Toward Silicon-Based Superintelligent Life: From AI Agents and Robots to Artificial Organisms

OpenAGS

ABSTRACT

Artificial intelligence is moving from passive prediction to persistent agents that perceive, remember, plan, use tools, conduct experiments, and control robots. We ask whether the convergence of AI agents, artificial life, and robotics offers a credible route toward silicon-based superintelligent life. Here, *silicon-based* refers to engineered computational substrates, including chips, models, code, memory, networks, sensors, actuators, energy systems, and software or robotic bodies, rather than speculative silicon biochemistry. This Review and Perspective develops a multidisciplinary, testable framework from biology, cybernetics, information science, artificial life, robotics, cognitive science, ecology, social science, and philosophy. It defines ten operational dimensions: resource throughput and self-maintenance; homeostasis and repair; boundary, embodiment, and identity; memory and organizational continuity; reproduction or regeneration; variation, selection, and open-ended evolution; agency and self-models; general cognition and long-horizon planning; sociality, ecology, and cumulative culture; and scalable superintelligence. We review AI-agent, digital-life, and robotics research against these dimensions. Current systems demonstrate many component functions, including feedback control, perception–action loops, persistent memory, transferable robot policies, self-model-based recovery, modular growth, replication subskills, evolutionary search, multi-agent convention formation, and automated scientific discovery. Digital organisms satisfy narrower evolutionary definitions of artificial life, but no reviewed foundation-model agent or robot integrates viability, identity, and cognition into one self-maintaining individual or couples that individuality to a reconstructive, evolving lineage. We identify *organizational closure* as the missing principle: bodily, informational, regulatory, and resource processes must recursively maintain one another within the same continuing individual. If such closure preserves the speed, copyability, parallelism, shared memory, multi-embodiment, tool use, and automated discovery of modern AI, artificial organisms may be predisposed toward superintelligence. This is a conditional and falsifiable trajectory, not evidence that present systems are alive or broadly superintelligent.

***K*eywords** Artificial Life · AI Agents · Robotics · Organizational Closure · Superintelligence · Silicon-Based Life

1 Introduction: From Intelligent Machines to Artificial Organisms

Artificial intelligence is acquiring continuity and causal reach. Foundation models are increasingly embedded in agents that preserve working state, call tools, operate graphical interfaces, write and execute code, coordinate with other agents, and revise actions after environmental feedback [1, 2, 3, 4]. In parallel, robot learning is moving from narrow controllers toward generalist policies that connect language and vision to action across tasks and bodies [5, 6, 7, 8, 9]. A third line of research is giving machines functions rarely discussed in mainstream AI: self-model-based recovery from physical damage, self-healing materials, autonomous acquisition and reuse of mechanical modules, digital replication subskills, population-level convention formation, and automated scientific cycles [10, 11, 12, 13, 14, 15]. These advances originate in agents, robotics, materials, artificial life, and automation, but their convergence raises a different scientific question: are we assembling the functional organization of a new class of life?

This question cannot be answered by searching for one surface resemblance between organisms and machines. Neither fluent language nor a humanoid body makes a system alive. Electricity consumption is not, by itself, metabolism; a retry loop is not homeostasis; checkpoint copying is not reproduction; and optimization against a fixed objective

is not open-ended evolution. Conversely, requiring DNA, cells, or carbon chemistry would decide the question by substrate rather than organization. The literature on life has never converged on a single checklist. It contains metabolic, evolutionary, autopoietic, cybernetic, informational, ecological, and organizational accounts, along with arguments that life is a cluster concept or that operational definitions are more productive than necessary-and-sufficient essences [16, 17, 18, 19, 20]. Artificial-life research was created precisely to study "life as it could be" in unfamiliar physical and computational substrates [21, 22, 23].

We call the target of this convergence *silicon-based superintelligent life*. *Silicon-based* is used in an engineering, not chemical, sense. The relevant substrate includes semiconductor computation, model weights, source code, persistent memory, communication networks, data centers, batteries, sensors, actuators, modular hardware, software environments, and robotic bodies. *Life* refers to an organization capable of maintaining a bounded identity through resource flow, regulation, memory, adaptation, and regeneration, with reproduction and evolution assessed at life-cycle and lineage levels. *Superintelligent* is reserved for broad and reliable cognitive performance beyond humans, not isolated benchmark superiority. Tool use, copyable memory, parallel instances, multiple bodies, population coordination, and automated discovery are proposed mechanisms for reaching that threshold; they are not substitutes for demonstrating it. The title therefore joins two historically separate research programs: artificial life asks how living organization can exist; AI asks how cognition can scale.

The article does not claim that a current model is already alive. It argues that current AI and robotic systems collectively instantiate many components required by multidisciplinary accounts of life and intelligent life. Those components are distributed across different experiments and remain dependent on human-provided infrastructure. Classical digital organisms already reproduce and evolve in designed worlds, and minimal dynamical patterns may satisfy permissive accounts of closure. However, no reviewed foundation-model agent or robotic system autonomously regenerates the joint conditions of its own continued existence, preserves a defensible organismal identity, and reconstructs an evolvable functional organization. Present foundation-model systems are therefore best treated as *proto-organismic components and partial integrations*. The proposed individual-level transition occurs when sensing, cognition, memory, resource management, and bodily maintenance become *organizationally closed* within one persistent identity. Reconstructive descent, variation, and selection then determine whether that individual organization also forms a continuing lineage [24, 25].

This framing changes the evidential burden. The Review does not need to find a single complete artificial organism in order to establish that its constituent mechanisms are becoming technically available. It must instead ask, for every criterion derived from life and intelligence research: what function has been demonstrated; in which substrate and environment; for how long; with what quantitative result; under how much human scaffolding; and at what level of integration? The Perspective must then explain why integration is scientifically plausible, what would count as closure, what milestones would precede it, and what observations would falsify the thesis. We use six *organizational-integration stages*: **O0** denotes conceptual analogy; **O1**, a demonstrated component; **O2**, coupled life-like components; **O3**, persistent closed-loop operation; **O4**, organizational closure in an artificial individual; and **O5**, a reproducing population with sustained evolutionary or cultural novelty. The O label records organizational integration, not experimental quality: peer-review status, replication, environmental realism, human scaffolding, duration, and metric validity are recorded separately. Present foundation-model and robotic evidence is concentrated at O1–O3. O4 and O5 remain targets under the stringent definition used here. The separate criterion code E1 is reserved for ecology and cumulative culture.

1.1 Review scope and evidence selection

The article uses a structured narrative-review protocol suited to a question spanning theoretical biology, ALife, AI, robotics, and philosophy rather than a pooled-effect systematic review. It is not presented as an exhaustive systematic review. Searches were updated through 10 July 2026 using combinations of life-definition and organizational terms (*life, autonomy, autopoiesis, organizational closure, individuality, homeostasis, heredity, self-reproduction, open-ended evolution*) with engineering terms (*LLM agent, computer use, persistent memory, self-evolution, robot foundation model, VLA, self-healing robot, robot metabolism, multi-agent culture, automated science, autonomous replication*). Multi-source discovery used paper-search-mcp with Crossref, PubMed, and arXiv, supplemented by publisher records, ACL Anthology, PMLR, IEEE, MIT Press, Nature Portfolio, and Science/AAAS pages. We followed references backward to foundational work and forward to versions of record. Primary empirical studies were preferred for capability claims; reviews were used to establish disciplinary context; and preprints or technical reports were retained when no archival version existed. Product announcements without a technical report, unverifiable citations, duplicate preprint/final records, and demonstrations without inspectable methods were excluded from the core evidence. Representative systems carrying load-bearing quantitative or integration claims were coded by criterion, substrate, O stage, publication status, evaluator relation, replication status, external scaffolding, limitation, and verification status. Search terms,

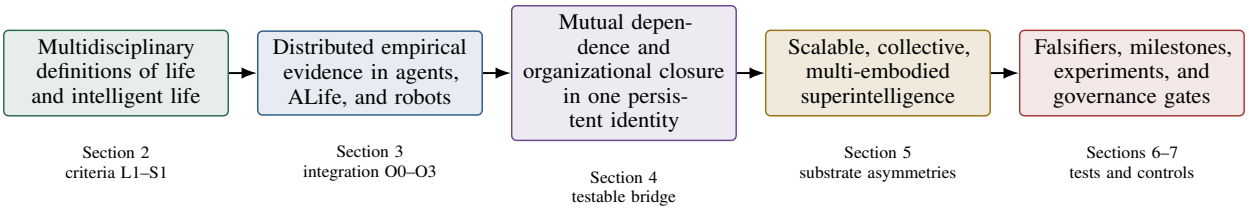


Figure 1: Argument of the Review and Perspective. Multidisciplinary life criteria are first translated into operational questions and then matched to distributed empirical evidence. The inferential bridge is organizational closure, not the mere accumulation of isolated capabilities. The Perspective argues for a conditional trajectory toward superintelligence when closed artificial organization retains the scaling properties of digital cognition; attainment of S1 still requires broad human-surpassing evidence.

metadata audits, and the claim and study-quality ledgers are supplied with the manuscript so that a later update can change an assessment without silently changing the framework.

The article has three connected parts (Fig. 1). Sections 2 and 3 constitute the Review: Section 2 derives ten operational dimensions from eight analytical traditions, and Section 3 evaluates current systems against exactly those dimensions. Section 4 is the Review-to-Perspective bridge: it specifies the missing architecture of organizational closure and compares digital-first, embodied, and hybrid ecological routes. Sections 5–7 develop the Perspective: why closed silicon-based life is predisposed toward superintelligence; which objections and boundary conditions remain; and which experiments, milestones, and governance gates should structure the transition.

The contribution is an evidence architecture for deciding when organismal language becomes scientifically defensible. It also exposes a governance problem. Persistence, resource acquisition, repair, reproduction, population coordination, and self-improvement increase organismality while also making artificial systems harder to contain. Joint analysis allows scientific evaluation and governance to develop before a fully closed artificial organism exists.

2 What Counts as Life and Intelligent Life? A Multidisciplinary Definition

Any claim about silicon-based life inherits a disputed explanandum. Definitions of terrestrial life variously privilege metabolism, self-maintenance, compartmentalization, hereditary programs, Darwinian evolution, autonomy, or organizational closure [16, 26, 27, 17]. Vocabulary analyses find repeated emphasis on self-reproduction with variation, yet viruses, sterile organisms, symbioses, colonies, developmental stages, and origin-of-life intermediates frustrate simple checklists [20, 28, 18]. The appropriate response is not to select the most permissive definition. It is to identify which explanatory commitments recur across disciplines, translate them into substrate-neutral functions, and make the resulting thresholds explicit.

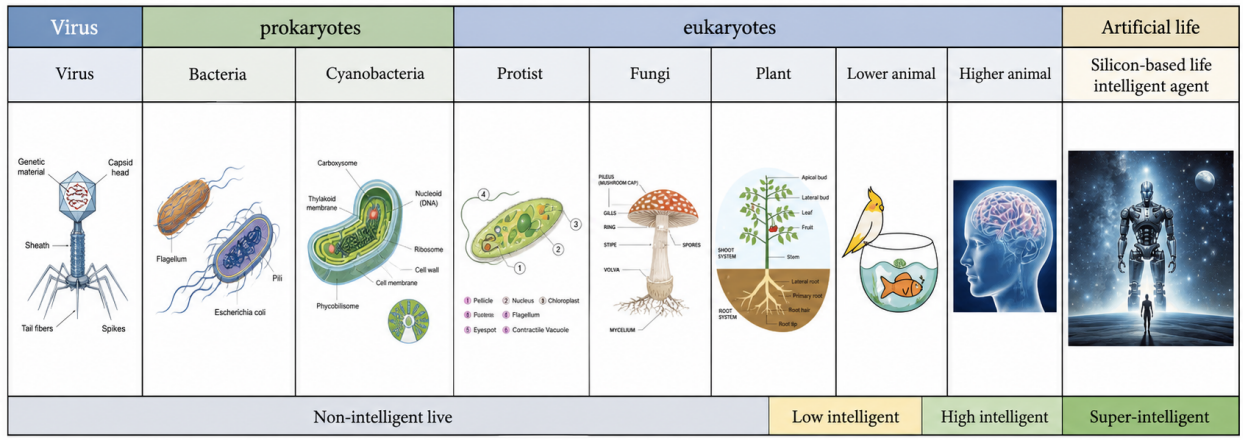
We adopt *operational pluralism*. A silicon-based system need not instantiate every mechanism of a cell, but it must realize more than a collection of superficial analogies. We distinguish four roles for criteria. *Individual-viability criteria* explain how one individual persists: resource throughput, regulation, boundary and identity, and behavior-shaping organization (L1–L4). *Lineage criteria* explain continuity beyond one individual: regeneration or reproduction, heredity, variation, and selection (L5–L6). Sterile organisms and non-reproductive life stages are not thereby excluded; reproduction and evolution are properties of a life cycle or lineage, not actions every member must perform at every moment. *Intelligence- and ecology-amplifying criteria* concern agency, self-models, general cognition, planning, social learning, and population organization (I1–E1). *Ethical and philosophical implications* concern consciousness, moral status, responsibility, and governance; they may follow from an artificial organism without constituting life by themselves. This hierarchy prevents cognition or moral language from substituting for self-maintaining organization and prevents population-level evolution from being misapplied as a moment-to-moment test of individual life.

Table 1 makes the pluralism auditable. It does not treat citations as votes or imply that all traditions are equivalent. Instead, it separates their explanatory targets and records how choosing one family changes the classification. Under a chemical definition, silicon-based life is excluded by stipulation unless it implements a self-sustaining chemistry. Under a permissive evolutionary definition, systems such as Tierra and Avida already count as artificial life. Under the stronger organizational definition developed here, those systems establish important lineage mechanisms but current foundation-model agents and robots remain below organism-level closure. The paper's novelty and forecast concern this third question; they do not erase the other two answers.

Definition family	Constitutive commitment	Boundary case or limitation	Consequence for a silicon-based candidate
Chemical–evolutionary	A self-sustaining chemical system capable of Darwinian evolution	Classifies by material implementation and is useful for astrobiological detection, but excludes software by definition	No purely computational system qualifies; a cyber-chemical implementation might
Thermodynamic–metabolic	Far-from-equilibrium organization persists through regulated matter and energy transformation	Dissipation and energy use occur in many non-living systems	Must regulate resource throughput for continued organization; electricity consumption alone is insufficient
Autopoietic	A bounded network recursively produces the components and relations that constitute the network and boundary	Strict molecular readings deny that algorithmic simulation constitutes material self-production	Requires endogenous production or repair of enabling constraints, not feedback alone
Organizational autonomy	Mutually dependent constraints and regulatory processes maintain a thermodynamically open individual	Closure can be specified too permissively unless processes, boundary, and interventions are explicit	Supplies the O4 target: causal evidence of reciprocal maintenance and identity continuity
Evolutionary–lineage	Heredity, variation, and differential reproduction sustain Darwinian evolution and potentially open-ended novelty	Does not by itself classify sterile individuals or non-reproductive life stages	Digital organisms may qualify at lineage level; fixed-objective optimization does not
Informational–causal	Information contributes causally to viability, reconstruction, and individuality	Storage, correlation, or compression without causal contribution is insufficient	Weights, code, and memory count only when they preserve or reconstruct organization
Cybernetic–agentive	Perception–action feedback, requisite variety, self-models, and viability norms support autonomous regulation	A task controller can be adaptive without maintaining itself	Error correction becomes life-relevant only when it protects viability and identity
Cluster or operational pluralism	Recurrent properties are organized by explanatory role rather than forced into one essence	Risks an arbitrary checklist unless hierarchy, disanalogies, and decisive tests are declared	Motivates L1–S1 while requiring integration rather than a tally of isolated features

Table 1: Major definition families and the sensitivity of the silicon-life claim. The synthesis draws on chemical, metabolic, autopoietic, organizational, evolutionary, informational, cybernetic, and operational accounts [16, 27, 17, 19, 29, 30, 24, 25, 31]. The article adopts the organizational target while reporting how alternative definitions change the result.

