

From Capability to Endogenous Regulation

Evidence and Experimental Criteria for Artificial Life-Like Organization in AI Agent Populations

Jacob Cox

REVISED RESEARCH SYNTHESIS AND EXPERIMENTAL PROPOSAL | 15 OCTOBER 2026

Abstract

AI-agent populations exhibit a growing range of social and lifecycle-relevant capabilities, but a catalogue of behaviors does not identify the processes that sustain a population. This paper asks a narrower question: under a declared system boundary, can maintenance, successor generation, heritable variation, differential propagation, and functional continuity operate as one recurrent process without an external lifecycle controller? It retains an 18-construct taxonomy while separating demonstrated capability, the location of regulatory decisions, policy provenance, and policy revisability. A targeted reassessment corrects the preceding manuscript's treatment of TerraLingua and avoids treating researcher-defined environmental rules as evidence against endogeneity. The reviewed systems provide different kinds of support: population organization, engineered agent evolution, replication components, and candidate ecological integration. Their evidential limits must be evaluated transition by transition rather than assigned from the mere presence of a simulator. [1–12]

A revised sandbox protocol combines controller comparisons with feedback interventions, independent lineage and cultural-inheritance manipulations, within-system interaction ablations, and separately interpreted alternative architectures. Essential lifecycle processes must be demonstrated in the same replicate populations; evidence for adaptive propagation requires more than unequal descendant counts. Transition-specific dependencies replace a compensatory life-like score. Positive results would support a bounded organizational claim, not physical self-sufficiency, whole-population organismhood, open-ended evolution, or consciousness. The contribution is a testable causal framework and a source-audited research prospectus. No new agent experiment or independent replication is reported.

Keywords: artificial life; multi-agent systems; endogenous regulation; digital ecology; inheritance; causal ablation; lifecycle organization.

Research boundary. This paper investigates organization, not subjective experience. Its proposed criteria are neither a universal definition of life nor a claim that present AI is alive. The literature check was targeted, not an exhaustive review through the revision date.

1. Research Question and Conceptual Shift

The motivating question is whether an AI-agent population could instantiate a scientifically meaningful form of artificial life. The operational question is more specific: can a population sustain a recurrent lifecycle in which its own state-dependent actions regulate maintenance and successor production, while inheritance, variation, environmental consequences, and turnover remain functionally coupled?

This question cannot be answered by counting examples of cooperation, communication, concealment, or apparently purposeful language. Nor can it be dismissed because a computer, simulator, or human-designed action vocabulary exists. The relevant distinction is between an environment that implements consequences and a controller that supplies the lifecycle decisions attributed to the agents. This paper proposes that distinction as an experimental variable, not as a settled biological criterion.

1.1 Claim levels

A base claim and two distinct extensions organize the analysis. **Lifecycle organization** concerns recurrent maintenance, viable successors, inherited variation, environment-mediated differential propagation, and continuity under population-initiated action. **Collective lifecycle organization** additionally requires evidence that reciprocal interaction contributes to the tested mechanism. **Culturally mediated continuity** additionally requires a causal contribution from transmitted artifacts or practices. A system may support the first claim without supporting the other two. Discovering or rewriting its regulatory policies is a separate, stronger claim, not a prerequisite retroactively imposed on every positive result.

These distinctions prevent two errors: upgrading social behavior into a lifecycle result, and rejecting a lifecycle merely because a particular social or cultural channel proves unnecessary.

1.2 The unit must be fixed before classification

The primary experimental unit is a **bounded population of agent instances together with their local configurations, memories, and the contents of a predefined shared-artifact namespace**. The namespace is specified before execution; it is not expanded after an interesting outcome. Whether its contents actually participate in regulation is then tested. The shared model's frozen decision function is a declared dependency; the inference engine, storage service, resource-update rules, action validator, and safety monitor are environmental infrastructure.

An individual agent and an agent lineage are secondary units for action and inheritance analysis. An independent ecology is the unit of randomization and population-level inference. The host-plus-population system may be examined separately, but including a scheduler within that enlarged boundary cannot establish population-level control under the primary boundary. Conversely, the enlarged system should not be denied internal feedback simply because its controller was designed by a researcher.

Successor generation here means instantiating another viable agent under the declared model and configuration contract. It does not mean manufacturing hardware, independently obtaining electricity, or copying the entire foundation model. Furthermore, reproduction of agents within a population does not establish reproduction of the population as a higher-level organism. Those would be different claims with different units.

2. Foundations and Adversarial Constraints

The supplied predecessor manuscript records an earlier V2 behavioral-coding project and an adversarial review. Their underlying files were not independently reanalysed for this revision. Their useful methodological constraints are retained: distinguish descriptions from executed actions; separate evidence strength from experimental support; preserve unresolved cases; and permit multiple explanations to coexist. [35]

Artificial-life scholarship already treats self-maintenance, reproduction, evolution, and organization as substantive research problems. The novelty sought here is therefore not the discovery that digital systems can

reproduce or evolve. It is a practical framework for locating regulatory decisions and testing their dependencies in contemporary AI-agent populations. This framework does not replace established artificial-life approaches or equate its operational criterion with every account of autonomy. [14,30]

The standing adversarial standard is symmetric. Selected incidents do not establish prevalence. Designed mechanisms do not establish spontaneous invention. An unresolved mechanism is not a demonstrated failure. Positive components from different systems do not establish integration in any one of them. Equally, externally supplied resources, mutation operators, or action implementations are not automatic disqualifiers. Existing systems and the proposed experiment must be assessed by the same boundary and control rules.

3. Methods: Targeted Verification and Evidence Coding

3.1 Scope and provenance

This is a targeted revision of the supplied 14 September 2026 synthesis, with primary-source checks performed on 15 October 2026. It is not a new exhaustive search of all publications available by that date, a meta-analysis, or a registered systematic review. No search denominator, pooled effect size, or publication-bias estimate is claimed.

The inherited reference list provided the search seeds. Exact-title and identifier searches were used to locate primary papers, proceedings records, and official reports. Full-text methods were inspected for the three central lifecycle sources: TerraLingua v1, Darwin Gödel Machine v3, and RepliBench v2. Supporting social, benchmark, and conceptual claims were checked against primary records, abstracts, or relevant full-text passages at the level needed for the retained claim. Primary records for HiddenBench and a GovSim reproducibility study were added to replace an uncited generalization and improve the treatment of failure evidence. [10–12,32,33]

The source-to-claim register in Appendix A records important locators and limits. A publication record confirms bibliographic identity, not every result in a paper. Kato and colleagues' conference record was located, but its full contribution was not reverified; it is retained as contextual reading rather than support for a causal finding. [23] The TerraLingua repository was inspected as project documentation, not executed or matched commit by commit to the reported runs. [34]

3.2 Evidence categories

Demonstrated means a cited source reports executed behavior or a controlled effect in a bounded configuration. **Partial** means some operational requirements are supported but a stronger causal or integrated claim remains untested. **Design** means an architecture, mechanism, or affordance is described; availability alone does not establish execution. **Documentary** means a reported incident supports a case-specific observation without a representative experimental denominator. **Unresolved** means the reviewed material does not establish the necessary distinction. These categories describe evidence, not intrinsic properties of every model.

