International Journal of Advanced Multidisciplinary Application | IJAMA
Volume 3 Issue 11 Nov 2026
ISSN No: 3048-9350

INTELLIGENCEOGENESIS:A Multilevel Theoretical and Computational Framework for Intelligence Evolution Beyond AGI

Abdelouahab ZALAGHI, Hamid TOUFIK and Fatima LAMCHOURI

MPCAE Laboratory - Materials, Natural Substances, Environment and Modeling Team (MSNEM), Faculty of Applied Sciences of Taza (FSAT), Sidi Mohamed Ben Abdellah University (USMBA) of Fez, B.P.: 1223 Taza-Gare. Taza, Morocco.

Corresponding Author:
Dr. Abdelouahab ZALAGHI, PhD
E-mail: dr.a.zalaghi@gmail.com

Abstract

Artificial intelligence has traditionally been evaluated through capability, accuracy, generalization, optimization, autonomy, reasoning, and task performance. Evolutionary biology, complexity science, collective intelligence, distributed cognition, artificial life, systems theory, ecological science, and trustworthy AI provide complementary accounts of how organization can persist, diversify, coordinate, and transform through interaction. Recent scholarship explicitly frames human–AI interaction as a reciprocal feedback process and asks whether evolvable artificial systems could participate in processes analogous to major evolutionary transitions [1–4]. This manuscript proposes Intelligenceogenesis as an interdisciplinary research program for studying how intelligence emerges, differentiates, integrates, transforms, and potentially reorganizes across biological, human, artificial, ecological, and institutional levels through time. The central question is: what can intelligence become when it no longer evolves alone?

The framework is integrative rather than replacementist. It does not claim that existing theories of artificial intelligence, evolution, collective intelligence, cognition, artificial life, complex adaptive systems, or consciousness are incomplete in themselves; rather, it proposes a common object of investigation: the evolving organization of intelligence across interacting levels. Five central constructs are formalized: Intelligence Ecogenesis, Bio-Eco-Artificial Intelligence, Trans-Individual Intelligence, Emergent Agency, and Emergent Trustworthiness. The framework also introduces UMET, an equilibrium-oriented computational perspective inspired by organizational principles associated with microbiomes and adaptive biological systems, while explicitly separating biological analogy from validated computational mechanism.

A state-space formulation, coupled human–AI and ecological dynamics, proposed indices for emergence and irreducibility, multi-objective equilibrium optimization, experimental regimes A–E, matched-resource controls, ablation protocols, and falsifiable hypotheses H1–H15 are specified. The mathematical formalism is intentionally modular: the equations are not presented as universal laws, but as operational scaffolds that can be implemented, stress-tested, rejected, or replaced. The manuscript therefore presents theoretical and methodological contributions only: a vocabulary, a formal architecture, candidate measures, experimental designs, and criteria for distinguishing coordination from stronger forms of system-level organization.

The principal contribution is a testable bridge between AI, evolutionary organization, complexity science, ecological coupling, human–AI coevolution, systems theory, and AI governance. The framework proposes that the unit of analysis may need to shift, in some experimental regimes, from an isolated model to a coupled intelligence–environment system. Any stronger claim—that a new evolutionary individual, emergent agency, ecosystem-level intelligence, or artificial consciousness has actually arisen—remains an empirical question requiring controlled experiments and independent evidence [1–4].

Keywords: ***Intelligenceogenesis; Intelligence Ecogenesis; Bio-Eco-Artificial Intelligence; evolvable AI; human–AI coevolution; collective intelligence; trans-individual intelligence; emergent agency; emergent trustworthiness; dynamic equilibrium; UMET; artificial life; major evolutionary transitions; trustworthy AI; complex adaptive systems.Introduction***

1.1 Background and Problem Definition

Artificial intelligence is commonly framed as an engineering discipline concerned with constructing systems that learn, predict, generate, plan, reason, interact, and act. This framing has enabled measurable progress because capabilities can be associated with datasets, objectives, benchmarks, computational resources, and evaluation protocols. Yet the increasing embedding of AI into social and technical environments changes the object of analysis. A model may affect users who generate subsequent data; those data can alter future model behavior; model outputs can influence institutions and infrastructures; and environmental responses can feed back into future decisions. Pedreschi et al. describe human–AI coevolution as continuous mutual influence and emphasize the importance of feedback-loop mechanisms and systemic outcomes [1]. The issue is therefore not whether component-level evaluation remains useful—it does—but whether component-level evaluation is sufficient for every scientific question concerning intelligence.