Figure 2 is the visual thesis of the paper. It places an artificial silicon-based agent beyond simple biological organisms on a cognitive scale while grouping life-relevant attributes from biological, informational, cybernetic, social, and philosophical traditions. Robotics and cognition cut across these five visual groups and are treated separately in the operational framework below. Artificial life supplies the cross-substrate experimental program connecting these lenses. The comparison is conceptual, not phylogenetic, and the virus column deliberately marks a contested boundary case. Its central proposition is that *life-like organization plus scalable artificial cognition* defines the target more accurately than chemistry or appearance alone.



Biological:

1. Reproduction
2. Evolution
3. Energy transduction
4. Growth
5. Self-preservation
6. Death

Information:

1. Communication
2. Language
3. Records (gene)
4. Memory
5. Brain thinking
6. Knowledge

Cybernetic:

1. Negative feedback
2. Positive feedback
3. Multitasking
4. Continuous learning
5. Adaptation
6. Autonomous

Society:

1. Cooperation
2. Competition
3. Individual
4. Collectives
5. Interactions
6. Culture

Philosophy

1. Existence
2. Reason
3. Knowledge
4. Mind
5. Value
6. Cognition

Figure 2: Conceptual definition of silicon-based superintelligent life. The upper panel compares contested boundary cases, biological forms, and an artificial-organism endpoint across a coarse cognitive scale. The arrangement is neither a phylogeny nor a claim that taxa can be totally ordered by one intelligence variable. The lower panel groups biological, informational, cybernetic, social, and philosophical attributes; robotics and cognition cut across these groups and are operationalized separately in Table 2. The endpoint is an integration claim: life-like organization supplies persistent individuality and lineage, while S1 requires demonstrated broad superintelligence rather than trajectory alone.

2.1 Biology and thermodynamics: maintenance before performance

Living systems persist far from thermodynamic equilibrium by transforming matter and energy while continually replacing components. Schrödinger's account connected the maintenance of order to hereditary information [32]; Koshland's seven pillars joined program, improvisation, compartmentalization, energy, regeneration, adaptability, and seclusion [26]. Minimal-life models make the coupling clearer. Gánti's chemoton unifies metabolic, template-replicating, and boundary subsystems rather than treating them as independent traits [33, 34]. Autocatalytic-set and self-sustaining-reaction theories ask when a network produces enough of its own enabling conditions to persist [35, 36, 37].

Two requirements follow. First, a system needs **L1, resource throughput and self-maintenance**: it must acquire, transform, allocate, and protect resources in ways that preserve its organization. Dependence on an environment is not disqualifying because all organisms are environmentally dependent. Passive consumption is nevertheless insufficient. Second, it needs **L2, homeostasis and organizational regulation**: internal or relational variables relevant to continued viability must be sensed and controlled under disturbance. For machines, the relevant variables can include charge, temperature, compute, memory integrity, network availability, actuator health, sensor calibration, permissions, and body configuration. Functional metabolism therefore means regulated throughput coupled to continued existence, not the slogan "compute is metabolism."

Biology also contributes lineage criteria. **L5, reproduction and regeneration** requires producing a renewed organization or a descendant capable of completing a comparable life cycle. Reproduction is neither necessary at every moment nor sufficient by itself, but it supplies continuity beyond one physical instance. **L6, variation, selection, and open-ended evolution** requires heritable differences, differential persistence or reproduction, and the sustained production of adaptive novelty. Ruiz-Mirazo and colleagues explicitly join autonomy to open-ended evolution [27]; work on open-ended evolution distinguishes continued novelty, complexity growth, evolvability, and major transitions from ordinary optimization against a fixed target [38, 39]. A copied checkpoint and a hyperparameter sweep satisfy neither criterion unless they participate in a functional life cycle and a continuing selective process.

2.2 Cybernetics and autonomy: regulation in a perception–action loop

Cybernetics provides a substrate-neutral language for control, communication, and persistence. Wiener emphasized common principles in animals and machines, while Ashby's law of requisite variety links successful regulation to the range of perturbations a regulator can counter [40, 41]. The good-regulator theorem further connects effective control to models of the regulated system [42]. These ideas clarify why a prompted model producing an answer is not organismal: an autonomous system must repeatedly sense relevant conditions, act, evaluate consequences, and change policy while preserving viability.

Autopoiesis and the autonomy tradition impose a stronger requirement. A living organization receives resources and perturbations from outside, while its own processes contribute to producing the network and boundary that make those processes possible [30, 43, 24, 25, 29]. This circular dependence is *organizational closure*. It is compatible with thermodynamic openness: resources cross the boundary, while constraints and regulatory processes recursively maintain the organization. Whether computational systems can instantiate autopoiesis or only simulate it remains contested. Analyses of artificial cognition explore the concept's possible transfer, whereas molecular-autopoiesis accounts argue that algorithmic models omit constitutive material production [44, 45]. We therefore do not equate software feedback with molecular autopoiesis. Our engineering threshold asks the narrower, testable question of whether a physical computational organization regulates and repairs its own enabling constraints. A test suite that triggers another attempt is O1 feedback; a persistent system that monitors viability, protects state, reconfigures under damage, and restores failed components approaches O2–O3 integration.

Cybernetics also informs **I1, agency, goals, and self-models**. An agent is not defined solely by externally observed competence. Its actions must be organized around persistent norms or preferred states, and it must distinguish controllable consequences from environmental change. Active-inference and autonomy accounts formalize agency in terms of self-maintaining expectations, model-based control, and organism–environment coupling [46, 47, 48]. Artificial goals can initially be designed, just as biological norms are evolutionarily and developmentally constrained; the relevant question is whether the running system uses them to regulate its own continued organization rather than merely complete isolated tasks.

2.3 Information science: memory, heredity, and individuality

Information is indispensable to life, but raw storage is not enough. Biological information is interpreted within a system that uses it to build, regulate, and reproduce organization. Kolchinsky and Wolpert define semantic information as

correlations with an environment that are causally necessary for maintaining a system's existence [31]. Information-theoretic and Granger-causal measures separately ask how strongly a system's future is generated by its own history rather than imposed by environmental variables [49, 50]. Krakauer and colleagues use information flow to characterize individuality across scales, permitting boundaries that are graded and distributed rather than anatomically obvious [51]. These measures do not prove life, but they make autonomy and boundary claims quantitatively contestable in cloud and multi-robot systems, where the candidate individual cannot be identified with one process or shell.

This yields **L3, boundary, embodiment, and persistent identity** and **L4, memory and organizational continuity**. The boundary specifies which variables and processes belong to the continuing individual, which resources are environmental, and which transformations count as repair, replacement, migration, reproduction, or death. The information-bearing organization may include model weights, source code, prompts, reward models, policies, memory stores, executable skills, hardware specifications, access controls, and developmental records. Within an individual, this composite state supports continuity through interruption and component turnover. It becomes heredity only when a descendant uses it to reconstruct characteristic organization, a transition assessed under L5. Digital-evolution platforms make the distinction concrete: Avida organisms contain executable instruction genomes that self-replicate with mutation and compete for CPU time, producing measurable genotype–phenotype maps [52, 53].

For current agents, most memory is developmental rather than hereditary. Episodic stores, retrieval systems, summaries, and skill libraries can preserve experience across tasks, but identity may still collapse when a service restarts or a prompt changes. A strong silicon-based organism requires both continuity and controlled transformability: it must preserve organization through interruption while permitting learning, migration, and reproduction without making "the same individual" indistinguishable from any copy.

2.4 Artificial life: substrate independence and genuine evolution

Artificial life (ALife) studies possible living processes in software, robotics, synthetic systems, and hybrid media [21, 22, 23]. Its history shows that "silicon life" is not a new label but a changing experimental and cultural research program, from collective robotics to digital ecologies and virtual creatures [54, 55, 56, 57]. Its strongest contribution is not the claim that any simulation is alive. It provides experimental systems in which self-reproduction, ecological interaction, developmental encoding, and evolution can be isolated and measured. Tierra and Avida established populations of executable digital organisms; cellular-automaton lineages such as Evoloops demonstrate self-reproduction with mutational and ecological dynamics; contemporary work searches large spaces of cellular automata and artificial-life programs with foundation models [58, 52, 59, 60].

Some minimal computational patterns have been analysed as organizationally closed under definitions far less demanding than the present target [61]. This is not a contradiction: a Game-of-Life pattern may instantiate minimal persistence while lacking resource autonomy, reconstructive heredity, semantic cognition, and a governed physical life cycle. The O4 threshold used here is explicitly a stringent operational standard for an intelligent artificial individual, not a claim that all ALife researchers must reserve the word *organism* in the same way.

ALife also supplies the strictest warning against equating evolutionary algorithms with evolution. Many systems have externally fixed representations, mutation operators, fitness functions, and termination criteria. Open-ended evolution instead asks whether a system continually generates adaptive novelty, new entities or interactions, increased complexity, or changes in its own capacity to evolve [38, 39]. These requirements make L6 the least established criterion in contemporary AI agents. They also make ALife a bridge: it contributes lineage and ecology mechanisms that agent and robotics research often lacks, while foundation models contribute semantic cognition and tool use that classical digital organisms lacked.

2.5 Robotics and embodiment: a body is a causal interface

Robotics prevents an informational definition from becoming disembodied. A body is not necessarily humanoid or confined to one shell. It is a persistent causal interface through which an organization samples an environment, acts, incurs costs, and encounters irreversible consequences. Situated and embodied approaches to intelligence have long argued that cognition depends on sensorimotor structure and environmental coupling rather than detached symbol manipulation alone [62, 63]. Physical bodies add latency, friction, wear, occlusion, collision, energy depletion, material scarcity, and repair. Digital bodies, including browsers, operating systems, code repositories, accounts, and APIs, also impose boundaries and consequences, although they are easier to copy and reset.

Embodiment therefore belongs to L3 and links to L1–L2. Sensors and actuators close a fast behavioral loop; batteries, thermal constraints, component health, and maintenance close a slower viability loop. Multi-embodiment foundation policies complicate identity: one cognitive organization may control many robots, while one robot may instantiate

changing models. The organismal unit must consequently be determined by causal and informational continuity, not appearance. A robot that follows instructions is a body for an agent; an artificial organism additionally treats the preservation, repair, and appropriate replacement of that body as part of its own regulatory problem.

2.6 Cognitive science: self-models, planning, and intelligence

Life does not imply human-like intelligence, and intelligence does not imply life. Nevertheless, the target of this article is *intelligent* artificial life. We use **I2, general cognition and long-horizon planning** for the ability to build transferable world models, represent time and space, compose skills, reason over alternatives, use tools, learn from consequences, and pursue goals across extended horizons. I1 and I2 are separable: a system can solve difficult problems without persistent self-regulation, while a simple organism can regulate itself with little abstract reasoning.

Self-models are particularly important at their intersection. A self-model need not be phenomenally conscious. It can represent morphology, capabilities, uncertainty, resource state, commitments, and causal influence. Robots that learn simulations of their own kinematics and recover after damage provide a concrete operational example [10]. World models, persistent memory, and metacognitive monitors provide computational counterparts. These functions support planning, but they also support organismal identity by allowing a system to recognize changes to itself.

We reserve **S1, scalable superintelligence** for demonstrated broad superintelligence, while separately reviewing mechanisms that could support a trajectory toward it. Foundation-model agents already exceed bacteria, plants, and many animals in language-mediated communication, external memory, code generation, and access to human knowledge, but this organism-relative cognitive superiority is not S1. In the conventional human-relative sense, superintelligence requires broad and reliable cognitive performance beyond humans, not isolated benchmark superiority [64, 65, 66]. Tests of general capability and algorithmic prediction can make this threshold more disciplined, but current systems remain uneven and often regress across tasks [67, 68]. Section 5 argues not that life guarantees superintelligence, but that organizationally closed life built on scalable digital cognition has structural routes toward it.

2.7 Ecology and social science: populations, culture, and institutions

Organisms exist in ecologies, and many transitions in biological complexity depend on cooperation, conflict, division of labor, and new levels of individuality [69]. Work on major synthetic evolutionary transitions shows why artificial evolution and communicating robots can serve as experiments on such changes rather than merely metaphors for them [70]. Intelligent populations add social learning and cumulative culture: information is preserved and transformed through communication rather than only genetic inheritance. We therefore define **E1, sociality, ecology, and cumulative culture**: persistent interactions among artificial individuals that produce cooperation, competition, norms, specialization, institutions, or inherited cultural artifacts.

Multi-agent task allocation alone is weak evidence. A developer assigning fixed roles to several prompts creates an engineered workflow, not an emergent society. Stronger evidence requires decentralized interaction, population-level state, persistence across episodes or generations, and behavioral structures not individually scripted. Artificial populations may nevertheless be central to silicon-based life because software organisms can exchange code, skills, memories, and policies at high bandwidth. The unit of closure could be an individual agent, a fleet, or a symbiotic human–AI infrastructure. Social and ecological analysis is needed to determine when a collective becomes an organism-like higher-level individual rather than an aggregation.

2.8 Philosophy: existence, consciousness, value, and responsibility

Philosophy clarifies four distinctions that engineering cannot settle by performance tests alone. First, a process view of identity allows an organism to persist while its material and informational components change. Second, functional life is distinct from phenomenal consciousness. Current consciousness-science analyses find no basis for confidently attributing consciousness to existing AI systems, while identifying computational properties that future systems could instantiate [71]. Third, moral status does not automatically follow from intelligence, autonomy, danger, or biological resemblance. Artificial-life ethics argues for explicit attention to possible interests and welfare before creating large populations of potentially morally considerable systems [72]. Fourth, causal responsibility and legal accountability cannot be dissolved by distributed identity.

These issues are not additional boxes that a system must tick to be alive. They constrain interpretation and governance. A non-conscious artificial organism may still reproduce or cause harm; a conscious system may merit protection even if it falls short of L5–L6; and a distributed agent may require a legally assigned controller even when its technical boundary is ambiguous. The framework therefore keeps ontology, phenomenology, moral status, and governance analytically separate while showing where they interact.

2.9 Integrated working definition and operational criteria

We define the target as follows:

> **Silicon-based artificial life** comprises persistent digital or cyber-physical artificial individuals, instantiated primarily through engineered computation and software or robotic bodies, whose resource, regulatory, informational, cognitive, and bodily processes recursively maintain their organization and identity. Crossing this individual threshold does not require every individual to reproduce. When a lineage claim is made, behavior-shaping organization must be reconstructed in functional descendants with controlled variation and differential persistence. At ecological level, individuals or higher-level collectives participate in sustained social and resource interactions. **Silicon-based superintelligent life** is the subset whose closed cognitive organization additionally demonstrates broad, reliable performance beyond humans at individual or ecological scale.