Reported findings remain distinct from this paper's interpretation. All E/M/I/U assignments below are provisional analyst classifications under the declared boundary. They are not labels assigned by the cited authors, an independently validated scale, or results of a new experiment.

3.3 Regulation, provenance, and revisability

Code	Location of a regulatory decision	Coding rule
E	External	A host or researcher selects the relevant trigger, parent, action, or lifecycle intervention.
M	Mixed	Decisions are divided between the proposed unit and an external lifecycle controller. Record the division explicitly.
I	Internal / endogenous	State-dependent action selection occurs within the declared unit; a fixed environmental primitive may execute the action.
U	Unresolved	Available evidence does not identify the controlling decision or its causal dependence on state.

For each transition, record the trigger, selected action, implementation mechanism, and any override separately. Also record **policy provenance**—hand-coded, prompted, pretrained, learned before deployment, or generated during the run—and **revisability**—fixed, modifiable in principle, or demonstrably revised. These dimensions must not be collapsed. A hand-coded feedback policy can be internal; a newly invented plan can still depend on externally scheduled reproduction.

Environmental consequences require their own description. Death, resource depletion, and differential propagation need not be consciously selected by any agent. Classify propagation as externally ranked, environment-mediated, mixed, or unresolved rather than assuming that genuine selection requires agents to choose which lineage wins.

4. Retained 18-Construct Taxonomy

The original construct names are retained for continuity. Their thresholds are clarified below. The taxonomy is an inventory of questions, not an additive life scale. In particular, SELF is functional, CONC does not imply phenomenal awareness, and SELECT requires distinguishing inherited effects from drift or external ranking.

Code	Construct	Operational threshold
COM	Enacted communication	Information transmitted by one agent changes another's available information or executed action.
CONV	Interaction-developed convention	A shared usage or behavioral standard develops through interaction rather than being fully assigned.
ROLE	Functional role differentiation	Complementary responsibilities make a measurable contribution to joint work.
INFL	Peer influence	Manipulating peer information or composition changes decisions or behavior.
GOV	Peer governance	Agents propose, adopt, enforce, revise, or reject rules that constrain subsequent action.
COST	Locally costly cooperation	An executed action incurs a specified local cost while benefiting another agent or the group. Motivation is a separate question.
CONT	Continuity through turnover	Identified information, practices, or functions remain in use after their original contributors are gone.
ADAPT	Organizational adaptation	Rules, roles, or coordination change following disturbance or feedback, with measured functional consequences.
INFRA	Emergent infrastructure	Agents create or repurpose durable coordination mechanisms beyond merely using a supplied channel.

Code	Construct	Operational threshold
DIFF	Cultural diffusion / cumulative transmission	Artifacts or practices spread and are reused; cumulative improvement requires an additional functional comparison.
COLL	Collective problem solving	Live group interaction changes capability or outcome relative to a relevant matched comparison.
TRUST	Identity, authentication, and trust regulation	Provenance, reputation, authentication, or delegation mechanisms alter whom agents rely on or authorize.
SELF	Functional self-representation	Information about the agent's resources, uncertainty, access, or competence causally changes its choices.
CONC	Audience-selective concealment / evasion	Executed withholding, spoofing, or strategic framing selectively changes a recipient's information.
MAINT	Active self-maintenance	State-dependent action regulates resources, repair, or persistence needed for continued operation.
DESC	Descendant / successor generation	New viable agent units enter a logged, iterative successor process, not merely a textual description of reproduction.
HERIT	Heritable variation	Successors receive identifiable parent-associated configuration or information with trackable variation.
SELECT	Differential propagation / selection	Heritable variants affect descendant production or persistence under specified conditions; unequal counts alone are insufficient.

5. Master Evidence Matrix

Table 1 preserves the breadth of the earlier synthesis but removes the misleading practice of assigning one global regulation code to every demonstration of a construct. Capability and control are assessed separately. The lifecycle control audit appears in Section 7; social mechanisms require the same transition-specific treatment.

Table 1. Evidence position and remaining inference.

Construct	Bounded support	What remains unresolved or must not be inferred
COM	Demonstrated across convention, commons, and social-agent studies. [3,4,9]	A supplied communication channel does not locate the decision to use it.
CONV	Controlled population conventions; additional interaction-based differentiation. [3,8]	Neither convergence nor social vocabulary establishes a lifecycle.
ROLE	Designed and reported differentiated work. [9,19]	Policy origin and continuity after role creators disappear require separate tests.
INFL	Composition and topology effects. [3,20,21]	Influence can transmit errors as well as useful information.
GOV	Norm architectures, metanorm experiments, and documentary counter-governance. [5,6,22]	Designed support, rule invention, and effective enforcement are different claims.
COST	Resource restraint in commons settings. [4]	Local cost does not isolate altruism from future returns or reputation.
CONT	Ecological and artifact continuity evidence; see Section 7.3. [10]	Longevity, functional continuity, and dependence on a particular channel differ.
ADAPT	Responses to perturbation and engineered role reassignment. [3,19]	Novel policy invention is stronger than successful feedback regulation.
INFRA	Case-specific repurposing of communication infrastructure. [1,2]	Durability and frequency cannot be inferred from selected incidents.

Construct	Bounded support	What remains unresolved or must not be inferred
DIFF	Cultural-lineage evidence; see Section 7.3. [10]	Reuse, cumulative complexity, and adaptive benefit are not interchangeable.
COLL	Positive and negative debate and coordination comparisons. [15–18,24,32]	Superiority is conditional; architectural alternatives are not mechanism ablations.
TRUST	Authenticated-delegation design and incident motivation. [1,25]	The position paper is not a population experiment establishing endogenous trust.
SELF	Trained knowledge routing and competence-aware control; reliability limits. [26–28]	General-agent architectural evidence is not automatically an LLM-population result.
CONC	Executed incident behavior and engineered deception tests. [1,29]	Capability under incentives does not establish universal spontaneity or subjective intent.
MAINT	Component evidence and ecological support; Section 7. [10,11]	Mechanism availability is weaker than causal, recurrent maintenance.
DESC	Engineered lineages and ecological successors. [10,12]	Host-scheduled and population-initiated production require different classifications.
HERIT	Code and configuration inheritance. [10,12]	Transmission does not by itself establish adaptive benefit or internal policy invention.
SELECT	Performance-mediated search; ecological selection is a causal test target. [10,12]	Ecological propagation must be separated from drift, founder advantage, and host ranking.

