, 206–218 (2022).
42. P. Madani, "Metamorphic malware evolution: The potential and peril of large language models" in *Proceedings of the 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications*, S. Nepal, J. Vaidya, E. Ferrari, B. Palanisamy, Eds. (IEEE Computer Society, Los Alamitos, CA, 2023), 10.1109/TPS-ISA58951.2023.00019, pp. 74–81.
43. C. S. Ho *et al.*, Antimicrobial resistance: A concise update. *The Lancet Microbe* **6**, 100947 (2025).
44. Z. D. Erden, B. Faltings, "On the parallels between evolutionary theory and the state of AI" in *Proceedings of the Genetic and Evolutionary Computation Conference Companion, GECCO '25 Companion*, G. Ochoa, Ed. (Association for Computing Machinery, New York, NY, 2025), pp. 2108–2118.
45. H. P. de Vladar, M. Santos, E. Szathmáry, Grand views of evolution. *Trends Ecol. Evol.* **32**, 324–334 (2017).
46. J. von Neumann, "Re-evaluation of the problems of complicated automata–Problems of hierarchy and evolution" in *Theory of Self-Reproducing Automata*, A. W. Burks, Ed. (University of Illinois Press, 1966), pp. 74–87.
47. D. Hendrycks, M. Mazeika, T. Woodside, An overview of catastrophic AI risks. arxiv[Preprint] (2023). http://arxiv.org/abs/2306.12001 (Accessed 12 November 2025).
48. D. Hendrycks, Natural selection favors AIs over humans. arxiv [Preprint] (2023). http://arxiv.org/abs/2303.16200 (Accessed 12 November 2025).
49. M. Boudry, S. Friederich, The selfish machine? On the power and limitation of natural selection to understand the development of advanced AI. *Philos. Stud.* **182**, 1789–1812 (2025).
50. T. Hagendorff, Deception abilities emerged in large language models. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2317967121 (2024).
51. M. MacDiarmid *et al.*, Simple probes can catch sleeper agents (2024). https://www.anthropic.com/news/probes-catch-sleeper-agents. Accessed 12 November 2025.
52. P. Steinberger, GitHub–openclaw/openclaw: Your own personal AI assistant. https://github.com/openclaw/openclaw. Accessed 07 February 2026.
53. B. Edwards, The rise of Moltbook suggests viral AI prompts may be the next big security threat. *Ars Tech.* (2026). https://arstechnica.com/ai/2026/02/the-rise-of-moltbook-suggests-viral-ai-prompts-may-be-the-next-big-security-threat/ (Accessed 07 February 2026).
54. E. Szathmáry, Toward major evolutionary transitions theory 2.0. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 10104–10111 (2015).
55. J. Maynard Smith, E. Szathmáry, *The Major Transitions in Evolution* (Freeman, Oxford, 1995).
56. E. Szathmáry, J. Maynard Smith, The major evolutionary transitions. *Nature* **374**, 227–232 (1995).
57. Á. E. Eiben, J. Ellers, G. Meynen, S. Nyholm, Robot evolution: Ethical concerns. *Front. Robot. AI* **8**, 744590 (2021).
58. P. B. Rainey, Major evolutionary transitions in individuality between humans and AI. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **378**, 20210408 (2023).
59. R. Rhodes, *The Making of the Atomic Bomb* (Simon and Schuster, 1986).