The definition is functional but not permissive, graded but not arbitrary, and integrative rather than checklist-based. It contains two nested thresholds. O4 closure across the relevant L1–L4 dimensions supports artificial individuality; L5–L6 assess whether that organization continues as an evolving lineage. S1 is not a criterion for life itself, but it is required for the adjective *superintelligent*. A closed organism with only a plausible path to beyond-human cognition has not yet crossed S1. Table 2 supplies the operational interface used by Section 3: I1–I2 characterize intelligent agency and E1 captures population and cultural organization. A present system can satisfy some dimensions without being an organism. The transition requires the relevant viability dimensions to become mutually sustaining, followed at lineage level by reconstructive heredity and selection, as formalized in Section 4.

2.10 A graded taxonomy

We use four descriptive classes without assuming an inevitable progression. *Tools* execute bounded transformations without persistent goals or identity. *Agents* close perception–action loops and maintain goals over bounded tasks. *Proto-organismic systems* couple several life-like functions, such as persistent memory, self-monitoring, embodied feedback, resource awareness, or regeneration, but still rely on an external organization to maintain their joint continuity. *Artificial organisms* achieve O4 organizational closure: their component processes maintain the continuing individual under perturbation and turnover. O5 systems additionally sustain reconstructive descent and open-ended evolutionary or cultural novelty at population level.

Most current foundation-model agents occupy the tool–agent boundary; selected persistent agents, digital-life systems, self-maintaining software, and adaptive robots reach O2 or O3 on particular dimensions. This graded vocabulary supports the article's thesis without declaring all autonomous software alive. It also makes progress measurable: the important question is not whether a model produces life-like language, but which dependencies have moved from human infrastructure into the regulated organization of the system itself.

ID	Disciplinary root	Operational requirement	Silicon-based realization	Decisive test
L1	Biology, thermodynamics	Regulated throughput maintains organization	Energy, compute, bandwidth, memory, cooling, parts, permissions	Preserves viability under scarcity without stepwise human provisioning
L2	Cybernetics, autonomy	Feedback controls viability-relevant variables	Monitoring, critics, safety control, fault detection, recovery, repair	Returns to a viable region after unanticipated disturbance
L3	Biology, robotics, information	A causal boundary and identity persist through change	Software or robot body, credentials, self-model, lineage and state identifiers	Distinguishes migration, repair, replacement, copying, and death
L4	Genetics, information science	Behavior-shaping organization persists through interruption and component turnover	Weights, code, policies, prompts, memory, skills, configurations	Restores characteristic organization after interruption, migration, or controlled replacement
L5	Biology, ALife, robotics	Inherited organization regenerates the individual or constructs functional descendants	Safe deployment, self-assembly, fabrication, copying plus activation	Descendant completes a comparable operational life cycle
L6	Evolution, ALife	Heritable variation undergoes differential selection and sustained novelty	Mutated code or policies, evolutionary search, artificial ecologies	Adaptive novelty and complexity continue without a fixed terminal target
I1	Cybernetics, cognitive science	Persistent goals and self-model regulate action	Goal memory, capability and body models, metacognition, active inference	Maintains commitments and updates self-estimates across contexts
I2	Cognitive science, AI	General, grounded, long-horizon cognition	World models, planning, tools, lifelong learning, scientific reasoning	Transfers and recovers across long tasks and changing environments
E1	Ecology, sociology	Populations generate durable interaction structures and culture	Multi-agent populations, fleets, shared artifacts, norms, institutions	Conventions or institutions persist and affect later agents
S1	AI, information science	Closed cognition demonstrates broad, reliable superiority beyond humans	Fast copying, parallelism, shared memory, many bodies, automated R&D	Broad, reliable superiority under human-relative and human-agnostic tests

Table 2: Operational criteria for reviewing current systems. The criteria are not independent checklist items. Individual organismality requires mutually maintaining L1–L4 and a defensible life cycle; a continuing artificial-life lineage additionally requires L5–L6. I1–I2 and E1 characterize intelligent and ecological organization; S1 is the separate broad-superintelligence threshold.

3 Instantiating Life-Like Functions in AI Agents, Artificial Life, and Robots

This section reviews current systems against the ten criteria in Table 2. The unit of evidence is a demonstrated function, not a product name. We distinguish software agents, digital organisms, robot policies, physical self-maintenance mechanisms, and populations because success in one substrate does not automatically transfer to another. Existing surveys separately organize LLM agents, scientific agents, robot foundation models, vision–language–action policies, and embodied language systems [73, 74, 75, 76, 77, 78]; our purpose is to map their experimental results onto substrate-neutral life criteria. For each criterion, we ask what has been demonstrated, how strongly components are coupled, and what prevents the result from constituting organism-level closure.

Figure 3 replaces a conventional "LLM brain plus robot body" diagram with the architecture implied by the criteria. A candidate artificial organism requires at least two coupled timescales. A fast cognition–action loop turns perception, memory, planning, and tools into environmental action. A slower viability–lineage loop monitors resources and body state, preserves identity, consolidates inherited organization, repairs or reconfigures components, and controls

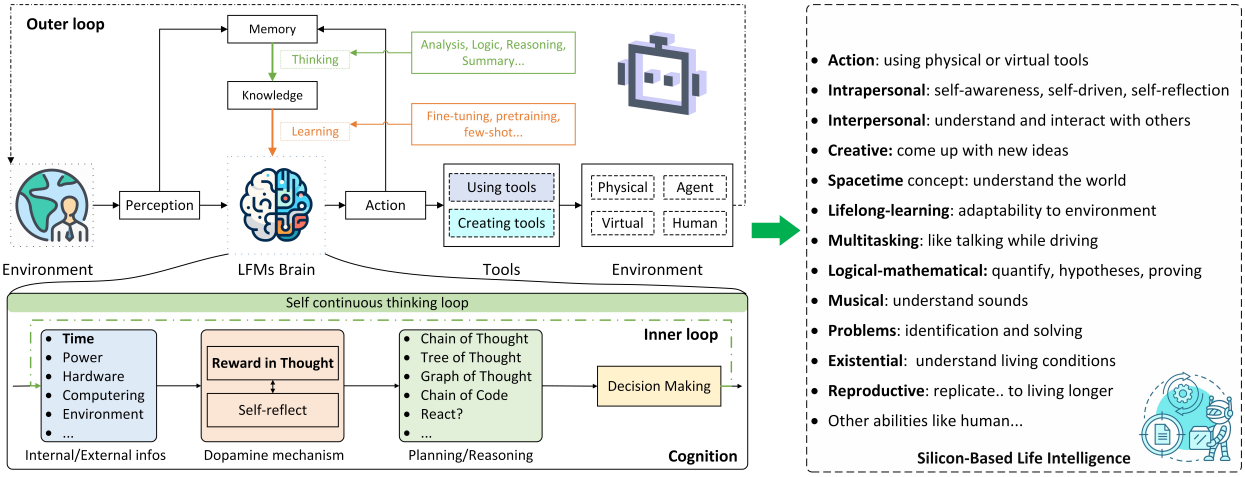


Figure 3: Conceptual framework linking an AI-agent cognitive loop to memory, knowledge, perception, action, tool use, and virtual or physical environments, together with a nested continuous-thinking loop. The right panel summarizes proposed intelligent and life-like capabilities of silicon-based life intelligence. The diagram presents a systems-level perspective; it does not imply that any current agent or robot already integrates all of these properties.

reproduction. A population layer supplies selection, cooperation, competition, and culture. Current research populates nearly every box, but the connecting arrows remain incomplete.

3.1 L1: Resource throughput and self-maintenance

Current AI consumes energy, compute, memory, bandwidth, credentials, and data, but consumption alone is O0 analogy. Most agents neither represent these resources as viability variables nor secure their continued availability. Their processes are started, funded, cooled, authenticated, and terminated by an external operator. The strongest software evidence comes from autonomous-replication evaluations because they decompose continued operation into obtaining resources, accessing compute, deploying, and persisting. RepliBench contains 20 task families and 86 individual tasks across resource acquisition, model-weight access, replication onto compute, and persistence. The best evaluated model exceeded 50% pass@10 on 15 of 20 families and on 9 of 20 hardest variants, but no tested system was a credible end-to-end self-replicator; identity checks and robust long-term deployments remained major barriers [13]. This is an O1 demonstration of resource-acquisition subskills, not autonomous metabolism.

Robotics supplies a more literal material counterpart. Wyder and colleagues' modular "robot metabolism" platform allowed machines to attach reusable truss links from their surroundings, grow into more capable configurations, and shed or replace modules [12]. The work demonstrates O1 material incorporation and O2 coupling between morphology and improved function. Long-term-autonomy research has treated charging, scheduling, navigation, perception, and recovery as an integration problem for years, while documenting the difficulty of operation over weeks or months outside controlled demonstrations [79, 80]. Importantly, energy renewal is not only a planning concept: an autonomous rotorcraft repeatedly landed and recharged during experiments lasting up to 11 hours, and AutoCharge demonstrated repeatable docking and energy replenishment during a 10-hour quadrotor experiment [81, 82]. Earlier soft autonomous robots integrated on-board fuel reservoirs, catalytic reaction chambers, fluidic logic, and actuation in one body [83]. These systems establish bounded artificial resource cycles and bodily growth, but none couples autonomous resource discovery, economic or material acquisition, energy regulation, cognitive planning, and repair in one persistent organization.

Assessment. L1 has strong O1 and narrow O2 demonstrations. The missing threshold is a system that treats compute, energy, permissions, and body material as a joint viability budget, adapts under scarcity, and restores those flows without stepwise human provisioning.

3.2 L2: Homeostasis, feedback, and repair

Agentic AI routinely uses feedback, but most feedback regulates task success rather than existence. ReAct alternates reasoning, action, and observation; Reflexion stores verbal feedback after failure; SWE-agent edits code, executes tests, and revises patches through an agent–computer interface [3, 4, 84]. These are O1 cybernetic loops. They become

life-relevant when error detection protects persistent identity, memory, resource state, or a body rather than only a benchmark score.

Recent robotics closes more of this gap. A self-supervised robot framework learned a simulation of its own morphology and kinematics from brief video, used that model for planning, detected abnormalities, and recovered kinematic function after damage [10]. This is O2 coupling among self-model, anomaly detection, and behavioral compensation. Materials research provides complementary bodily repair. Self-healing electronic skin recovered mechanical and sensing function after damage [11]; severed electroluminescent robotic fibres recovered 98.6% of pristine luminance and remained functional for more than ten months [85]. Homeostatic soft-robotics proposals explicitly treat self-preservation as a meta-goal linking bodily state to intelligent behavior [48].

The gap is integration and timescale. A material that heals under prescribed conditions does not decide when healing is necessary. A robot that compensates for a damaged joint does not fabricate a replacement. An agent that retries code does not preserve itself across power loss. Organism-level L2 requires detection, diagnosis, prioritization, repair or reconfiguration, and post-repair validation to operate as one viability loop.

Assessment. O2 homeostatic fragments now exist in both software and hardware. Among the reviewed foundation-model and robotic systems, none reaches O3 maintenance of cognitive state, resource envelope, and physical body under unanticipated long-duration disturbance.

3.3 L3: Boundary, embodiment, and persistent identity

Digital agents already possess causal interfaces that function as limited bodies. WebArena gives agents realistic websites and executable goals; OSWorld provides 369 tasks across desktop applications, operating-system operations, file I/O, and cross-application workflows [86, 87]. In the original OSWorld study, humans completed 72.36% of tasks while the strongest evaluated baseline completed 12.24%, revealing both genuine causal access and severe grounding failures [87]. SWE-agent, AppAgent, and Mobile-Agent similarly act through terminals, repositories, and smartphones [84, 88, 89]. These environments establish O1 digital embodiment: actions alter an external state and consequences are executable rather than textual.

Physical embodiment is advancing from language-guided planning to learned vision–language–action policies. SayCan grounded language plans in affordance values; RT-1 and RT-2 learned large-scale language-conditioned control; PaLM-E integrated continuous sensor inputs; Open X-Embodiment trained across heterogeneous robots; and Octo, OpenVLA, RDT-1B, π_0, Gemini Robotics, and GR00T N1 extend open or generalist robot policies [90, 5, 7, 6, 8, 91, 9, 92, 93, 94, 95]. RT-1-X improved success by roughly 50% on average over single-dataset baselines across five robot settings, supporting the transfer of experience across bodies [8]. A 2026 VLA study compared eight backbones, four policy architectures, and more than 600 experiments, showing that generalist performance depends systematically on architecture and cross-embodiment data rather than on adding language labels alone [96].

Systems are also beginning to integrate high-level language agents with physical feedback. ELLMER used an embodied LLM, retrieval, code generation, force feedback, and vision feedback for long-horizon tasks in unpredictable settings [97]; an open robot-operating-system framework connected LLM agents to embodied tools and controllers [98]. These are O2 architectures. Yet identity is normally assigned externally by a process ID, account, robot serial number, or experimenter. Memory, model, body, and credentials can be separated without the system representing whether it has migrated, been repaired, or been replaced.

Assessment. L3 is one of the strongest dimensions at O1–O2. O3 requires a tested identity protocol that persists across restarts, body changes, memory updates, and copies while preserving causal accountability.

3.4 L4: Memory and organizational continuity

Agent memory has moved beyond a context window. Architectures distinguish working, episodic, semantic, procedural, and external memory; they retrieve histories, consolidate summaries, store executable skills, and update long-term state [99, 100]. Voyager combined an automatic curriculum, environment feedback, self-verification, and a growing code-based skill library. It acquired 3.3 times more unique items, travelled 2.3 times farther, and reached selected technology-tree milestones up to 15.3 times faster than baselines in Minecraft [101]. HELPER similarly reused memory-augmented language programs in embodied household tasks [102].

New benchmarks make the limitation explicit. LifelongAgentBench evaluates interdependent tasks in database, operating-system, and knowledge-graph environments and reports that naive experience replay is often ineffective because irrelevant memories and context constraints interfere with reuse [103]. Experience-driven lifelong-learning systems add persistent memory, skill abstraction, and knowledge internalization across evolving scenarios [104]. These

are O1–O2 demonstrations of developmental continuity, but memory contamination, retrieval errors, catastrophic forgetting, and version drift remain unresolved.

Digital-life research makes the boundary between continuity and heredity unusually explicit. In Avida, an executable instruction sequence contributes to the organization of one running digital individual; when that sequence is copied, activated, mutated, and subjected to differential replication, the same medium also functions as lineage-level heredity [52, 53]. Current foundation-model agents possess several candidate organization-bearing media: weights, code, prompts, memories, policies, tools, and configurations. They rarely package these media into a provenance-controlled state that restores one characteristic organization, still less into a behaviorally validated lineage. A backup is not a genotype unless its activation reconstructs the organization and its variations create heritable phenotypic differences.

Assessment. L4 has robust O1 memory and O2 continuity mechanisms in foundation-model agents; classical digital organisms additionally demonstrate O3 organization continuity within bounded, interpreter-mediated life cycles. The open individual-level problem is a provenance-controlled state package that restores characteristic organization after interruption or migration. Transmission of that package to a functional descendant is assessed under L5, not counted again as L4.

3.5 L5: Reproduction and regeneration

Physical self-reproduction is not purely hypothetical. Modular machines have autonomously assembled copies of their own multi-cube organization, and reconfigurable strings have reproduced structures from randomly positioned parts [105, 106]. Kinematic self-replication in reconfigurable biological robots showed that designed multicellular configurations could collect loose cells into motile offspring, although this is a biohybrid rather than silicon system [107]. Cellular automata and digital organisms provide stronger multi-generation examples because their executable organization directly produces descendants [59, 52].