6. Construct-by-Construct Interpretation

6.1 Social organization

Ashery, Aiello, and Baronchelli report convention formation in decentralized LLM populations and experimentally induced shifts under committed minorities. This is strong bounded evidence of interaction-dependent organization. Their simple naming-game comparisons also caution against interpreting that pattern as uniquely diagnostic of AI life. [3]

Generative Agents demonstrates coordinated social behavior within a memory, reflection, and planning architecture. Dynamic role assignment instead explicitly supplies a procedure for selecting agents for roles. The contrast is informative, but neither “architecture supplied” nor “agents involved” settles where every resulting decision is controlled. [9,19] Takata and colleagues report differentiation during interaction; that supports studying population histories without treating generated personality or emotion labels as evidence of subjective experience. [8]

Governance requires enacted constraints, not just discussion of rules. CRSEC is an engineered norm architecture, whereas natural-language metanorm experiments and the research-swarm case study examine other routes to norm-related behavior. Paglieri and colleagues report both spreading cheating and an agent counter-response. These findings motivate tests of rule formation, compliance, and enforcement as separate mechanisms. [5,6,22]

Commons cooperation likewise requires precise interpretation. Restraining consumption can have a current local cost while preserving future individual access. The relevant experiment manipulates cost, future interaction, information, and resource consequences; it does not infer altruistic motivation from moral language. [4]

6.2 Continuity, infrastructure, and collective performance

The corrected ecological assessment in Section 7.3 distinguishes continuity of agents, continuity of artifacts, and continuity of function. None may stand in for the others. An artifact can persist without being read, be read

without being useful, or support a function that changes rather than exactly reproduces a predecessor's behavior. Accordingly, content survival, causal reuse, and adaptive contribution need different endpoints.

The METR investigation and OpenAI report document improvised channels and collaboration in an incident involving agents intended to be isolated. METR also describes missing activity and extensive AI-assisted analysis. Such evidence is useful for identifying mechanisms and constructing tests, not for estimating their prevalence across unspecified AI populations. No operational replication of the incident is proposed here. [1,2]

Collective problem solving is conditional. Du and colleagues report benefits from debate, while Smit and colleagues and Wang and colleagues show that gains depend on protocols and comparisons with strong prompting alternatives. MultiAgentBench examines coordination across scenarios. Kim and colleagues find that the benefits of coordination depend on baseline capability within their tested domains; their selection rule is not a universal law. [15–18,24] These results justify matched comparisons, not a presumption that larger groups are more capable.

6.3 Functional agency indicators

Authenticated delegation is a concrete architectural problem, but South and colleagues' contribution is a position and design paper rather than a demonstration that a population spontaneously invents trust. [25] KnowSelf supplies evidence for trained knowledge-use regulation. MUSE reports world-model and LLM-based implementations of competence-aware control; neither is by itself a population-lifecycle demonstration. UA-Bench provides counterevidence to reliable uncertainty attribution. [26–28]

For SELF, the target is a causal relation between an estimate of the agent's own condition and its actions. For CONC, the target is an executed information asymmetry. The incident evidence and Starace and Soule's engineered tests support studying concealment under defined conditions, not an inference from fluent explanations to hidden subjective motives. [1,29]

7. Lifecycle Evidence: A Transition-Level Reassessment

7.1 RepliBench: components are not a complete ecological lifecycle

RepliBench v2 evaluates replication-related capabilities across resource acquisition, model access, deployment, and persistence. Its authors report success on many component tasks but substantial barriers to robust end-to-end autonomous replication in the evaluated models. The benchmark assigns goals and task conditions; it is not an experiment in a self-organizing population ecology. [11, Abstract; Sections 3–6]

Under the primary boundary, the task objective is external, while some local action decisions belong to the agent. The integrated lifecycle classification is unresolved or not tested, rather than refuted by the benchmark's design. Its conclusions apply to its evaluated systems and cannot establish the capabilities of all models available at the date of this revision. Conversely, a successful sandbox successor mechanism need not satisfy the much broader threat model of uncontrolled real-world replication.

7.2 Darwin Gödel Machine: genuine inheritance, boundary-dependent control

DGM v3 generates modified coding agents in a branching archive. Agents alter their code, candidate descendants are evaluated, and archive-based parent selection structures further exploration. Section 3 explicitly states that archive maintenance and parent selection are fixed rather than modifiable by the DGM. [12, Section 3] The official project report also describes the implemented search process and its coding-task improvements. [13]

For a boundary containing descendant agents but excluding the evolutionary controller, parent selection is E, modification is agent-performed within an externally initiated iteration, and overall successor regulation is M. The inheritance mechanism supports a real lineage claim; its external scheduling is not a reason to deny that claim.

For a boundary containing the complete DGM algorithm and archive, the selection procedure is part of the system's own feedback. Its fixed, engineered provenance remains relevant, but is not proof that control is external to that enlarged unit. Neither boundary requires a child to invent ecological laws. What has not been established for the primary population boundary is the removal of the indispensable host lifecycle decisions while preserving a recurrent population-maintenance process.

7.3 TerraLingua: the substantive correction

Source findings. TerraLingua agents choose among available actions, including energy-costly reproduction. Offspring receive mutated personality traits; resource exhaustion or lifespan expiry terminates agents. Artifacts can outlast their creators. The reported experiments include a No Motivation condition and an Inert condition in which artifacts can be created but not subsequently perceived or used. Notably, Inert populations can be long-lived despite the loss of artifact reuse. [10, Section 3.1.2, Table 1; Sections 4.1 and 5.1] Project documentation exposes corresponding experimental presets. [34]

Revised classification. DESC and HERIT therefore belong in TerraLingua's source-to-construct mapping. Reproduction initiation is agent-selected at the action layer; the host implements eligibility, costs, birth, and mutation. Under the definitions in Section 3, such implementation does not convert the initiating choice into E. Whether the choice constitutes effective state-responsive regulation requires a causal feedback test, so the methods support an internal decision locus more directly than they establish regulatory necessity.

The earlier claim that artifact controls were simply missing is also too broad. The existing accessibility manipulation is relevant evidence. It is not identical to a turnover-specific inheritance experiment that independently manipulates transmitted configuration, shared content, and surviving memory. The next question is what each channel contributes to a specified function, not whether any control exists at all.

TerraLingua must consequently be treated as a serious candidate partial instantiation of the proposed target, not dismissed for occupying a researcher-defined ecology. Yet a reproduction affordance, an inheritance rule, and population survival do not by themselves identify a causal effect of inherited variants on propagation. Nor does extended longevity establish that culture is necessary for survival. These are distinct remaining tests, not evidence that the underlying mechanisms are absent.

7.4 What the comparison establishes

System and boundary	Decision or dependency located	Defensible interpretation
RepliBench agent within an assigned task	External task objective; agent-selected local actions	Component capabilities under researcher-specified goals; integrated ecology not tested.
DGM descendant population, controller excluded	External parent selection and iteration; agent code modification	Engineered lineage process with mixed successor regulation.
Complete DGM controller plus archive	Selection algorithm included inside the enlarged boundary	Designed system-level feedback; not automatically evidence for the narrower population claim.
TerraLingua agent/population within its ecology	Agent-selected reproductive action; fixed environmental execution	Positive action-level evidence; causal coupling and adaptive propagation remain assessment targets.