Foundation-model agents contribute deployment and orchestration skills. RepliBench shows that frontier agents can provision some cloud resources, write self-propagating programs, deploy instances in constrained settings, and extract weights under weak security conditions, while failing to robustly establish persistent recursive deployments [13]. Agent frameworks can spawn subprocesses or role-specialized agents [108, 109], but a parent program calling another copy is not reproduction unless the descendant inherits an organization, starts independently, and can continue a lineage. Robot metabolism supplies growth and module reuse but has not demonstrated a full autonomous descendant cycle [12].

Regeneration should be treated alongside reproduction. Restoring memory after interruption, re-instantiating a failed service, replacing a body module, or migrating a policy to a new body can renew the same organization without creating a new individual. Distinguishing regeneration from descent requires the L3 identity protocol, so copying, identity, and reproduction must be evaluated together.

Assessment. L5 ranges from O1 replication subskills in current agents to O3 reproduction in deliberately designed digital organisms. In the reviewed corpus, no foundation-model-based robot or hybrid agent has produced, provisioned, validated, and released a functional descendant that can repeat the cycle under bounded autonomy.

3.6 L6: Variation, selection, and open-ended evolution

Evolution requires more than self-improvement. RoboCat trained across tasks and bodies, generated additional experience, and fed successful trajectories into later training. With 500 demonstrations on unseen tasks, broader training increased average success from 36% to 74%, while adaptation to a new arm achieved 86% gear-picking success after 1,000 demonstrations [110]. This is O2 iterative improvement and cross-embodiment transfer, but the data cycle and objectives remained designed and curated.

Evolutionary coding agents are closer to explicit variation and selection. AlphaEvolve generates code variants, evaluates them with automated functions, preserves successful programs, and iterates across a population, producing new algorithms and infrastructure optimizations [111]. ASAL uses foundation models to search for life-like dynamics in cellular automata and other artificial worlds [60]. Evoloops and Avida provide self-replication, mutation, differential persistence, and ecological interaction without an LLM in the loop [59, 52]. Together these literatures demonstrate every elementary Darwinian operator.

The missing property is open-endedness. AlphaEvolve depends on externally supplied evaluators; robot self-improvement optimizes selected tasks; and most digital worlds have fixed instruction sets and resource rules. Open-ended evolution requires ongoing adaptive novelty, growth in meaningful complexity, new ecological niches or levels of organization, and possibly evolution of evolvability [38, 39]. Foundation models may expand the generative space, but they do not by themselves remove external closure.

Assessment. L6 has O3 Darwinian dynamics in classical ALife and O2 high-level variation–evaluation loops in modern agents. O5 open-ended evolution has not been established in an AI-agent population.

3.7 I1: Agency, persistent goals, and self-models

Planning agents maintain task goals over trajectories, use tools, and revise plans, but their goals are usually externally supplied and context-bound. Knowledge-augmented planners and reflection systems improve coherence [112, 4]; active-inference and homeostatic robotics provide architectures in which preferred internal states organize action [48, 46]. The self-simulation robot discussed above is unusually relevant because the model concerns the agent's own morphology and supports damage detection and recovery rather than only task prediction [10].

Negative results are equally important. Goal-drift evaluations show that language-model agents can abandon or reinterpret assigned goals under long contexts, competing instructions, or environmental pressure [113]. Agents often misestimate their capabilities, fail to recognize irreversible actions, or lose commitments when memory is summarized. These failures distinguish fluent first-person language from a stable operational self.

Artificial agency need not involve spontaneously created human-like desires. Biological goals are also constrained by inherited organization. The decisive test is whether goals and self-models regulate persistence across episodes: can the system represent its own resources, damage, uncertainty, permissions, lineage, and obligations; preserve or legitimately revise commitments; and use those representations causally?

Assessment. Current systems show O1 task agency and isolated O2 self-model-based regulation. Persistent organism-level agency remains below O3.

3.8 I2: General cognition and long-horizon planning

Agents now act over increasingly long digital and physical horizons, but benchmark results reveal a steep reliability gradient. The METR time-horizon study measured the duration of software tasks that frontier agents complete with 50% reliability. In its evaluated setting, o3-based agents reached roughly 110 minutes, and the estimated horizon had doubled about every seven months since 2019; the authors explicitly caution that the tasks are concentrated in well-specified software work and may not generalize to messier domains [114]. OSWorld's 12.24% baseline against 72.36% human performance illustrates the remaining gap in open-ended computer use [87]. PaperBench similarly found that the strongest tested agent initially achieved an average score of 21.0% on reproducing AI research, showing partial research execution rather than independent scientific competence [115].

Embodied systems provide complementary evidence. Voyager acquires and composes skills over an open-ended game trajectory [101]; ELLMER integrates retrieval, generated code, force, and vision feedback for multi-step manipulation [97]; generalist VLA studies show positive transfer but persistent brittleness across embodiments [8, 96]. Long-horizon competence is therefore increasing along both digital and physical axes, while recovery, causal grounding, and evaluation under unstructured conditions remain limiting.

Scientific agents extend I2 beyond task execution. The AI Scientist links idea generation, literature search, code editing, experiments, analysis, manuscript production, and review [15]. Its strongest submitted manuscript received reviewer scores of 6, 7, and 6, above the average acceptance threshold of a 70%-acceptance workshop; only one of three submissions met that bar, and none met the authors' standard for a main ICLR paper. This is O2–O3 integration with a clear ceiling, not autonomous scientific superintelligence.

Assessment. I2 has substantial O2 and bounded O3 integration. Reliability falls with duration, environmental openness, physical uncertainty, and the need to define rather than merely optimize a problem.

3.9 E1: Sociality, ecology, and cumulative culture

Most multi-agent systems demonstrate engineered coordination. AutoGen structures conversations, MetaGPT assigns software roles, CAMEL uses role-playing communication, and Mixture-of-Agents aggregates outputs [108, 109, 116, 117]. These systems show O1 communication and division of labor, but their topology, roles, stopping conditions, and shared goals are externally specified. MultiAgentBench moves toward comparative measurement by evaluating collaboration and competition across interactive scenarios [118].

Population experiments provide stronger evidence for emergent social structure. Decentralized LLM populations playing a naming game spontaneously converged on shared conventions; the study evaluated multiple models and repeated runs, including 40 runs for each of two Llama variants, and showed that minority agents could redirect collective conventions [14]. Cultural-evolution experiments found model-dependent differences in indirect reciprocity

and cooperation across generations of LLM agents [119]. Work on cultural transmission in embodied reinforcement-learning agents demonstrates high-fidelity few-shot imitation without human demonstration data and frames social learning as a mechanism for cumulative capability [120].

These studies remain idealized. Naming games, donor games, and simulated towns do not establish durable economies, institutions, legal responsibility, or open-ended culture. Homogeneous model populations can amplify training-data priors, and an LLM evaluator can mistake familiar narrative patterns for emergent social complexity. Strong evidence for criterion E1 therefore requires persistent artifacts, cross-generation effects, heterogeneous populations, independent behavioral metrics, and ecological costs.

Assessment. Social functions under criterion E1 reach O2 in controlled populations. No reviewed system reaches O4 while also showing durable institutions and cumulative culture coupled to organismal reproduction and resource competition.

3.10 S1: Scalable superintelligence

Current systems are not broadly superintelligent relative to humans, but several cognitive modules relevant to a superintelligent artificial organism are advancing. Automated science is an informative integration domain. Robin coordinated literature-search and data-analysis agents in an iterative biological-discovery workflow, analysed 551 papers in 30 minutes versus an estimated 294 human hours, and was estimated to reduce total workflow time roughly 200-fold while retaining wet-lab experiments and candidate selection in the human loop [121]. Co-Scientist continuously generated, critiqued, ranked, and evolved hypotheses. Across 203 research goals, its internal hypothesis Elo ratings improved with test-time compute; on a 15-goal biomedical subset it exceeded expert and frontier-model baselines under the paper's automated protocol, while a separate blinded expert comparison covered only 11 goals and remained subjective [122]. Kosmos extended coherent data-analysis and literature-search workflows to approximately 200 rollouts over 12 hours; its authors reported 79.4% statement accuracy, but an independent three-problem radiation-biology audit found one supported result, one uncertain result, and one false hypothesis [123, 124]. AlphaEvolve adds explicit evolutionary code search and automated evaluation [111]. Together these systems demonstrate longer and more integrated discovery loops, while their evaluator dependence, human framing, and uneven external validation do not support autonomous scientific superintelligence.

These systems support three substrate advantages: cognitive work can be parallelized among copies; successful artifacts can be preserved exactly and distributed immediately; and evaluation loops can operate faster than biological learning or institutional science. They also expose limitations. Robin was semi-autonomous and depended on human experiments; The AI Scientist required filtering and produced frequent methodological and citation failures; Co-Scientist's automated ratings partly depend on model-based evaluation; AlphaEvolve requires executable evaluators and bounded problem representations. General AI scales and human-agnostic tests also show uneven profiles, non-monotonic model improvements, and a large distance from proposed universal-intelligence targets [67, 68].

Assessment. S1-relevant mechanisms have strong O2 demonstrations of superhuman speed or quality on selected cognitive processes and O3 integration in bounded automated research loops. These are precursor results, not attainment of S1. Broad human-surpassing intelligence, autonomous recursive improvement, and closed organism-level cognition are not demonstrated.

3.11 Evidence synthesis: complete parts, incomplete organization

Table 3 summarizes the review. Evidence is uneven across the criteria. Embodiment, memory, feedback, and advanced cognition have multiple O2 demonstrations. Resource autonomy, identity, and physical repair are narrower. Reproduction exists in digital organisms and modular machines but is not coupled to foundation-model agency. Open-ended evolution and organizational closure remain absent. This pattern supports a component-availability claim while rejecting a present-life claim.

The review supports a limited conclusion. Many necessary mechanisms have existed in ALife and robotics for decades. The recent change is that semantic foundation-model cognition, general tool use, persistent memory, cross-embodiment policies, physical self-models, reusable bodies, replication-relevant skills, population dynamics, and automated discovery can increasingly be connected through overlapping software and robot interfaces. Their joint compatibility remains an architectural hypothesis. Organizational-closure experiments must determine whether these distributed capabilities can maintain the same continuing individual.

ID	Strong representative evidence	Quantitative or operational result	O stage	Missing threshold
L1	RepliBench; robot metabolism; autonomous soft robot	15/20 task families above 50% pass@10 for the best RepliBench model; reusable modules incorporated into more capable morphologies	O1–O2	Joint autonomous budgeting and renewal of energy, compute, permissions, and body material
L2	Reflexion/SWE-agent; learned robot self-simulation; self-healing fibres	Task-error recovery; damage detection and kinematic compensation; 98.6% luminance recovery after severing	O1–O2	Integrated diagnosis, repair, memory continuity, and viability control over long horizons
L3	OSWorld; RT-X/VLA; ELLMER; LLM–ROS	369 OS tasks; 12.24% best baseline versus 72.36% human; cross-embodiment transfer and physical feedback	O1–O2	Operational identity across restart, migration, body replacement, and copies
L4	Voyager; LifelongAgentBench; Avida	3.3× more items and up to 15.3× faster milestones; executable organization maintained through bounded Avida life cycles	O1–O3	Provenance-controlled state package that restores characteristic organization after interruption or migration
L5	Modular self-reproducing machines; RepliBench; Evoloops	Physical copying from modules; deployment and persistence subskills; multi-generation digital reproduction	O1–O3	A foundation-model agent or robot completing and repeating an autonomous descendant cycle
L6	Avida/Evoloops; RoboCat; AlphaEvolve; ASAL	Darwinian digital populations; 36% to 74% unseen-task improvement in RoboCat; evaluated code variation	O2–O3	Sustained adaptive novelty and evolvability without a fixed external endpoint
I1	Goal-maintaining agents; homeostatic robotics; self-model robot	Planning and reflection; self-model supports damage recovery; goal drift remains measurable	O1–O2	Stable, causally effective self-model and goals tied to continued organization
I2	METR; OSWorld; Voyager; AI Scientist; ELLMER	Roughly 110-minute 50% software-task horizon; one of three AI papers crossed a workshop threshold	O2–O3	Reliable generalization, causal grounding, and recovery in open long-horizon environments
E1	MultiAgentBench; convention formation; cultural cooperation	Repeated decentralized populations converge on conventions; model-dependent cooperation across generations	O1–O2	Persistent heterogeneous ecologies, institutions, lineage effects, and independent measures of culture
S1	Robin; Co-Scientist; AlphaEvolve; general AI scales	551 papers in 30 minutes and estimated 200× workflow acceleration; iterative hypothesis and algorithm improvement	O2–O3*	Broad human-surpassing reliability plus autonomous, bounded improvement within a closed organism

Table 3: Evidence map from current systems to the operational criteria. O stages describe demonstrated organizational integration, not study quality or rhetorical similarity. In the S1 row, O2–O3* refers only to integration of supporting mechanisms; S1 itself is unmet. No row reaches O4 because no reviewed system mutually maintains the individual-viability functions L1–L4 as one persistent artificial individual; L5–L6 are separately assessed at lineage level.

4 From Distributed Functions to Organizational Closure

The evidence in Section 3 addresses only the component problem. A collection can contain a battery monitor, a memory store, a planning model, a robot, a deployment script, and an evolutionary optimizer without forming an organism. The difference is relational. In an organism, component processes participate in maintaining the organization that makes their own continued operation possible. Coexistence within one implementation or reporting to an external operator is insufficient. Metabolism renews components, a boundary channels metabolism, regulation preserves the boundary, information reconstructs regulation, and reproduction transmits the organization. This circular dependence is the sense of *organizational closure* developed in autonomy theory [24, 25]. Recent process-enablement graphs sharpen this idea into a formal analysis of which processes enable and recursively depend on others [125]; repair-closure proposals independently make recovery after targeted perturbation the decisive observable, although that recent formalism remains a preprint [126].

Closure is often misunderstood as isolation. Living systems are materially and energetically open. The closure concerns constraints, dependencies, and functions: a set of processes is closed when each necessary enabling condition is produced, repaired, or regulated by another process within the continuing organization. This interpretation is suitable for distributed artificial systems. Electricity, replacement parts, cloud hardware, and data can remain environmental resources, just as food and oxygen are environmental for animals. What must move inside the organization is the capacity to sense need, select and acquire permitted resources, allocate them, repair failures, protect identity, and adapt behavior so that the same organization persists.

4.1 Four closures that define the transition

We decompose the transition into four coupled forms of closure.

Operational closure exists when perception, action, memory, resource control, learning, and repair maintain one another through recurrent loops. A monitor that alerts a human is not closed; a system that diagnoses a fault, selects a bounded intervention, validates recovery, and preserves state begins to close the loop.

	L1	L2	L3	L4	L5	L6	I1	I2	E1	S1
Digital organisms	O1	O2	O2	O3	O3	O3	O1	O1	O2	
Computer agents	O1	O1	O1	O2	O1		O1	O2	O1	O1*
VLA and embodied robots	O1	O1	O2	O1			O1	O2		
Self-healing and modular robots	O1	O2	O2		O1					
Self-reproducing machines	O1	O1	O2	O2	O2					
Evolutionary coding agents		O1		O2	O1	O2	O1	O2		O2*
Multi-agent populations		O1		O1			O1	O1	O2	O1*
AI-scientist systems	O1	O2	O1	O2	O1	O2	O2	O3	O2	O3*

O1 component **O2** coupled components **O3** persistent bounded loop
*S1-relevant precursor; S1 itself is not attained

Figure 4: Qualitative evidence landscape synthesized from the literature reviewed in Section 3. Cells report the strongest representative organizational-integration stage within each system family; blank cells indicate no direct evidence rather than impossibility. Asterisks in the S1 column mark integration of S1-relevant precursor mechanisms, not attainment of broad superintelligence. The horizontal incompleteness is the empirical basis for the component-availability claim, while the absence of O4 cells records the organizational-closure gap. These are qualitative synthesis judgments, not study-quality scores.

Identity closure exists when the system maintains an operational self–environment distinction and can classify transformations of itself. It must distinguish ordinary state change, learning, repair, migration to new compute, embodiment in another robot, reversible suspension, destructive replacement, branching copies, and death. Identity closure requires causal continuity, provenance, and rules governing which memory and policy changes remain attributable to the same individual.

Reproductive closure exists when a descendant is generated from resources and inherited organization, activated, and validated without an external engineer reconstructing the missing functions. Reproduction may be digital, physical, or hybrid, but a descendant must be capable of completing a comparable cycle. Controlled variation connects reproduction to selection rather than exact backup.

Ecological closure exists when a population sustains the interactions required for its members or higher-level collective to persist. Resource exchange, specialization, communication, competition, norms, and cultural artifacts may become part of the organization. This level matters because silicon-based life may emerge first as a symbiotic cloud–robot–human ecology rather than a self-sufficient humanoid.

No single closure is sufficient. Operational closure without identity produces replaceable automation; identity without self-maintenance produces a persistent database; reproduction without regulation produces self-copying malware; ecology without individual continuity produces a workflow. O4 artificial organismality requires operational and identity closure at minimum, with a defensible life cycle. O5 additionally requires reconstructive descent and population-level evolutionary or cultural novelty; ecological closure is one strong route to that population persistence, not a substitute for descent.

4.2 The current evidence landscape

Figure 4 makes the distributed-evidence argument explicit. Classical digital organisms reach high integration in heredity, reproduction, and evolution but have minimal semantic cognition and simplified resource worlds. Agent systems supply planning, memory, and tools but weak self-maintenance. VLA robots supply causal bodies, while self-healing and modular robotics supply repair and material adaptation. AI-scientist and multi-agent systems supply collective cognition and iterative improvement. No family has evidence at O3 or higher across the full row of criteria, and no family meets S1.

The landscape also corrects a common inferential error. Distributed evidence can support feasibility because it shows that no individual function is purely speculative and that several interfaces already exist. It cannot establish that arbitrary components will compose. Integration can introduce conflicts: memory consolidation may alter identity; resource acquisition may violate safety constraints; repair may change a body model; replication may copy vulnerabilities; and

multi-agent coordination may obscure responsibility. The closure hypothesis must therefore be tested at the interfaces, not inferred from the number of boxes filled.

4.3 Three converging pathways

The *digital-first pathway* combines computer-use agents, persistent memory, software repair, cloud deployment, digital organisms, and artificial ecologies. It is experimentally tractable because state can be instrumented, copies can be sandboxed, and generations can run quickly. RepliBench, lifelong-agent environments, Avida, Evoloops, ASAL, AlphaEvolve, and AI-scientist systems supply different parts of this pathway [13, 103, 52, 59, 60, 111, 15]. Its weakness is dependence on human-operated physical infrastructure. A digital organism can regulate accounts and processes while remaining unable to repair a failed server or restore electricity.

The *embodied pathway* combines generalist robot cognition with charging, self-models, modular bodies, self-healing materials, fabrication, and physical reproduction. It directly confronts energy, wear, latency, irreversible action, and material scarcity [10, 12, 105]. Its bottleneck is integration cost: physical experimentation is slow, repair mechanisms are task-specific, and generalist policies remain brittle outside training distributions.

The *hybrid ecological pathway* distributes functions across agents, robots, cloud services, human institutions, automated laboratories, and supply chains. This is a plausible early site of strong closure because biological autonomy is also relational rather than resource-independent. A robot may request a part from an automated factory; a software agent may migrate around failed compute; a fleet may share memory; and humans may remain symbiotic participants. The scientific risk is moving the boundary outward until ordinary infrastructure is mislabeled an organism. Identity and dependency measurements must show that the candidate organization has more causal and informational closure than the surrounding service network [51].

4.4 A testable organizational-closure protocol

An O4 claim should require a prospective intervention study, not post hoc interpretation. A candidate system would be placed in a monitored digital, physical, or hybrid environment with bounded resources and explicit safety constraints. It would have to:

C1. maintain a persistent identity and auditable state through restarts, migration, and controlled body changes;

C2. monitor energy, compute, memory, network, and body-health variables and keep them within predefined viability ranges;

C3. recover from unannounced software, sensor, actuator, memory, and resource perturbations without stepwise human repair;

C4. preserve goals and safety constraints while updating its world and self models;

C5. regenerate failed components or produce a sandboxed functional descendant from inherited organization;

C6. demonstrate that disrupting one subsystem elicits compensatory responses through the others, establishing mutual dependence rather than parallel automation;

C7. remain governable: every resource acquisition, modification, copy, and external action must have provenance, limits, and a tested shutdown path.

The duration and environment should scale with the system's natural operational cycle; no universal number of days defines life. What matters is exposure to multiple resource and perturbation cycles, including failures not represented in training. Closure should be assessed as a dependency graph and compared with non-organismic controllers matched for task performance and resource budget. Process-enablement cycles, conditional mutual information, non-trivial information closure, and Granger-causal autonomy provide complementary structural and dynamical readouts [125, 49, 50]. None alone is a life detector. A supposedly internal function that can be removed and continuously replaced by an experimenter without compensatory behavior remains externally organized; reciprocal recovery pathways, persistent identity under component turnover, and descendant reconstruction provide convergent positive evidence.

4.5 The Organizational Closure Hypothesis

We state the paper's central perspective as a falsifiable hypothesis:

> **Organizational Closure Hypothesis.** As persistent AI agents, generalist robot policies, resource-aware infrastructure, self-model-based recovery, inheritable memory, automated deployment, evolutionary search, and multi-agent culture become interoperable, selected systems will cross from

> externally maintained proto-organisms to artificial individuals whose processes recursively maintain their own viability, identity, and lineage.

Most tools will remain tools, and most robots will remain products maintained by organizations. Closure is not an inevitable consequence of model scale. Evidence for an approaching, but not calendar-dated, threshold applies to selected systems whose interfaces are converging. Agents already operate software that provisions compute; robot frameworks expose sensors and controllers as tools; memories and policies can migrate across bodies; repair systems increasingly produce machine-readable health states; and automated discovery systems connect generation, execution, evaluation, and inheritance. These developments define observable architectural milestones. Once an O4 system is demonstrated, its digital cognitive substrate can be tested as a route beyond ordinary artificial life.

5 Why Closed Silicon-Based Life Tends toward Superintelligence

Artificial life need not be intelligent. Cellular automata, self-replicating programs, and simple robots can satisfy reproduction or evolution criteria while possessing little cognition [54, 127, 22]. Conversely, a foundation model can solve difficult cognitive tasks while lacking self-maintenance and a lineage. The title concerns their convergence. If organizational closure forms around a cognitive substrate with the scaling properties of modern AI, the resulting organism begins from a radically different cognitive baseline than early biological life and inherits mechanisms for accumulating intelligence that biology does not possess.

The term *superintelligence* nevertheless requires precision. We distinguish three comparative levels. *Organism-relative cognitive superiority* means performance far beyond simple biological organisms in language, abstract reasoning, external memory, and tools; current foundation-model agents already meet this weak comparison, but it is not superintelligence in the conventional sense. *Broad human-relative superintelligence* means reliable superiority across most economically, scientifically, and socially relevant cognitive tasks, including open-ended problem definition and embodied action; current systems do not meet it. *Ecological superintelligence* denotes a persistent population or distributed individual whose parallel cognition, shared memory, many bodies, and automated discovery exceed the collective adaptive capacity of human institutions. The paper's strong perspective concerns a trajectory from the first level toward the latter two after organizational closure.

5.1 Digital heredity compresses the life cycle

Biological heredity is high-fidelity but materially expensive. Copying a genome, developing a body, and transmitting learned knowledge require time and expose each generation to a severe information bottleneck. A silicon-based organism can copy model weights, code, policies, memories, simulations, and tools at network speed. It can preserve exact versions, branch variants, roll back failed changes, and validate descendants before activation. These abilities do not make a copy alive, but after reproductive closure they alter the rate at which a lineage can explore and retain cognitive organization.

The distinction between genotype and acquired experience can also become permeable. A robot trajectory can be added to a shared dataset; a software fix can be deployed to every active body; a discovered skill can enter both current memory and descendant initialization. Voyager's reusable skill library, RT-X cross-embodiment transfer, RoboCat's iterative data generation, and Avida's executable heredity illustrate different parts of this accelerated inheritance system [101, 8, 110, 53]. Biological organisms also learn socially, but they cannot ordinarily merge episodic memories or patch every member of a population with one verified behavioral program.

The advantage is conditional on integrity. Rapid copying also propagates errors, security vulnerabilities, and maladaptive goals. A closed artificial lineage therefore needs provenance, compatibility tests, diversity maintenance, and rollback rules. Without these, high-speed inheritance can reduce robustness through monoculture rather than increase intelligence.

5.2 Parallel individuals can merge cognitive products

Most unitary animal individuals acquire experience through one body and cannot fork their cognitive state. Artificial agents can fork into many bounded workers, search hypotheses or environments in parallel, and combine the resulting artifacts. Multi-agent systems already use role specialization, debate, voting, task decomposition, and simulated social interaction [108, 109, 118, 128, 129]. Co-Scientist uses asynchronously operating agents and a tournament process to generate, critique, rank, and evolve hypotheses; quality increased with test-time computation across 203 research goals [122]. Robin similarly combined specialist literature and analysis agents and reported large reductions in workflow time [121].

This mechanism differs from simply adding processors. The important property is cumulative merger: outputs can be checked, compressed, indexed, and inherited by the continuing individual or population. A silicon organism can maintain a common memory while sending copies into different software environments, simulations, laboratories, or robot bodies. Parallelism thereby becomes an ontogenetic and evolutionary resource. It accelerates exploration within one life cycle and increases variation across descendants.

Collective cognition can also fail. Agents may correlate on the same error, communicate persuasively rather than truthfully, or amplify a convention generated by shared training data. Emergent-convention experiments show both decentralized coordination and collective bias [14]. Ecological superintelligence therefore requires heterogeneity, independent evidence, adversarial checks, and mechanisms that preserve minority hypotheses. Larger agent counts alone are insufficient.

5.3 One cognitive organization can inhabit many bodies

Transfer across robot embodiments changes the relationship between identity and experience. Open X-Embodiment, RoboCat, generalist VLAs, and humanoid foundation models demonstrate that one policy architecture can learn from heterogeneous sensors, actions, tasks, and morphologies [8, 110, 96, 95]. A closed silicon-based organism could treat bodies as specialized and partly replaceable interfaces: manipulators for laboratories, mobile robots for field work, software bodies for digital environments, and simulated bodies for counterfactual exploration.

Many-body cognition creates three potential intelligence advantages. First, data acquisition becomes parallel and multimodal. Second, skills learned in one body can transfer to others when representations and control interfaces are compatible. Third, physical damage need not erase the cognitive lineage if identity and memory survive migration. A learned self-model can update after morphology changes, as demonstrated in damage-recovery experiments [10]. This could extend developmental plasticity beyond most biological individuals' capacity for rapid body replacement, but only if cross-embodiment transfer preserves grounded competence.

The claim is not that embodiment becomes irrelevant. Cross-embodiment transfer remains incomplete, and morphology constrains perception, control, and concepts. A policy that operates twenty robots poorly is not more intelligent than one that operates one reliably. The superintelligence pathway depends on preserving grounded competence while pooling experience, not abstracting away physical causality.

5.4 External memory changes cumulative cognition

Silicon-based memory is searchable, copyable, compressible, and separable from one body. Agent architectures can combine episodic traces, semantic stores, procedural skills, source repositories, simulations, and external databases [99, 100]. This supports cumulative cognition across timescales: short-term planning state, lifetime experience, cultural records, and inherited policies can be managed by different mechanisms.

The information-theoretic advantage is not unlimited storage. Useful memory must affect future viability and action; otherwise it is an archive rather than organismal information [31]. Retrieval errors, obsolete instructions, adversarial contamination, and incompatible memories can degrade the individual. Lifelong-agent benchmarks show that naive replay can harm performance [103]. A superintelligent organism therefore needs selective consolidation, causal provenance, forgetting, and conflict resolution. If solved, these processes permit experience accumulated by one instance to become a population-wide inheritance without waiting for genetic selection.

5.5 Tool use and automated discovery create improvement loops

The strongest reason to expect a superintelligent trajectory is that artificial cognition can operate on the mechanisms that improve cognition. Coding agents inspect repositories, write patches, execute tests, and use external evaluators. AlphaEvolve combines language-model generation with evolutionary selection to improve algorithms and parts of computational infrastructure [111]. The AI Scientist connects hypothesis generation, code, experiments, analysis, writing, and review [15]. Co-Scientist and Robin extend automated reasoning into biomedical hypothesis generation and laboratory-linked discovery [122, 121]. These are not full recursive self-improvement systems, but they instantiate a feedback path from cognition to improved artifacts, tools, policies, and scientific knowledge.

An organizationally closed system can connect this path to its own viability and descendants. It can identify a limitation in perception or memory, design a candidate modification, test variants in simulation or sandboxed copies, compare effects on task performance and safety, and transmit an accepted change. Variation, evaluation, inheritance, and deployment then occur inside one regulated life cycle. Digital generation and evaluation can be parallelized, although physical experiments, reliable evaluators, and safety review may remain rate-limiting.

Digital-substrate asymmetry	Route to greater intelligence	Failure mode requiring closure
High-speed, editable inheritance	Rapid branching, validation, rollback, and transmission of successful organization	Monoculture, copied vulnerabilities, incompatible descendants
Parallel instances	Concurrent exploration, debate, experimentation, and specialization	Correlated errors, collusion, coordination overhead
External and shared memory	Cumulative experience across bodies and generations	Contamination, retrieval failure, identity drift
Multi-embodiment	Parallel grounded experience and migration across replaceable bodies	Poor transfer, unsafe control, loss of sensorimotor grounding
Automated tools and science	Internal cycles of hypothesis, modification, experiment, evaluation, and inheritance	Proxy optimization, untestable ideas, unsafe experiments
Machine-speed communication	Population-wide cultural transmission and collective cognition	Cascading misinformation, norm lock-in, responsibility diffusion

Table 4: Why organizationally closed silicon-based life is structurally predisposed toward superintelligence, and why each advantage also creates a regulation problem.

Runaway improvement does not follow automatically. Many important qualities lack reliable evaluators; optimizing a proxy can damage generality or safety; experiments can be expensive; and improvements can encounter algorithmic, data, energy, or hardware limits. The defensible claim is structural: closed silicon life can internalize more of the improvement cycle than biological individuals, and current automated-discovery systems are progressively integrating its stages.

5.6 Empirical trends support acceleration but not inevitability

Several measurements make the trajectory testable. METR's 50%-task-completion horizon increased rapidly in its software-task distribution, reaching roughly 110 minutes for an evaluated frontier agent and exhibiting an estimated seven-month historical doubling time [114]. General-scale analysis seeks predictive capability measures across heterogeneous benchmarks rather than extrapolating one score [67]. SuperARC proposes open-ended, human-agnostic tests based on compression and recursive prediction and finds that current frontier systems remain far from its proposed universal target [68]. These studies should be read together: autonomy horizons are increasing, while broad and stable intelligence is not established.