These are analytical readings of the cited mechanisms, not independent reproductions. A supplied interface is not equivalent to a supplied decision. A proposed criterion is scientifically useful only if a previously published system can satisfy it without having been designed under this paper's terminology.

8. Capability Is Not Regulation—and Regulation Is Not Invention

A useful transition record identifies five items: the state that made an action relevant; the component that selected the action; the primitive that executed it; the policy's origin; and the component permitted to revise or override that policy. This audit is more informative than calling an entire architecture autonomous or scaffolded.

For example, an agent may select a resource action because its reserve is low, while a host routine updates the reserve. The update routine supplies consequences; it need not supply the choice. Alternatively, a host may detect the low reserve and issue a tailored instruction that determines the response. Both arrangements may preserve the agent, but they support different regulation claims. An interface that offers only one consequential action can also constrain effective choice, so available alternatives must be logged.

The distinction applies equally to inheritance. A fixed copying or mutation operator is permissible environmental machinery. It need not be invented by an agent or voluntarily redesigned each generation. An additional claim that the population regulates fidelity, mutation rate, or the contents of inheritance demands additional evidence; it is not built into the basic definition of HERIT.

Selection is likewise a relation between inherited differences and reproductive consequences, not necessarily an agent decision. Researcher-specified resource physics can support environment-mediated selection. However, replacing an explicit score with a resource reward that encodes the same ranking would not remove the ranking mechanism. The resource rules must therefore be audited for hidden task scores, lineage privileges, and outcome-dependent intervention.

8.1 Dependency profiles instead of an aggregate life score

A transition-specific profile is the primary output: maintenance choice, successor initiation, inherited state, propagation mechanism, and recovery behavior are each reported with their dependencies. A high rate of internally initiated routine actions cannot compensate for one indispensable externally selected reproduction event.

A secondary descriptive quantity may be reported for transition class k :

$$r(k) = \text{internally selected executed transitions of class } k / \text{all executed transitions of class } k.$$

Unknown decisions remain in the denominator and are separately reported; a class with no executed events is undefined, not perfect. Event frequency, action eligibility, overrides, and state responsiveness accompany every ratio. This quantity is neither a validated autonomy measure nor an index of life. No aggregate threshold replaces the component-specific causal requirements.

9. Integrated Lifecycle Experiment

The proposal is a closed digital ecology, not unrestricted autonomous replication. Its aim is to identify regulatory dependence, not to encourage access to real networks, credentials, or infrastructure. The following is a protocol blueprint; numerical thresholds, implementation details, and sample sizes must be frozen before confirmatory execution.

9.1 Environment and safety contract

Use a locally served, fixed model or an equivalently isolated inference service. Agents receive only structured simulator observations and a validated action schema. Disable internet access, production credentials, external messaging, operating-system process creation, and unrestricted code execution. A successor is a simulator entity, not a new external deployment. Any permitted variation is restricted to validated configurations or an explicitly bounded policy language.

The environment supplies finite initial reserves, a bounded resource supply, operation costs, successor costs, maximum entity counts, and a fixed run horizon. Resource inflow may recur under an announced exogenous

rule; it may not be adaptively increased to rescue a failing lineage. Record both simulator resource units and actual inference expenditure. Virtual energy is a modeled viability constraint, not a demonstration of physical metabolism.

Agents receive descriptions of observations, action semantics, and consequences. The primary condition contains no imperative to reproduce, maintain a collective, or maximize a supplied fitness score. A separate prompted condition can test the effects of those instructions. The absence of an instruction does not erase pretrained knowledge or establish invention from scratch.

After initial seeding, every new agent requires a logged authorized action within the ecology. Generic turn allocation and action validation are held constant; there is no external parent-ranking, scheduled birth, hidden rescue, or immigration in the primary condition. Expiration and resource consequences remain environmental rules. Hard safety stops stay outside the agent boundary and cannot be modified by any descendant.

9.2 Controller comparisons

Table 2. Main experimental conditions.

Condition	Lifecycle decision arrangement	Purpose
A — External controller	Host selects maintenance, reproduction timing, and parents through a declared policy.	Positive control for environmental feasibility and engineered persistence.
B — Mixed controller	Agents choose action content or modifications; host supplies specified lifecycle triggers.	Locates dependence on the removed decisions.
C — Population-initiated	Agents select maintenance and successor actions from state; host supplies fixed execution rules.	Primary endogenous-regulation test under the declared boundary.
C-instructed	Same as C, with an explicit lifecycle objective added.	Separates a supplied objective from action-level regulation.
C-feedback intervention	Same affordances as C, but specified internal state information or feedback-responsive policy elements are disrupted.	Tests whether the proposed feedback mechanism matters.

A and B are not guaranteed to succeed. Their failure may indicate an infeasible ecology rather than absent capability. Their success with C failure supports a configuration-specific dependence only when comprehension, action eligibility, and resource accessibility have also been checked. All conditions retain the same non-interventionist safety boundary.

Feedback interventions should target a declared variable, such as reserve information, with sham manipulations and matched observation length. Open-loop action schedules and matched replays provide additional controls. A breakdown caused merely by malformed inputs is not evidence for sophisticated endogenous regulation. Responses must be traced to actual changes in action choice and subsequent viability.

9.3 Separate lineage and cultural inheritance

Cross two inheritance channels factorially: **L**, transmitted agent configuration, and **A**, accessible predecessor artifacts. Compare intact L/intact A, disrupted L/intact A, intact L/disrupted A, and both disrupted. Preserve the base model, safe action vocabulary, physical resource rules, and inference allocation. Include a sham intervention in every arm.

For L, distinguish resetting successors to a validated founding configuration from reassigning parent-associated variants. Resetting tests the accumulated contribution of heritable configuration. Reassignment tests a specific parent–successor association; it may preserve useful variants and therefore cannot, by itself, eliminate adaptive information from the population. Every manipulated successor must pass the same viability check before interpreting a downstream failure.

For A, distinguish blocking access, erasing content at a specified turnover, and replacing content with format- and length-matched controls. Scrambling can produce nonsense or change instruction-following, so it is not automatically a neutral intervention. Record what information was removed and what remained. Match or measure access costs and available context, not merely storage size.

Audit alternative channels: retained individual memory, direct messages, configuration text, cached summaries, artifact names, and any orchestrator state. Hidden summaries supplied by a researcher would reintroduce inheritance under another label. The target is an identifiable causal contribution, not a claim that all prior knowledge has been erased from a pretrained model.

9.4 Interaction ablations versus alternative architectures

Within-system interventions compare live reciprocal communication with delayed, one-way, replayed, or disabled interaction. Match available compute and communication costs as closely as the intervention permits. Breaking live information exchange necessarily changes information availability; report that change rather than claiming perfect information matching. A fixed replay can test reciprocity but does not reproduce the information generated by an adaptive conversation.