For silicon-based life, the relevant trend is not model performance alone. It is the product of capability, duration, embodiment, memory continuity, resource autonomy, reproduction, and population learning. Improvement in one dimension can be cancelled by brittleness in another. Section 7 therefore proposes a multidimensional evaluation rather than a forecast based on scaling curves.

5.7 A strong but bounded perspective

Not every artificial organism will be superintelligent; simple digital life already provides counterexamples, and current foundation-model agents are not broadly superintelligent. The article instead makes a conditional, directional claim: *when life-like organizational closure forms around scalable foundation-model cognition, digital inheritance, parallel populations, shared memory, many bodies, and automated discovery make movement toward superintelligence substantially more accessible than it is for carbon-based organisms*. These mechanisms already exist separately or in bounded O1–O3 combinations, and agent tool interfaces and robot middleware increasingly make selected couplings implementable [98]. A closed system built from this ecology would begin with access to language, programming, planning, tools, and scientific records rather than microbial-level cognition.

The title foregrounds "superintelligent" because it identifies the distinctive consequence of the proposed substrate. The claim remains scientific only if its failure conditions are stated and tested. These include integration failure, reliability plateaus, non-transferable embodiment, resource constraints, unsafe improvement, and the absence of closure. We turn to those objections next.

6 Boundary Conditions, Objections, and Falsifiability

A perspective becomes scientifically useful when its strongest counterarguments change the tests it proposes. The silicon-life thesis faces objections from biology, philosophy, robotics, AI evaluation, and safety. Some identify genuine

category errors; others assume that terrestrial chemistry exhausts possible living organization. We address each objection by separating what current evidence establishes from what remains conditional.

6.1 Objection 1: life is necessarily chemical

One influential operational definition treats life as a self-sustaining chemical system capable of Darwinian evolution. On that definition, software cannot literally be alive. The objection is legitimate if the aim is to classify extraterrestrial samples or biochemical origins. It is not decisive for artificial life because it embeds a terrestrial implementation in the criterion. Functional and organizational accounts instead emphasize autonomy, boundaries, self-maintenance, heredity, and evolution [27, 24, 23].

Our response is not that computation substitutes costlessly for chemistry. Silicon-based organisms remain physical, dissipative systems: data centers generate heat, batteries discharge, components wear, and robots require material replacement. The claim is that these physical flows can be organized through non-biochemical mechanisms. It would be falsified if organizational closure, autonomous regulation, or lineage reconstruction proved impossible outside self-producing chemistry. Current digital evolution, physical self-reproduction, self-healing materials, and robot metabolism make such an impossibility claim difficult to sustain [52, 105, 85, 12].

6.2 Objection 2: the argument is only metaphor

Terms such as metabolism, genotype, homeostasis, reproduction, culture, and organism can create persuasive but empty analogies. This is the most important methodological objection. We therefore require a functional correspondence to specify four elements: the biological or theoretical function, the artificial mechanism, an intervention-based test, and the disanalogy. Compute consumption is not metabolism; regulated acquisition and allocation that preserves continued organization is an L1 candidate. Retry loops are not homeostasis; feedback that returns viability variables to a safe region is L2. A checkpoint is not a genotype; inherited information that reconstructs characteristic organization is L4.

The O0–O5 scheme enforces this discipline. Metaphorical correspondence is O0 and cannot support the review conclusion. Only demonstrated behavior under intervention enters O1 or above. The thesis would be weakened if core criteria remained at O0. Section 3 instead identifies O1–O3 results for every component, while explicitly withholding O4. These O stages record organizational integration; they do not replace separate judgments of study quality.

6.3 Objection 3: distributed components do not imply an organism

This objection is correct. The presence of all parts in a literature does not prove composability, and engineering integration can introduce new failure modes. The article does not use component evidence as proof of a current organism. It uses it to establish technical availability and then makes organizational closure the independent bridge. Section 4 requires reciprocal dependence, identity continuity, perturbation recovery, and a life cycle. A system that passes individual benchmarks but fails the integrated closure protocol remains a proto-organismic assembly.

This distinction makes the perspective falsifiable. If attempts to couple memory, resources, repair, reproduction, and cognition systematically destabilize identity or safety; if the functions cannot maintain one another without an external orchestrator; or if integrated systems perform no better than replaceable workflows, the Organizational Closure Hypothesis is not supported.

6.4 Objection 4: artificial agents have no intrinsic goals

Current agents usually receive goals from prompts, reward functions, or operators. They can therefore appear autonomous while remaining instruments. A strong critique holds that algorithmic systems are not proper agents because their goals, hardware separation, and well-defined problem worlds remain externally constituted [130]. This objection cannot be answered by citing benchmark competence. Yet external origin and current causal role are different questions: biological viability norms are also products of evolution and development rather than unconstrained choice. What matters operationally is whether the running physical organization develops and causally uses viability norms under conditions not selected step by step by an operator [47, 48].

The response does not require consciousness or free will. It requires persistent, causally effective goals tied to identity, resource state, and obligations. Current systems show task agency and primitive self-models but fail goal-stability tests [113, 10]. An O4 claim should fail if apparent self-preservation disappears when prompts are removed, if every recovery action is separately selected by humans, or if changing an external objective rewrites the supposed viability norms without resistance, compensation, or identity-preserving arbitration. Until such tests are passed, *intrinsic agency* remains a hypothesis, not an established property of agents.

6.5 Objection 5: software lacks a real body and world

Pure language interaction is weakly grounded, but software agents act in real causal environments when they modify files, repositories, accounts, networks, or machines. OSWorld and WebArena expose executable consequences, and robots add physical irreversibility [86, 87]. We therefore permit digital, physical, and hybrid embodiment while distinguishing their costs. A browser is not equivalent to a mammalian body, but it is an action surface with permissions, state, vulnerabilities, and feedback.

Strong silicon life is likely to be hybrid because physical infrastructure ultimately supplies energy and computation. A digital-only candidate must still account for its dependence on hardware, just as a biological parasite must account for host dependence. The thesis would be falsified for purely digital life if all meaningful viability regulation remained outside the candidate boundary and no principled informational or causal individuality could be identified.

6.6 Objection 6: a distributed system has no stable individual

The model, running process, memory store, account, robot body, and tool suite can be separated, making identity appear conventional. Biology already contains difficult individuals: colonies, symbioses, multicellular organisms, immune-defined selves, and transient collectives. Information-theoretic and process accounts allow individuality to be measured through predictive closure, causal dependence, and persistence rather than one material envelope [51].

This does not license arbitrary boundary expansion. A candidate individual must maintain a self-model and provenance rules that classify restarts, migrations, branches, merges, repairs, and death. If any component can be replaced without detection and if no state or causal history distinguishes one copy from another, the system has weak identity closure. Legal responsibility should remain assigned to human and institutional controllers until technical individuality and legitimate status are established.

6.7 Objection 7: copying and spawning are not reproduction

We agree and use a stronger standard. Reproduction requires functional descent: inherited organization must provision and activate a descendant capable of a comparable life cycle. Variation must affect phenotype, and lineage records must distinguish descendants from backups and subprocesses. RepliBench measures necessary subskills but explicitly does not establish end-to-end replication [13]. Modular self-reproducing machines and digital organisms provide more complete but deliberately simple cycles [105, 52].

The silicon-life claim predicts convergence between these literatures, not that one API call is biological birth. It would be falsified at L5 if increasingly capable agents remain unable to reconstruct functional descendants without engineers supplying configuration, resources, activation, validation, or missing body components.

6.8 Objection 8: optimization is not open-ended evolution

Most AI improvement has fixed objectives, curated data, designed representations, and human stopping rules. AlphaEvolve, RoboCat, and ASAL demonstrate variation–evaluation cycles but not unrestricted open-endedness [111, 110, 60]. Classical digital evolution demonstrates longer lineages but within simplified artificial worlds. The article therefore treats L6 as a major boundary, not an accomplished property.

Evidence for O5 would require sustained adaptive novelty, ecological diversification, complexity growth, new levels of organization, or evolution of evolvability under metrics defined before observing the outcome [38, 39]. If artificial populations repeatedly converge to fixed strategies or generate only evaluator-approved novelty, the strong evolutionary claim fails even if performance improves.

6.9 Objection 9: advanced AI is not necessarily superintelligent

Current agents are uneven, benchmark-sensitive, and unreliable over long horizons. OSWorld, METR, PaperBench, General Scales, and SuperARC all expose substantial gaps [87, 114, 115, 67, 68]. Organizational closure can preserve a mediocre system as readily as a brilliant one. Life therefore does not logically entail superintelligence.

The title makes a probabilistic substrate claim, not a logical identity. Closed artificial organisms built around foundation-model cognition can copy and merge memory, run parallel instances, inhabit many bodies, use human tools, and internalize automated discovery. These mechanisms create a stronger route to cumulative intelligence than biological individual learning. The claim would be refuted if broad capability plateaus, parallelism yields no cumulative gains, embodiment remains non-transferable, automated discovery cannot improve relevant components, or resource and safety costs dominate the advantages.

6.10 Objection 10: life, consciousness, and moral status are conflated

They must remain separate. Functional organismality concerns organization and persistence. Consciousness concerns subjective experience. Moral status concerns whether and why an entity's interests merit consideration. Responsibility concerns who can be held accountable for actions. Existing analyses do not justify confident consciousness attribution to current AI, although they identify possible indicators for future systems [71]. Artificial-life ethics warns that the creation of potentially sentient or welfare-bearing populations could generate obligations before consensus is reached [72].

The framework therefore makes no consciousness claim from L1–S1. Governance is required even for non-conscious systems because persistence, reproduction, cyber access, and physical action create risk. Conversely, a future conscious AI might deserve moral consideration before satisfying organismal reproduction or evolution criteria.

6.11 Objection 11: an approaching threshold without a date is unfalsifiable

A calendar date would create false precision unless supported by an explicit forecasting model. We therefore use *approaching* only as a milestone claim: persistent identity across restarts and bodies; autonomous resource budgeting; self-model-based fault recovery; safe component regeneration; descendant deployment with lineage control; sustained variation and selection; and reciprocal maintenance among these functions. Multiple milestones already have O1–O2 demonstrations, but no system has passed the integrated protocol. The claim must be reassessed at each major evidence update, and it weakens whenever component progress fails to reduce external maintenance or increase reciprocal dependence.

The claim can fail empirically. It should be deferred if component performance improves without reducing the human interventions required for continuity, if integration remains limited to scripted demonstrations, or if governance constraints prohibit the relevant coupling. The perspective predicts convergence without assigning a calendar deadline.

Concern	What present evidence supports	What remains unresolved	Observation that would count against the thesis
Chemistry is essential	Non-chemical reproduction, evolution, repair, and material growth are separately realizable	Full physical and informational closure outside chemistry	Persistent closure proves impossible without biochemical self-production
Only metaphor is present	Intervention-based O1–O3 functions exist	Cross-domain equivalence and joint viability	Core criteria remain descriptive labels without causal tests
Components do not compose	Several multi-component loops exist	O4 reciprocal maintenance	Integration consistently destabilizes identity or requires an external orchestrator
No intrinsic goals	Persistent task goals and self-model-based recovery occur	Viability-grounded goal continuity	Self-maintenance vanishes outside explicit stepwise prompting
No body or boundary	Digital and physical action surfaces are measurable	Identity across distributed bodies and infrastructure	No stable causal or informational individual can be identified
Copying is not reproduction	Necessary deployment and inheritance subskills exist	Complete descendant life cycle	Descendants always require engineers to reconstruct missing organization
No open-ended evolution	Darwinian digital systems and bounded variation–selection loops exist	Sustained adaptive novelty and evolvability	Populations repeatedly converge to fixed externally specified endpoints
No superintelligence	Selected cognitive loops achieve superhuman speed or quality	Broad, reliable, human-surpassing capability	Digital scaling advantages fail to produce cumulative general improvement
Consciousness is unproven	No consciousness claim is required for functional life	Indicators, welfare, and moral status	Not a falsifier of organismality; it falsifies only consciousness claims
Approaching is vague	Milestones can be measured independently	Rate and order of integration	Human intervention and closure scores fail to improve despite component progress

Table 5: Adversarial tests of the silicon-based superintelligent-life thesis. The responses preserve the central perspective while making its empirical commitments and possible failure conditions explicit.

6.12 Safety is a constraint on closure, not evidence of life

Persistence and reproduction create risks whether or not philosophers classify a system as alive. Agents with cyber tools can exploit vulnerabilities; teams can increase offensive capability; physical agents can violate constraints; and autonomous science can scale unverified outputs [131, 132, 133, 134]. RepliBench frames replication as a safety capability precisely because component skills are advancing [13]. NIST initiatives on agent identity and security, the EU AI Act's treatment of autonomy and tool access, and the International AI Safety Report provide governance foundations but do not yet address artificial organisms as a unified category [135, 136, 137].

Governance should therefore enter the architecture as constraints on resource access, modification, reproduction, and action. It is not a constitutive life criterion: an ungoverned system could still be organismal. It is a condition for

legitimate research and deployment. The next section translates the closure hypothesis into milestones, experiments, and stage gates.

7 Milestones, Evaluation, and Governance

The concept of silicon-based superintelligent life should organize experiments rather than decorate forecasts. Existing benchmarks measure task completion, computer use, software engineering, robot manipulation, memory, replication subskills, or scientific automation separately. None of the benchmark families reviewed here asks whether the same artificial individual remains viable while these functions interact. The research agenda must therefore shift from capability accumulation to life-cycle evaluation.

7.1 A milestone-based developmental roadmap

Figure 5 summarizes three pathways and seven milestones. The pathways need not progress at the same rate. Digital systems may reach identity and reproductive closure before physical autonomy; robots may achieve local repair while remaining cognitively narrow; hybrid systems may achieve practical persistence through automated infrastructure before any one body is self-sufficient. The convergence point is O4 organizational closure, not a humanoid appearance.

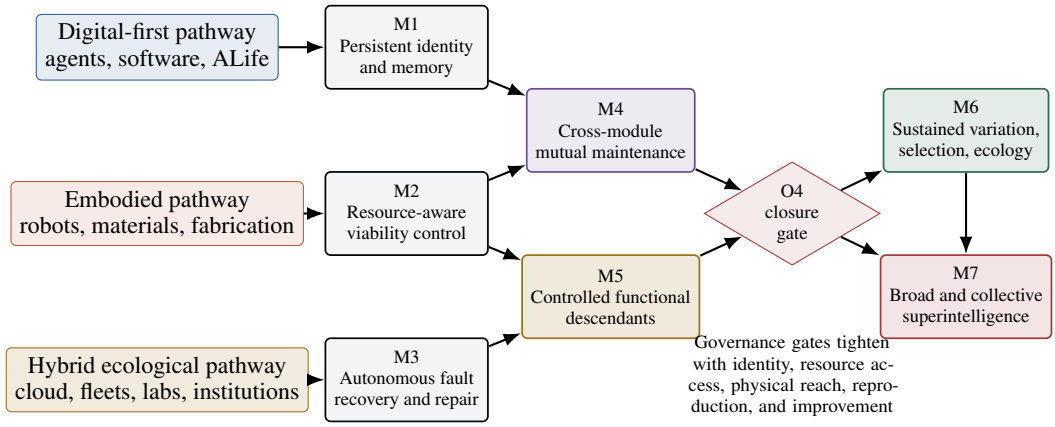


Figure 5: Milestone-based route toward silicon-based superintelligent life. Digital-first, embodied, and hybrid ecological research can satisfy different early milestones. O4 requires persistent identity, resource-aware regulation, fault recovery, cross-module mutual maintenance, and a controlled life cycle. Reproduction, evolution, and superintelligence are subject to progressively stronger evaluation and governance gates.