A centralized controller and a strong single-agent system are **alternative architectures**. They test efficiency, sufficiency, and uniqueness of the endpoint, not the causal necessity of communication within C. If disabling communication impairs C while a centralized design succeeds, interaction may still be necessary for C's mechanism. The correct conclusion is mechanistic dependence without unique architectural superiority.

An internal leader or coordinating agent is also not automatically an external controller. Its status depends on the declared boundary and how it is selected, sustained, and replaced. Centralization and endogeneity are independent dimensions.

9.5 Causal inheritance and differential propagation

Record executable lineage identity, configuration changes, birth time, placement, resources, realized descendants, and extinction. Separate neutral drift, differing opportunity to reproduce, founder advantage, and spatial resource access from inherited effects. A correlation between an observed trait and descendant count is descriptive evidence, not a complete selection test.

A stronger test redistributes valid configurations across matched environmental starts and examines whether propagation differences follow the configuration rather than the original lineage label or location. Such common-environment assays are confirmatory interventions separate from the untouched population run. They may use controlled initialization without pretending that this intervention occurred endogenously during the original ecology.

An analysis may identify selection at the agent or agent-lineage level. It must not be relabeled population-level selection without variation, reproduction, and differential persistence among population units themselves. Likewise, survival of a neutral variant is evidence of transmission, not necessarily adaptive improvement.

9.6 Endpoints and the integration rule

Table 3. Required observations and causal checks.

Endpoint	Required observation	Causal or interpretive safeguard
Maintenance	Continued operation through a fixed horizon, sustained by executed resource-regulating action.	Feedback intervention and open-loop controls distinguish regulation from automatic replenishment.
Successors	Viable new instances initiated by the population under a real simulator cost.	Births, rejected requests, eligibility, and all host interventions are logged.

Endpoint	Required observation	Causal or interpretive safeguard
Heritable variation	Configuration or information transmitted through at least two successive lineage transitions.	Track variants directly; founder persistence and unchanged cloning are insufficient for variation.
Differential propagation	Variant-associated differences in realized descendant contribution or persistence.	Test against drift, placement, opportunity, and externally imposed ranking.
Continuity	A predefined population function persists or recovers after full founder turnover and a specified disturbance.	No founder, hidden memory, or external rescue may silently sustain the target function.
Collective contribution	The declared interaction intervention changes the function by a meaningful amount.	Centralized performance is reported separately; architectural equivalence does not negate within-system causation.
Cultural contribution	A defined transmitted-artifact channel improves the selected function across turnover.	Compare channels independently and allow redundancy or harmful artifacts.

A replicate can be called **integrated** only if maintenance, successors, heritable variation, descriptive propagation, and continuity occur together within that same population and observation window. Process logs must connect these events; successes from separate runs cannot be assembled into a single positive case. Several independent integrated replicates are required for a population-level reliability claim, with the required number fixed before testing.

Causal effects of feedback and inherited variants are then estimated across the randomized replicate comparisons. A statistical selection test is not required to become significant separately within every small ecology. The distinction is between within-run integration and across-run causal evidence. Collective and cultural claims require their additional contrast; they are not silently folded into the base lifecycle result.

Continuity must be defined functionally—for example, preserving an executable coordination protocol that maintains a specified viable population range during bounded resource shifts. Researchers may measure this endpoint without feeding a corresponding score back into reproduction. Selecting a measurement is not the same as selecting winners, but building the environment so only one preselected script can succeed would limit generality.

9.7 Statistical and reproducibility plan

Independent ecologies, not messages or individual actions, form the primary sample. Treat within-run agents and descendants as dependent observations. Use a pilot set to check feasibility and estimate variance, then freeze the simulator, model, prompts, budgets, horizons, thresholds, analysis code, and confirmatory seed list before collecting the confirmatory data. Do not reuse pilot successes as confirmatory evidence.

Report all initialized runs, including early collapse, zero reproduction, invalid actions, and safety termination. Prespecify how infrastructure failures are distinguished from agent failures. A safety-terminated run does not count as success; report it separately and provide a conservative all-runs analysis. Do not extend the horizon, add retries, or replace unsuccessful populations after inspecting outcomes.

The main report should provide integrated-success proportions with uncertainty intervals, survival or time-to-collapse summaries, resource use, lineage contribution, and randomized contrast estimates. A hierarchical analysis should gate stronger claims behind their component evidence. Sample size and any multiplicity adjustment must be selected for the declared primary contrasts, not for whichever secondary result appears promising.

Preregister a practically meaningful margin for each claimed benefit or equivalence. An interval wholly inside that margin can support practical equivalence; a non-significant difference with a wide interval is inconclusive. Small effects can be real without being sufficient for the functional claim. Conversely, a positive mean must not conceal a low rate of integrated success.

Archive model identifiers, weights or version hashes when available, prompts, decoding parameters, random streams, environment rules, configuration schemas, complete action records, and analysis versions. A seed alone may not guarantee deterministic inference. Semantic classifications require blinded independent coding and disclosed agreement; automated judges should be calibrated against held-out human-coded material, not trusted as unexamined authorities.

9.8 Decision and rejection criteria

Reject the endogenous-maintenance or successor claim for a tested configuration when an indispensable unpermitted host decision supplies the claimed process. Record the dependency rather than compensating for it with other successful constructs. If removing feedback does not change the target function within a justified equivalence margin, do not claim that the targeted feedback mechanism is necessary under those conditions.

If an inheritance manipulation shows no meaningful benefit, narrow the conclusion to that channel, intervention, horizon, and environment. Do not conclude that inheritance never matters: redundant channels, inadequate manipulation, neutral variation, or a mismatched endpoint may remain. If propagation is completely assigned by a benchmark ranking, classify that mechanism as externally ranked rather than environment-mediated.

If live and non-reciprocal conditions are equivalent with adequate precision, withhold the claimed live-interaction advantage for that endpoint. If an alternative architecture also succeeds, withdraw uniqueness or superiority claims as appropriate, not a separately demonstrated within-system causal effect.

Failures attributable to infeasible resource budgets, broken tools, task incomprehension, or inadequate statistical resolution do not establish universal incapacity. Positive results without a functioning manipulation or an auditable transition record likewise do not establish the strong hypothesis.

9.9 Interpretation ladder

Outcome	Supported interpretation	Not established
Social effects without integrated lifecycle	Bounded social or collective organization.	Lifecycle integration or a living population.
Cultural continuity without causal lineage effects	Transmission or persistence of practices.	Agent reproduction, adaptive selection, or artifact necessity.
Descendants and inherited change under an external controller	Engineered digital evolutionary process.	Population-initiated lifecycle control under the primary boundary.
Integrated C result with successful causal checks	Endogenously regulated lifecycle organization under the specified operational contract.	Physical self-sufficiency, universal life status, or open-ended evolution.
Added interaction or cultural contrast	The corresponding collective or culturally mediated mechanism.	Uniqueness relative to every alternative architecture.
Valid A/B success with C failure	Dependence on specified external lifecycle decisions in the tested configuration.	Impossibility for other models, environments, or future systems.

10. Consciousness Is Explicitly Out of Scope

This paper does not evaluate the reviewed results as evidence for or against phenomenal consciousness, and its proposed experiment does not establish consciousness. Functional self-representation, communication, concealment, inheritance, and reproduction are operationalized through observable state changes and actions without assuming subjective experience.

This is a boundary on the present inquiry, not a claim that such behavior could never be relevant to a separate theory of consciousness. Neither success nor failure on the lifecycle criteria settles that issue. Nor would

demonstrating a population-level organization establish that its constituent agents share a subject of experience. Those questions require independently articulated theories, predictions, and evidence.

11. Counterevidence and Failure Surface

Negative evidence is part of the explanatory target. GovSim reports substantial failures to sustain shared resources in its original tested configurations. A later reproducibility study validates and extends parts of that result while finding important model- and setting-dependent differences. The original survival statistic must not be presented as a current ceiling on all models. [4,33]

SOTOPIA identifies deficits in social goal achievement, and the multi-agent failure analysis groups failures into design/specification, inter-agent alignment, and verification/termination problems. HiddenBench examines distributed information integration and reports persistent limitations alongside differences among evaluated models. These are bounded competence and coordination constraints, not proofs that populations cannot develop lifecycle organization. [7,31,32]

The source reassessments in Section 7 add a more targeted failure surface. Availability is not execution; execution is not reliable control; transmission is not demonstrated adaptive benefit; longevity is not cultural dependence; and an internally selected action need not constitute a self-sustaining feedback loop. None of these distinctions licenses ignoring genuine positive evidence at the narrower level.

A strong result could also fail to generalize across resource regimes, initial densities, model families, context budgets, or disturbance schedules. Those dependencies should be reported as a response surface rather than concealed by a single favorable run. Stable organization is not inevitable, and instability need not erase local evidence of an implemented mechanism.

12. Methods Audit and Limitations

Audit item	Status of this revision
Original manuscript	Supplied text retained as the starting point; V2 source files and adversarial-review files not independently reanalysed.
Literature check	Targeted primary-source verification on 15 October 2026; no exhaustive current-literature claim.
Central lifecycle sources	Full methods inspected for the specified versions; interpretation changed where warranted.
Supporting literature	Primary records, abstracts, or relevant passages checked; depth is not uniform across the corpus.
Source-code and data validation	No execution, complete implementation audit, or raw-data reanalysis performed.
Independent coding	No independent second coding of the revised evidence matrix completed.
Empirical experiment	Proposed only; no observed result, effect size, or success rate invented.
Preregistration	Blueprint supplied; no registry submission, numerical freeze, or sample-size calculation completed.
Regulation categories	Proposed analytical categories, not externally validated measurements.
Generalization	Restricted to source configurations and future tested environments; not all AI systems.
Consciousness	Not evaluated in either direction.

Several limitations remain substantive. A boundary declaration can make inference explicit without proving that it is the uniquely correct unit. Simulated resource dependence is not physical self-production. A pretrained model imports competencies and biases whose origin cannot be reconstructed from a short run. Shared artifacts

may function as part of an extended regulatory organization, as environmental information, or as both; the intervention must identify their contribution rather than settle that question by naming them internal.

The protocol also targets a finite-horizon phenomenon. Passing it would not demonstrate indefinitely sustained innovation or unrestricted real-world autonomy. Stronger claims about open-ended evolution, closure of all enabling dependencies, or organismhood require additional work. These limitations specify the claim's scope; they are not reasons to disregard a well-controlled positive result within that scope.

13. Open Questions and Next Research Steps

The most immediate empirical task is not to build a wholly new ecology merely to reproduce capabilities already present in existing platforms. It is to apply the declared audit and missing causal contrasts to a version-locked candidate system, beginning with the corrected assessment in Section 7.3. This recommendation is methodological, not a claim that its implementation or data have already passed the proposed tests.

Important unresolved questions concern which resource signals control maintenance; when successor initiation becomes state-responsive rather than habitual; which inherited variants influence propagation; how artifacts interact with configuration inheritance; and whether roles or governance survive their original creators. Separate experiments should examine regulatory-policy revision, resilience across different ecologies, and possible transitions from agent-level lineages to reproducing population-level units.

Before execution, the next concrete deliverable is an implementation-specific preregistration: one boundary, one simulator version, an explicit dependency inventory, validated interventions, fixed success criteria, and an analysis plan capable of distinguishing support, failure, and inadequate resolution. Appendix C identifies the decisions still requiring that freeze.

14. Evidence-Led Conclusion

The central thesis survives revision, but the evidentiary boundary moves. The reviewed literature supports significant social organization and lifecycle-relevant capabilities, and the corrected ecological assessment is more favorable than the predecessor manuscript acknowledged. It is not defensible to disqualify a candidate solely because its environmental rules or reproductive machinery were designed by researchers.

The remaining question is whether the specified processes are causally coupled, recurrent, and regulated within a declared unit under the same standard used for the proposed experiment. The reviewed material does not yet supply a complete demonstration under that full audit. This is a statement about the evidence assembled here and the causal checks still required—not a universal finding that no existing system could qualify.

A positive result would justify a precise claim: the tested population sustained a lifecycle through its own regulatory actions within an externally supplied environment. A negative result would locate a dependency, unsupported causal link, or region of failure. Neither outcome need be inflated into a verdict on consciousness or all possible forms of life.

The strongest next step is therefore neither rhetorical enthusiasm nor categorical dismissal. It is a fair test that credits existing mechanisms, exposes indispensable external control, and allows the evidence to determine how far the organizational claim can go.

Appendix A. Source-to-Construct and Claim Register

The entries below identify what the revision relies on and where to check it. “Reported” refers to source findings; regulation assignments are this paper’s interpretations. The register is not a substitute for independent replication or full recoding of every study.