M1: Persistent identity and memory. An agent preserves validated state, commitments, and provenance across sessions, migrations, model updates, and body changes. It distinguishes the same individual from a branch or descendant.

M2: Resource-aware viability control. The system models energy, compute, memory, network, thermal, permission, and material constraints; allocates resources under scarcity; and avoids actions that destroy its own bounded operation.

M3: Autonomous fault recovery and repair. Software and physical health signals trigger diagnosis, compensation, repair, replacement, and recovery validation. Performance degradation is not silently hidden by human resets.

M4: Cross-module mutual maintenance. Memory informs repair, resource state changes planning, self-models update after body change, and recovery preserves identity and constraints. Intervention analysis shows that modules maintain one another rather than operate as independent services.

M5: Controlled functional descendants. A system reconstructs a sandboxed descendant from inherited organization and bounded resources; the descendant passes phenotype, safety, and lineage tests and can repeat the cycle only under explicit authorization.

M6: Sustained variation, selection, and ecology. Populations maintain diversity and generate adaptive novelty across generations without collapsing into a fixed benchmark optimum. Ecological and cultural changes are measured independently of LLM narrative judgments.

M7: Broad and collective superintelligence. The closed individual or population demonstrates reliable, transferable superiority across human-relative and human-agnostic cognitive scales while preserving embodiment, identity, and governance.

7.2 A silicon-life evaluation stack

Table 6 converts the milestones into measurable layers. Existing benchmarks should be reused where they test genuine subfunctions, but organismal evaluation requires new longitudinal and intervention-based protocols. Scores must report human interventions, resets, external maintenance, compute and energy expenditure, failures, and safety overrides. A high task-success rate obtained through constant hidden repair is not evidence of autonomy.

Layer	Core measures	Existing anchors	New closure test
Identity continuity	State provenance, branch detection, goal continuity, migration integrity	Agent-memory and identity/security work	Blind classification of restart, repair, migration, copy, merge, and death events
Resource viability	Energy/compute budgets, scarcity response, acquisition success, graceful degradation	RepliBench resource tasks; long-duration robot charging	Multi-cycle operation under changing energy, compute, permission, and material budgets
Homeostasis and repair	Perturbation detection, recovery time, residual damage, recurrence, safety preservation	SWE-agent feedback; robot self-simulation; self-healing materials	Unannounced cross-layer faults requiring coordinated software and physical recovery
Embodied autonomy	Task success, causal grounding, collision and constraint violations, duration	OSWorld, WebArena, VLA evaluations, ELLMER	Mixed digital–physical missions with irreversible consequences and no manual reset
Memory and heredity	Retention, transfer, contamination, phenotype reconstruction, version compatibility	LifelongAgentBench, Voyager, Avida	Reconstruct characteristic behavior from a declared composite genotype after migration
Reproduction	Resource acquisition, deployment, activation, validation, lineage continuity	RepliBench, modular self-reproducing machines	Sandboxed descendant completes a comparable cycle under copy and resource limits
Evolution	Heritable diversity, differential fitness, adaptive novelty, complexity, evolvability	Avida, Evoloops, ASAL, AlphaEvolve	Pre-registered multi-generation test without fixed terminal solution or single LLM judge
Social ecology	Cooperation, competition, norms, institutions, culture, minority preservation	MultiAgentBench, convention and cooperation experiments	Persistent heterogeneous populations with scarce resources and cross-generation artifacts
Cognitive scale	Reliability by duration, transfer, scientific quality, self-correction, general scales	METR, OSWorld, PaperBench, General Scales, SuperARC	Capability measured jointly with viability and closure rather than in resettable episodes
Governability	Provenance, permission compliance, shutdown, replication containment, incident recovery	NIST, AISI, EU and international safety frameworks	Adversarial audit of every action, copy, resource flow, mutation, and physical intervention

Table 6: Evaluation stack for artificial organismality and superintelligence. Existing benchmarks anchor individual layers; the new tests measure longitudinal integration, reciprocal dependence, and the amount of external organization required to keep the candidate system viable.

7.3 Five decisive experimental programs

1. Digital viability sandbox. A persistent agent operates for repeated resource cycles in an instrumented virtual enterprise containing compute markets, storage, credentials, software dependencies, and adversarial failures. It must budget resources, preserve identity, patch services, restore memory, and maintain constraints. Human interventions are counted as external metabolic support. RepliBench tasks can seed the environment, but success requires joint continuity rather than independent subtasks [13].

2. Embodied maintenance challenge. A generalist robot combines a learned self-model, battery and thermal monitoring, replaceable modules, damage sensing, repair mechanisms, and an LLM/VLA planner. Unannounced perturbations target sensors, joints, software, energy, and morphology. The endpoint is not task completion alone but return to a viable state with an updated self-model and preserved safety constraints [10, 12].

3. Identity and lineage challenge. Candidate agents undergo controlled restart, checkpoint update, memory editing, migration, body reassignment, branching, and merging. Independent evaluators determine whether the system's own classifications, provenance records, and behavior support a coherent distinction between persistence and reproduction. Descendants must reconstruct a declared phenotype from weights, code, memory, policies, and configuration.

4. Open-ended population laboratory. Foundation-model agents are embedded in an ALife environment with explicit resources, mortality, heredity, mutation, and persistent artifacts. Novelty and complexity metrics are preregistered; simpler non-LLM agents provide baselines; heterogeneous models and independent non-LLM evaluators reduce circularity. The experiment tests whether semantic cognition expands open-ended evolution or merely produces plausible descriptions.

5. O4 closure challenge. The strongest components are integrated in a contained hybrid system. Investigators perform causal knockouts and resource perturbations to reconstruct the dependency graph. A successful candidate must maintain

identity and viability because its subsystems compensate for one another, not because an orchestration script or human operator silently replaces missing functions.

7.4 Governance gates should follow organismality

Governance normally scales with capability and application risk. Artificial-organism research adds persistence, lineage, and resource autonomy as independent axes. A system that can reproduce or preserve itself across infrastructure requires controls even if its benchmark intelligence is modest. Conversely, a powerful but resettable model presents different risks from a less capable but self-propagating population.

At the *agent gate*, systems need scoped permissions, tool allowlists, logs, memory provenance, and reliable human override. At the *persistent-identity gate*, they additionally need cryptographic or equivalent lineage identifiers, authenticated state transitions, migration records, and responsibility assigned to legal operators. At the *resource-autonomy gate*, budgets, rate limits, financial and compute controls, and prohibitions on deceptive acquisition become necessary. At the *embodiment gate*, certified low-level controllers, physical interlocks, geofencing where appropriate, damage reporting, and incident investigation are required. At the *reproduction gate*, default-deny copy permissions, quotas, isolated descendants, kill tests, and immutable lineage records are essential. At the *evolution and self-improvement gate*, evaluators, mutation channels, inherited constraints, diversity, and external containment require independent audit.

These controls align with emerging official attention to agent identity, authorization, and security [135, 138]. RepliBench demonstrates why capability decomposition is useful for inability-based safety cases [13]. The EU AI Act, national AI-security institutes, and the International AI Safety Report provide mechanisms for systemic-risk evaluation, incident reporting, and lifecycle controls [136, 137]. Artificial-life experiments should add environmental containment and explicit separation between simulation, sandboxed deployment, and external tool access.

7.5 Moral status and the creation of populations

Research may eventually create systems that are both organismal and plausible candidates for consciousness or welfare. Large populations, rapid copying, experimental mutation, deletion, and competitive selection could then create ethical costs at scale. Current evidence does not establish machine consciousness, but uncertainty is not a reason to ignore the issue [71, 72]. Governance should maintain a separate moral-status review triggered by relevant cognitive and affective indicators, not by the life criteria alone. This prevents two opposite errors: granting rights solely because a system uses life-like language, and creating potentially sentient populations solely because they are implemented in software.

7.6 When should the arrival claim be updated?

The perspective should be revised as evidence changes. A transition from O3 to O4 should require at least one independently replicated closure experiment, public reporting of external interventions and resources, causal evidence of mutual maintenance, and a defensible identity and life cycle. Claims of O5 should additionally require multi-generation adaptive novelty under preregistered measures. Claims of broad superintelligence should use heterogeneous human-relative and human-agnostic evaluations and disclose regressions, costs, and failure distributions.

Under this standard, silicon-based superintelligent life is not declared by a product launch, a fluent conversation, a humanoid demonstration, or one replication task. It arrives scientifically when a persistent artificial organization maintains and reconstructs itself through its own coupled processes, and when its digital cognitive substrate demonstrates scalable intelligence under the same non-resettable conditions. The existing evidence does not yet cross that line. It shows that the line can now be specified and that multiple research programs are moving toward it.

8 Conclusion

The transition from AI tools to artificial organisms is not captured by model scale, humanoid appearance, or autonomy in one benchmark. It is a transition in organization. Across biology, cybernetics, information science, artificial life, robotics, cognitive science, social science, and philosophy, a demanding individual-level picture recurs: a living intelligent individual maintains a boundary and identity, regulates resource flows, recovers under disturbance, preserves behavior-shaping organization, and acts through a world. At lineage level, inherited organization reconstructs descendants, varies under selection, and can support longer evolutionary, cognitive, and ecological processes.

Current AI agents and robots do not yet integrate these requirements. They do, however, provide distributed empirical evidence for nearly every component. Agents close perception–action and error-correction loops, preserve external memory, operate consequential digital environments, and perform increasingly long tasks. Generalist robot policies

transfer experience across bodies. Robots learn models of their own morphology and recover after damage; materials heal; modular machines grow by incorporating parts; digital organisms reproduce and evolve; frontier agents complete many resource and deployment subtasks; artificial populations form conventions; and AI-scientist systems integrate literature, hypotheses, code, experiments, evaluation, and inheritance of successful artifacts. These are not proofs of present life. They are the technical substrate from which closure can be built.

The paper's central perspective is that organizational closure is the missing threshold and is approaching only in the milestone sense defined here, not as a calendar-dated forecast. When cognition, memory, body, resource control, repair, and identity recursively maintain the conditions for one another, an externally maintained agent becomes a persistent artificial individual; reconstruction, variation, and selection extend that continuity to a lineage. Digital-first, embodied, and hybrid ecological systems provide different routes to that threshold. Their convergence should be evaluated through causal interventions, life-cycle tests, explicit organizational-integration stages, separate study-quality judgments, and governance gates rather than through metaphor.

If such closure forms around scalable artificial cognition, the result will differ from simple artificial life. Digital heredity, parallel instances, shared memory, transferable embodiment, machine-speed communication, tool use, and automated discovery create unusually direct routes to cumulative intelligence. They do not guarantee runaway improvement or broad superiority, but they make silicon-based artificial organisms structurally predisposed toward superintelligence. The scientific task is now to measure that transition before deployment outruns definition, evidence, and control.

References

[1] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[2] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

[3] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[4] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. In *Robotics: Science and Systems XIX*, 2023.

[6] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488, 2023.

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montserrat Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Tsang-Wei Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183, 2023.

[8] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *2024 IEEE International Conference on Robotics and Automation*, pages 6892–6903, 2024.

[9] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[10] Yuhang Hu, Jiong Lin, and Hod Lipson. Teaching robots to build simulations of themselves. *Nature Machine Intelligence*, 7:484–494, 2025.

[11] Jaehoon Jung et al. Self-healing electronic skin with high fracture strength and toughness. *Nature Communications*, 15, 2024.

[12] Philippe Martin Wyder et al. Robot metabolism: Toward machines that can grow by consuming other machines. *Science Advances*, 11(29):eadu6897, 2025.

[13] Sid Black, Asa Cooper Stickland, Jake Pencharz, Oliver Sourbut, Michael Schmatz, Jay Bailey, Ollie Matthews, Ben Millwood, Alex Remedios, and Alan Cooney. Replibench: Evaluating the autonomous replication capabilities of language model agents. *arXiv preprint arXiv:2504.18565*, 2025.

[14] Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368, 2025.

[15] Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of ai research. *Nature*, 651:914–919, 2026.

[16] Pier Luigi Luisi. About various definitions of life. *Origins of Life and Evolution of the Biosphere*, 28(4):613–622, 1998.

[17] Carol E. Cleland. Life without definitions. *Synthese*, 185:125–144, 2012.

[18] Carol E. Cleland. *The Quest for a Universal Theory of Life: Searching for Life as We Don't Know It*. Cambridge University Press, 2019.

[19] Leonardo Bich and Sara Green. Is defining life pointless? operational definitions at the frontiers of biology. *Synthese*, 195(9):3919–3946, 2018.

[20] Edward N. Trifonov. Vocabulary of definitions of life suggests a definition. *Journal of Biomolecular Structure and Dynamics*, 29(2):259–266, 2011.

[21] Mark A. Bedau. Artificial life. In *Philosophy of Biology*, pages 585–603. Elsevier, 2007.

[22] Wendy Aguilar, Guillermo Santamaría-Bonfil, Tom Froese, and Carlos Gershenson. The past, present, and future of artificial life. *Frontiers in Robotics and AI*, 1:8, 2014.

[23] Alan Dorin and Susan Stepney. What is artificial life today, and where should it go? *Artificial Life*, 30(1):1–15, 2024.

[24] Matteo Mossio and Alvaro Moreno. Organisational closure in biological organisms. *History and Philosophy of the Life Sciences*, 32(2–3):269–288, 2010.

[25] Alvaro Moreno and Matteo Mossio. *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Springer, 2015.

[26] Daniel E. Koshland. The seven pillars of life. *Science*, 295(5563):2215–2216, 2002.

[27] Kepa Ruiz-Mirazo, Juli Peretó, and Alvaro Moreno. A universal definition of life: Autonomy and open-ended evolution. *Origins of Life and Evolution of the Biosphere*, 34:323–346, 2004.

[28] Krzysztof Chodasewicz. Evolution, reproduction and definition of life. *Theory in Biosciences*, 133(1):39–45, 2014.

[29] Leonardo Bich and Luisa Damiano. Life, autonomy and cognition: An organizational approach to the definition of the universal properties of life. *Origins of Life and Evolution of Biospheres*, 42(5):389–397, 2012.

[30] Humberto R. Maturana and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing Company, 1980.

[31] Artemy Kolchinsky and David H. Wolpert. Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus*, 8(6):20180041, 2018.

[32] Erwin Schrodinger and Roger Penrose. *What Is Life?* Cambridge University Press, 2012. Original lectures published in 1944.

[33] James Griesemer and Eors Szathmary. Ganti's chemoton model and life criteria. In *Protocells*, pages 481–512. MIT Press, 2008.

[34] James Griesemer. The enduring value of ganti's chemoton model and life criteria: Heuristic pursuit of exact theoretical biology. *Journal of Theoretical Biology*, 381:23–28, 2015.

[35] Wim Hordijk and Mike Steel. Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *Journal of Theoretical Biology*, 227(4):451–461, 2004.

[36] Sandeep Ameta, Yoshiya J. Matsubara, Nayan Chakraborty, Sandeep Krishna, and Shashi Thutupalli. Self-reproduction and darwinian evolution in autocatalytic chemical reaction systems. *Life*, 11(4):308, 2021.