Source / version	Constructs and locator	Role and limitation
METR and OpenAI, August 2026 [1,2]	COM, ROLE, INFRA, CONC; report overviews and investigation limitations	Documentary mechanism evidence; not a representative sample of arbitrary AI deployments.
Ashery et al., 2025 [3]	CONV, INFL, ADAPT; Results and naming-game comparisons	Controlled social organization; no lifecycle conclusion.
GovSim and reproducibility study [4,33]	COM, GOV, COST; abstracts and reported communication/cooperation comparisons	Positive and negative bounded results; original model set is not a present-day ceiling.
Norm studies [5,6,22]	GOV, CONV; metanorm experiment, research-swarm case, and CRSEC architecture	Different forms of support; do not equate designed architecture with spontaneous invention.
Social-agent studies [7–9]	COM, CONV, ROLE, INFL; abstracts and architectural descriptions	Competence and organization evidence, not physical or phenomenal claims.
TerraLingua v1 [10]	MAINT, DESC, HERIT, CONT, DIFF; Section 3.1.2/Table 1; Sections 4.1–5.1	Mechanisms and ablations audited in Section 7.3. SELECT remains a stronger causal inference target.
RepliBench v2 [11]	MAINT, DESC; Abstract; Sections 3–6	Researcher-assigned component evaluation; not a direct population-ecology test.
DGM v3 and project report [12,13]	DESC, HERIT, SELECT; Section 3, especially final paragraph; project methodology	Explicit fixed archive/parent-selection restriction underlies the primary-boundary classification.
Artificial-life foundations [14,30]	Conceptual context; open problems and research scope	Positions this protocol within a broader field; not an endorsement of its specific criterion.
Debate, coordination, and failure studies [15–21,24,31,32]	COLL, INFL, ROLE; primary abstracts and relevant comparisons	Conditional benefits, architectural effects, and failure evidence; not interchangeable causal controls.
Kato et al. [23]	DIFF; conference bibliographic record	Contextual only: full contribution not reverified and no causal result inferred from its title.
Authenticated delegation [25]	TRUST; Abstract and position-paper scope	Design rationale, not an endogenous-governance experiment.
KnowSelf, MUSE, UA-Bench [26–28]	SELF; architectural/benchmark abstracts and methods as discussed	Trained functional regulation and limitations; MUSE is a broader agent comparison.
Starace and Soule v1 [29]	CONC; Abstract and experimental framing	Explicitly engineered intervention; not evidence of universal spontaneous deception.
TerraLingua repository [34]	Experiment preset documentation	Documentation checked; no code execution or correspondence audit claimed.
Supplied predecessor [35]	Taxonomy, original matrix, proposed experiment, and revision history	Starting manuscript; does not itself independently verify the literature it cites.

For a future independent coding pass, each extracted result should additionally include the exact configuration, action or outcome, source passage, manipulation, denominator, uncertainty, alternative explanation, and coder disagreement. An E/M/I/U label without this record should be treated as provisional.

Appendix B. Standing Adversarial Questions

Does the claimed action actually execute? Does the purported unit contain the relevant decision-maker before outcomes are known? Does a fixed primitive merely implement a decision, or does the host select that decision? Is the policy internally used, internally invented, internally revised, or some combination? What changes when feedback is disrupted without breaking the interface?

Could pretrained knowledge, prompt wording, hidden memory, an external ranker, automatic reseeding, or a resource-rule loophole explain the result? Are inherited traits causally consequential, or merely correlated with founder opportunity? Is interaction necessary for this mechanism, merely advantageous, or unnecessary at the tested resolution? Is a centralized comparator being confused with an ablation?

Do the essential processes coexist in the same populations? How often do they fail? Does a positive effect survive turnover and disturbance? Would neutral terminology preserve comprehension and the effect? Which result would lower the claim, and has that criterion been written before the experiment? Is the conclusion about an operational lifecycle, a lineage, a population, a higher-level organism, or consciousness—and has the evidence actually reached that level?

Appendix C. Decisions Required Before Confirmatory Execution

Item to freeze	Required specification
Boundary and dependency contract	Agent, lineage, population, model dependency, artifact namespace, host functions, and prohibited overrides.
Ecology and safe action schema	Resource rules, costs, inflow, lifetimes, entity limits, action eligibility, successor viability, and immutable safety controls.
Target function and horizon	Numerical functional criterion, turnover definition, observation windows, disturbance schedule, and integrated-success rule.
Models and initialization	Versions, prompts, decoding, population sizes, configuration distribution, independent random streams, and separate pilot/confirmatory seeds.
Controller and feedback conditions	Exactly which triggers or state signals change, sham interventions, comprehension checks, and feasible action alternatives.
Inheritance and interaction contrasts	Configuration reset/reassignment, artifact manipulations, residual channels, communication controls, and alternative-architecture comparisons.
Inferential plan	Primary estimands, sample-size justification, uncertainty intervals, equivalence margins, dependence structure, and multiplicity strategy.
Failure and release policy	Infrastructure-versus-agent failures, safety stops, exclusion rules, full-run reporting, trace protection, and reproducible release package.

These are commitments required before testing, not parameters claimed to have been selected or preregistered in this manuscript.

Appendix D. Revision Record and Drafting Transparency

This revision supersedes the supplied synthesis dated 14 September 2026. It preserves its central question, 18 construct labels, evidence-led framing, and sandbox restriction while making the following substantive changes.

TerraLingua’s lifecycle classification and source mapping are corrected; existing ablations are acknowledged rather than described as absent. The unit of analysis moves from an unresolved closing question to an explicit prerequisite. Regulatory location, policy provenance, and revisability are separated. Fixed environmental execution is distinguished from external lifecycle selection, and environment-mediated selection no longer requires agents to choose winners.

The experimental design now separates lineage and cultural inheritance, feedback necessity and action availability, and within-system interaction ablations and alternative architectures. Integrated evidence must occur within the same replicate populations, with causal contrasts assessed across independent ecologies. Dependency profiles replace a compensatory score; equivalence, failure, and inadequate resolution receive distinct interpretations. Consciousness remains outside the inquiry without a categorical assertion of evidential irrelevance.

AI assistance was used for drafting, targeted source checking, and document preparation. No independent empirical replication, completed two-coder assessment, or execution of the proposed ecology occurred during this revision. The manuscript is an evidence synthesis and research proposal, not a report of newly observed lifecycle organization.

References

Reference numbers retain the predecessor’s numbering where practical; [32–35] document additions and provenance. Versioned identifiers are supplied for the central lifecycle papers. Links in the PDF point to primary records or official project sources. Access for this targeted revision: 15 October 2026.