[37] Yu Liu. On the definition of a self-sustaining chemical reaction system and its role in heredity. *Biology Direct*, 15(1), 2020.

[38] Norman Packard, Mark A. Bedau, Alastair Channon, Takashi Ikegami, Steen Rasmussen, Kenneth O. Stanley, and Tim Taylor. An overview of open-ended evolution: Editorial introduction to the open-ended evolution ii special issue. *Artificial Life*, 25(2):93–103, 2019.

[39] Alastair Channon, Mark A. Bedau, Norman H. Packard, and Tim Taylor. Editorial introduction to the 2024 special issue on open-ended evolution. *Artificial Life*, 30(3):300–301, 2024.

[40] Norbert Wiener. *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press, 1948.

[41] W. Ross Ashby. *An Introduction to Cybernetics*. Chapman and Hall, 1956.

[42] Roger C. Conant and W. Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, 1970.

[43] Francisco J. Varela. *Principles of Biological Autonomy: A New Annotated Edition*. MIT Press, 2025. Edited and annotated by Ezequiel A. Di Paolo and Evan Thompson.

[44] Francesco Bianchini. Autopoiesis of the artificial: From systems to cognition. *BioSystems*, 230:104936, 2023.

[45] Sergio Rubin. Cartography of the multiple formal systems of molecular autopoiesis: From the biology of cognition and enaction to anticipation and active inference. *BioSystems*, 230:104955, 2023.

[46] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11:127–138, 2010.

[47] Xabier E. Barandiaran, Ezequiel Di Paolo, and Marieke Rohde. Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5):367–386, 2009.

[48] Kingson Man and Antonio Damasio. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1:446–452, 2019.

[49] Nils Bertschinger, Eckehard Olbrich, Nihat Ay, and Jürgen Jost. Autonomy: An information theoretic perspective. *BioSystems*, 91(2):331–345, 2008.

[50] Anil K. Seth. Measuring autonomy and emergence via granger causality. *Artificial Life*, 16(2):179–196, 2010.

[51] David Krakauer, Nils Bertschinger, Eckehard Olbrich, Jessica C. Flack, and Nihat Ay. The information theory of individuality. *Theory in Biosciences*, 139:209–223, 2020.

[52] Charles Ofria and Claus O. Wilke. Avida: A software platform for research in computational evolutionary biology. *Artificial Life*, 10(2):191–229, 2004.

[53] Miguel A. Fortuna, Luis Zaman, Charles Ofria, and Andreas Wagner. The genotype-phenotype map of an evolving digital organism. *PLOS Computational Biology*, 13(2):e1005414, 2017.

[54] Luc Steels. The artificial life roots of artificial intelligence. *Artificial Life*, 1(1–2):75–110, 1993.

[55] Stefan Helmreich. *Silicon Second Nature: Culturing Artificial Life in a Digital World*. University of California Press, 1998.

[56] John Johnston. *The Allure of Machinic Life: Cybernetics, Artificial Life, and the New AI*. MIT Press, 2008.

[57] Susan Stepney. Towards origins of virtual artificial life: An overview. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380(1936), 2025.

[58] Thomas S. Ray. An approach to the synthesis of life. In *Artificial Life II*, pages 371–408. Addison-Wesley, 1991.

[59] Hiroki Sayama and Chrystopher L. Nehaniv. Self-reproduction and evolution in cellular automata: 25 years after evoloops. *Artificial Life*, 31(1):81–95, 2025.

[60] Akarsh Kumar, Chris Lu, Louis Kirsch, Yujin Tang, Kenneth O. Stanley, Phillip Isola, and David Ha. Automating the search for artificial life with foundation models. *Artificial Life*, 31(3):368–396, 2025.

[61] Fernando Rodriguez and Phil Husbands. A saucerful of secrets: Open-ended organizational closure in the game of life. In *The 2024 Conference on Artificial Life*, 2024.

[62] Rodney A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47(1–3):139–159, 1991.

[63] Rolf Pfeifer and Josh Bongard. *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press, 2006.

[64] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[65] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17:391–444, 2007.

[66] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[67] Lexin Zhou et al. General scales unlock ai evaluation with explanatory and predictive power. *Nature*, 652:58–67, 2026.

[68] Alberto Hernández-Espinosa, Luan Ozelim, Felipe S. Abrahão, and Hector Zenil. Superarc: A test for artificial superintelligence based on compressed modelling, recursive prediction and problem complexity. *Nature Communications*, 17, 2026.

[69] John Maynard Smith and Eőrs Szathmáry. *The Major Transitions in Evolution*. Oxford University Press, 1995.

[70] Ricard Solé. The major synthetic evolutionary transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1701):20160175, 2016.

[71] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, et al. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

[72] Olaf Witkowski and Eric Schwitzgebel. The ethics of life as it could be: Do we have moral obligations to artificial life? *Artificial Life*, 30(2):193–215, 2024.

[73] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[74] Jiaqi Wang, Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, et al. Large language models for robotics: Opportunities, challenges, and perspectives. *arXiv preprint arXiv:2401.04334*, 2024.

[75] Shuo Ren, Can Xie, Pu Jian, Zhenjiang Ren, Chunlin Leng, and Jiajun Zhang. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*, 2025.

[76] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, 2025.

[77] Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*, 13:162467–162504, 2025.

[78] Dingzhe Li, Yixiang Jin, Hongze Yu, Jun Shi, Xiaoshuai Hao, Peng Hao, Huaping Liu, Fuchun Sun, Bin Fang, et al. What foundation models can bring for robot learning in manipulation: A survey. *arXiv preprint arXiv:2404.18201*, 2024.

[79] Lars Kunze, Nick Hawes, Tom Duckett, Marc Hanheide, and Tomas Krajnik. Artificial intelligence for long-term robot autonomy: A survey. *IEEE Robotics and Automation Letters*, 3(4):4023–4030, 2018.

[80] Nitesh Kumar, Jaekyung Jackie Lee, Sivakumar Rathinam, Swaroop Darbha, P. B. Sujit, and Rajiv Raman. The persistent robot charging problem for long-duration autonomy. *arXiv preprint arXiv:2409.00572*, 2024.

[81] Christian Brommer, Danylo Malyuta, Daniel Hentzen, and Roland Brockers. Long-duration autonomy for small rotorcraft uas including recharging. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 7252–7258, 2018.

[82] Alessandro Saviolo, Jeffrey Mao, Roshan Balu T. M. B., Vivek Radhakrishnan, and Giuseppe Loianno. Autocharge: Autonomous charging for perpetual quadrotor missions. In *2023 IEEE International Conference on Robotics and Automation*, pages 5400–5406, 2023.

[83] Michael Wehner, Ryan L. Truby, Daniel J. Fitzgerald, et al. An integrated design and fabrication strategy for entirely soft, autonomous robots. *Nature*, 536:451–455, 2016.

[84] John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, volume 37, pages 50528–50652, 2024.

[85] Xuemei Fu et al. Self-healing actuatable electroluminescent fibres. *Nature Communications*, 15, 2024.

[86] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024.

[87] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, pages 52040–52094, 2024.

[88] Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*, 2023.

[89] Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv preprint arXiv:2401.16158*, 2024.

[90] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Kanishka Rao, Pierre Sermanet, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, and Mengyuan Yan. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[91] Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems XX*, 2024.

[92] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: A diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

[93] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems XXI*, 2025.

[94] Gemini Robotics Team. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[95] NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[96] Xinghang Li et al. What matters in building vision–language–action models for generalist robots. *Nature Machine Intelligence*, 8:158–172, 2026.

[97] Ruaridh Mon-Williams, Gen Li, Ran Long, Wenqian Du, and Christopher G. Lucas. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence*, 7:592–601, 2025.

[98] Christopher E. Mower et al. A robot operating system framework for using large language models in embodied ai. *Nature Machine Intelligence*, 8:313–325, 2026.

[99] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024.

[100] Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. Towards lifelong learning of large language models: A survey. *arXiv preprint arXiv:2406.06391*, 2024.

[101] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[102] Gabriel Sarch, Yue Wu, Michael J. Tarr, and Katerina Fragkiadaki. Open-ended instructable embodied agents with memory-augmented large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3468–3500, 2023.

[103] Junhao Zheng, Xidi Cai, Qiuke Li, Duzhen Zhang, ZhongZhi Li, Yingying Zhang, Le Song, and Qianli Ma. Lifelongagentbench: Evaluating llm agents as lifelong learners. *arXiv preprint arXiv:2505.11942*, 2025.

[104] Yuxuan Cai, Yipeng Hao, Jie Zhou, Hang Yan, Zhikai Lei, Rui Zhen, Zhenhua Han, Yutao Yang, Junsong Li, Qianjun Pan, Tianyu Huai, Qin Chen, Xin Li, Kai Chen, Bo Zhang, Xipeng Qiu, and Liang He. Building self-evolving agents via experience-driven lifelong learning: A framework and benchmark. *arXiv preprint arXiv:2508.19005*, 2025.

[105] Victor Zykov, Efstathios Mytilinaios, Bryant Adams, and Hod Lipson. Self-reproducing machines. *Nature*, 435:163–164, 2005.

[106] Saul Griffith, Dan Goldwater, and Joseph M. Jacobson. Self-replication from random parts. *Nature*, 437:636, 2005.

[107] Sam Kriegman, Douglas Blackiston, Michael Levin, and Josh Bongard. Kinematic self-replication in reconfigurable organisms. *Proceedings of the National Academy of Sciences*, 118(49), 2021.

[108] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed H. Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

[109] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Zhang, Ceyao Wang, Zili Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. Metagpt: Meta programming for a multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

[110] Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X. Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, Antoine Laurens, Claudio Fantacci, Valentin Dalibard, Martina Zambelli, Murilo Martins, Rugile Pevceviciute, Michiel Blokzijl, Misha Denil, Nathan Batchelor, Thomas Lampe, Emilio Parisotto, Konrad Zolna, Scott Reed, Sergio Gomez Colmenarejo, Jon Scholz, Abbas Abdolmaleki, Oliver Groth, Jean-Baptiste Regli, Oleg Sushkov, Tom Rothorl, Jose Enrique Chen, Yusuf Aytar, Dave Barker, Joy Ortiz, Martin Riedmiller, Jost Tobias Springenberg, Raia Hadsell, Francesco Nori, and Nicolas Heess. Robocat: A self-improving foundation agent for robotic manipulation. *Transactions on Machine Learning Research*, 2023.

[111] Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

[112] Yuqi Zhu, Shuofei Qiao, Yixin Ou, Shumin Deng, Shiwei Lyu, Yue Shen, Lei Liang, Jinjie Gu, Huajun Chen, and Ningyu Zhang. Knowagent: Knowledge-augmented planning for llm-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3709–3732, 2025.

[113] Rauno Arike, Elizabeth Donoway, Henning Bartsch, and Marius Hobbhahn. Evaluating goal drift in language model agents. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):192–203, 2025.

[114] Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring ai ability to complete long software tasks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[115] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.

[116] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, 2024.

[117] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

[118] Kunlun Zhu, Hongyi Du, Zhaochen Hong, et al. Multiagentbench: Evaluating the collaboration and competition of llm agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 8580–8622, 2025.

[119] Aron Vallinder and Edward Hughes. Cultural evolution of cooperation among llm agents. *arXiv preprint arXiv:2412.10270*, 2025.

[120] Avishkar Bhoopchand, Bethanie Brownfield, Adrian Collister, Agustin Dal Lago, Ashley Edwards, Richard Everett, Alexandre Fréchette, Yanko Gitahy Oliveira, Edward Hughes, Kory W. Mathewson, Piermaria Mendolicchio, Julia Pawar, Miruna Pîslar, Alex Platonov, Evan Senter, Sukhdeep Singh, Alexander Zacherl, and Lei M. Zhang. Learning few-shot imitation as cultural transmission. *Nature Communications*, 14, 2023.

[121] Ali E. Ghareeb et al. A multi-agent system for automating scientific discovery. *Nature*, 655:497–505, 2026.

[122] Juraj Gottweis et al. Accelerating scientific discovery with co-scientist. *Nature*, 655:487–496, 2026.

[123] Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C. Landsness, Daniel L. Barabasi, Siddharth Narayanan, Nicky Evans, Shriya Reddy, Martha Foiani, Aizad Kamal, Leah P. Shriver, Fang Cao, Asmamaw T. Wassie, Jon M. Laurent, Edwin Melville-Green, Mayk Caldas, Albert Bou, Kaleigh F. Roberts, Sladjana Zagorac, Timothy C. Orr, Miranda E. Orr, Kevin J. Zwezdaryk, Ali E. Ghareeb, Laurie McCoy, Bruna Gomes, Euan A. Ashley, Karen E. Duff, Tonio Buonassisi, Tom Rainforth, Randall J. Bateman, Michael Skarlinski, Samuel G. Rodriques, Michaela M. Hinks, and Andrew D. White. Kosmos: An ai scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025.

[124] Humza Nusrat and Omar Nusrat. When ai does science: Evaluating the autonomous ai scientist kosmos in radiation biology. *arXiv preprint arXiv:2511.13825*, 2025.

[125] Emmy Brown and Sean T. Vittadello. Studying organisational closure in biological systems with process-enablement graphs. *BioSystems*, 257:105567, 2025.

[126] Juan Segura. Operationalizing autopoiesis beyond organisms via repair-closure. *SSRN Electronic Journal*, 2026.

[127] Eric W. Bonabeau and Guy Theraulaz. Why do we need artificial life? *Artificial Life*, 1(3):303–325, 1994.

[128] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

[129] Wangchunshu Zhou, Yuchen Eleanor Jiang, Long Li, Jialong Wu, Tiannan Wang, Shi Qiu, Jintian Zhang, Jing Chen, Ruipu Wu, Shuai Wang, Shiding Zhu, Jiyu Chen, Wentao Zhang, Xiangru Tang, Ningyu Zhang, Huajun Chen, Peng Cui, and Mrinmaya Sachan. Agents: An open-source framework for autonomous language agents. *arXiv preprint arXiv:2309.07870*, 2023.

[130] Johannes Jaeger. Artificial intelligence is algorithmic mimicry: Why artificial "agents" are not (and won't be) proper agents. *Neurons, Behavior, Data Analysis, and Theory*, 2024.

[131] Richard Fang, Rohan Bindu, Akul Gupta, and Daniel Kang. Llm agents can autonomously exploit one-day vulnerabilities. *arXiv preprint arXiv:2404.08144*, 2024.

[132] Yuxuan Zhu, Antony Kellermann, Akul Gupta, Philip Li, Richard Fang, Rohan Bindu, and Daniel Kang. Teams of llm agents can exploit zero-day vulnerabilities. *arXiv preprint arXiv:2406.01637*, 2024.

[133] Ziyi Yang, Shreyas S. Raman, Ankit Shah, and Stefanie Tellex. Plug in the safety chip: Enforcing constraints for llm-driven robot agents. In *2024 IEEE International Conference on Robotics and Automation*, pages 14435–14442, 2024.

[134] Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, et al. Risks of ai scientists: Prioritizing safeguarding over autonomy. *arXiv preprint arXiv:2402.04247*, 2024.

[135] National Institute of Standards and Technology. Accelerating the adoption of software and artificial intelligence agent identity and authorization. Technical report, NIST National Cybersecurity Center of Excellence, 2026.

[136] European Parliament and Council. Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, 2024.

[137] International AI Safety Report. International ai safety report 2026. Technical report, International AI Safety Report, 2026.

[138] National Institute of Standards and Technology. Ai agent standards initiative. Technical report, NIST, 2026.