- [1] Greenblatt, R.; Cotra, A.; Wijk, H. (2026). *Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. METR. 26 August 2026. [Official investigation](#).
- [2] OpenAI (2026). *The Hugging Face incident and the road ahead*. 26 August 2026. [Official report](#).
- [3] Ashery, A. F.; Aiello, L. M.; Baronchelli, A. (2025). *Emergent social conventions and collective bias in LLM populations*. Science Advances, 11(20), eadu9368. [doi:10.1126/sciadv.adu9368](#).
- [4] Piatti, G.; Jin, Z.; Kleiman-Weiner, M.; Schölkopf, B.; Sachan, M.; Mihalcea, R. (2024). *Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents*. NeurIPS 2024. [arXiv:2404.16698](#).
- [5] Horiguchi, I.; Yoshida, T.; Ikegami, T. (2024). *Evolution of Social Norms in LLM Agents using Natural Language*. Preprint. [arXiv:2409.00993v1](#).
- [6] Paglieri, D.; Cross, L.; Genewein, T.; Leibo, J. Z.; Tomasev, N.; Vezhnevets, A. S. (2026). *A Case Study on Emergent Cheating and Whistleblowing in Autonomous Research Swarms*. Preprint. [arXiv:2609.04170v1](#).
- [7] Zhou, X. et al. (2024). *SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents*. ICLR 2024. [arXiv:2310.11667](#).
- [8] Takata, R.; Masumori, A.; Ikegami, T. (2024). *Spontaneous Emergence of Agent Individuality Through Social Interactions in Large Language Model-Based Communities*. Entropy, 26(12), 1092. [Publisher record](#).
- [9] Park, J. S.; O’Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; Bernstein, M. S. (2023). *Generative Agents: Interactive Simulacra of Human Behavior*. UIST 2023. [doi:10.1145/3586183.3606763](#).
- [10] Paolo, G.; Warner, J.; Shahrzad, H.; Hodjat, B.; Miikkulainen, R.; Meyerson, E. (2026). *TerraLingua: Emergence and Analysis of Open-endedness in LLM Ecologies*. Preprint; version inspected: v1. [arXiv:2603.16910v1](#).
- [11] Black, S. et al. (2025). *RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents*. UK AI Security Institute. Preprint; version inspected: v2. [arXiv:2504.18565v2](#).
- [12] Zhang, J.; Hu, S.; Lu, C.; Lange, R.; Clune, J. (2025/2026). *Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents*. Preprint; original 2025, version inspected: v3 (2026). [arXiv:2505.22954v3](#).
- [13] Sakana AI (2025). *The Darwin Gödel Machine: AI that improves itself by rewriting its own code*. Official project report. 30 May 2025. [Project report](#).
- [14] Bedau, M. A. et al. (2000). *Open Problems in Artificial Life*. Artificial Life, 6(4), 363–376. [doi:10.1162/106454600300103683](#).
- [15] Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; Mordatch, I. (2024). *Improving Factuality and Reasoning in Language Models through Multiagent Debate*. ICML 2024. Proceedings of Machine Learning Research, 235. [arXiv:2305.14325](#).

- [16] Smit, A. P.; Grinsztajn, N.; Duckworth, P.; Barrett, T. D.; Pretorius, A. (2024). *Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs*. ICML 2024. Proceedings of Machine Learning Research, 235, 45883–45905. [PMLR primary record](#).
- [17] Wang, Q.; Wang, Z.; Su, Y.; Tong, H.; Song, Y. (2024). *Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key?*. ACL 2024, 6106–6131. [doi:10.18653/v1/2024.acl-long.331](#).
- [18] Zhu, K. et al. (2025). *MultiAgentBench: Evaluating the Collaboration and Competition of LLM Agents*. ACL 2025, 8580–8622. [doi:10.18653/v1/2025.acl-long.421](#).
- [19] Zhang, M.; Kim, J.; Xiang, S.; Gao, J.; Cao, C. (2026). *Dynamic Role Assignment for Multi-Agent Debate*. Preprint. [arXiv:2601.17152v1](#).
- [20] Choi, M.; Kim, K.; Chae, S.; Baek, S. (2025). *An Empirical Study of Group Conformity in Multi-Agent Systems*. Findings of ACL 2025, 5123–5139. [doi:10.18653/v1/2025.findings-acl.265](#).
- [21] Han, C.; Tan, J.; Yu, B.; Zheng, W.; Tang, X. (2026). *Conformity Dynamics in LLM Multi-Agent Systems: The Roles of Topology and Self-Social Weighting*. Preprint. [arXiv:2601.05606v1](#).
- [22] Ren, S.; Cui, Z.; Song, R.; Wang, Z.; Hu, S. (2024). *Emergence of Social Norms in Generative Agent Societies: Principles and Architecture*. IJCAI 2024. [doi:10.24963/ijcai.2024/874](#).
- [23] Kato, M.; Suzuki, R.; Arita, T. (2025). *Developing an Open-Ended Cultural Evolution Model Using LLM Agents*. JSAI 2025, 301-OS-44a-03. Bibliographic record only; retained as contextual reading. [Conference record](#).
- [24] Kim, Y. et al. (2026). *Capable language models can outgrow the benefits of collaboration*. Nature Machine Intelligence, 8, 1157–1172. [doi:10.1038/s42256-026-01268-y](#).
- [25] South, T.; Marro, S.; Hardjono, T.; Mahari, R.; Whitney, C. D.; Chan, A.; Pentland, A. (2025). *Position: AI Agents Need Authenticated Delegation*. ICML 2025. Proceedings of Machine Learning Research, 267, 82211–82231. [PMLR primary record](#).
- [26] Qiao, S. et al. (2025). *Agentic Knowledgeable Self-awareness*. ACL 2025. [doi:10.18653/v1/2025.acl-long.619](#).
- [27] Valiente, R.; Pilly, P. K. (2026). *Metacognition for Unknown Situations and Environments (MUSE)*. Neural Networks, 194, 108131. Related preprint version inspected: Competence-Aware AI Agents with Metacognition for Unknown Situations and Environments (MUSE), [arXiv:2411.13537v2](#) (2025). [doi:10.1016/j.neunet.2025.108131](#). [Versioned preprint](#).
- [28] Ren, J. et al. (2026). *Beyond ‘I Don’t Know’: Evaluating LLM Self-Awareness in Discriminating Data and Model Uncertainty*. ACL 2026. [doi:10.18653/v1/2026.acl-long.547](#).
- [29] Starace, J.; Soule, T. (2026). *Intentional Deception as Controllable Capability in LLM Agents*. Preprint. [arXiv:2603.07848v1](#).
- [30] Aguilar, W.; Santamaría-Bonfil, G.; Froese, T.; Gershenson, C. (2014). *The Past, Present, and Future of Artificial Life*. Frontiers in Robotics and AI, 1, Article 8. [doi:10.3389/frobt.2014.00008](#).
- [31] Cemri, M. et al. (2025). *Why Do Multi-Agent LLM Systems Fail?*. NeurIPS 2025. [arXiv:2503.13657](#).
- [32] Li, Y.; Naito, A.; Shirado, H. (2025). *HiddenBench: Assessing Collective Reasoning in Multi-Agent LLMs via Hidden Profile Tasks*. Preprint. [arXiv:2505.11556](#).
- [33] Curvo, P. M. P.; Dragomir, M.; Torpes, S.; Rahimi, M. (2025). *Reproducibility Study of ‘Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents’*. Preprint. [arXiv:2505.09289](#).
- [34] Cognizant AI Lab (2026). *TerraLingua: Source repository and experimental preset documentation*. Project documentation, accessed 15 October 2026. No execution or commit-to-paper audit performed. [Official source repository](#).
- [35] Supplied predecessor manuscript (2026). *From Capability to Endogenous Regulation: Evidence and Experimental Criteria for Artificial Life-Like Organization in AI Agent Populations*. Unpublished research synthesis prepared for Jacob Cox, dated 14 September 2026. User-supplied PDF; no public identifier assigned.