The present framework begins from a narrower proposition than a claim of artificial consciousness or technological destiny. It asks whether intelligence can itself be studied as an evolving organizational phenomenon. Collective intelligence research has shown that groups can display measurable group-level performance [5]. Extended-mind and distributed-cognition approaches have argued that some cognitive processes can depend on relations among agents, artifacts, representations, and environments [6]. Major evolutionary transitions research asks when previously independent entities become integrated into a higher-level organization with altered units of selection [7–9]. Recent work on evolvable AI and possible human–AI evolutionary individuality makes the connection to artificial systems increasingly explicit [2,3]. Intelligenceogenesis uses these traditions as inputs while retaining their differences.

The shift proposed here is therefore a change in unit of analysis rather than a rejection of conventional AI. A language model, robot, recommender, or autonomous agent can still be assessed using task-level metrics. Intelligenceogenesis adds trajectory-level questions: how does the system change through feedback; which functions depend on coupling; what persists after perturbation; what information and resources are transferred; whether adaptation is reciprocal; whether selection acts at more than one level; and whether a system-level function remains after matched reconstruction. These questions are designed to make the framework empirically vulnerable. If apparent emergence disappears when compute, memory, information, specialization, redundancy, or communication are matched, the stronger interpretation should be rejected.

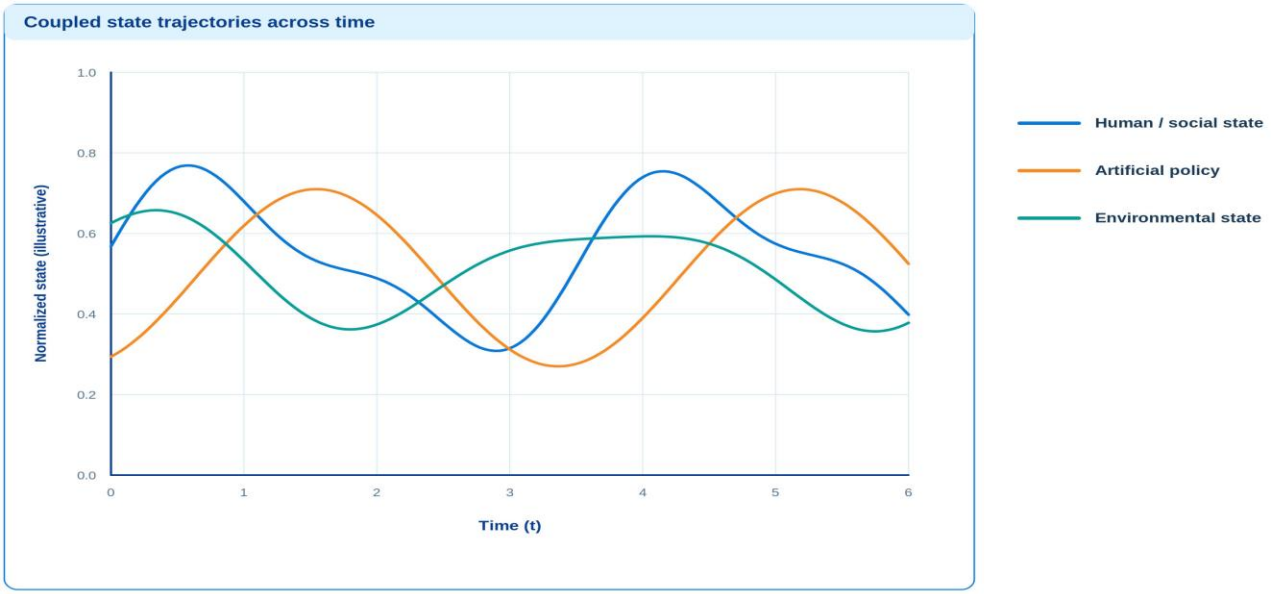

Figure 1. Coupled state trajectories: conceptual state coordinates, not empirical observations.

1.2 Scientific Positioning and Genealogy

The framework is positioned at the intersection of several established traditions. Evolutionary computation contributes variation, selection, adaptation, and search [10]. Artificial life contributes life-like organization, self-organization, evolutionary dynamics, and open-endedness. Multi-agent systems contribute interaction among autonomous computational entities. Collective intelligence contributes group-level performance and measurable aggregation effects. Distributed and extended cognition contribute analyses in which cognitive functions depend on people, artifacts, representations, and environments. Complex adaptive systems contribute nonlinear dynamics, feedback, attractors, adaptation, and emergence. Ecological science contributes coupling between organisms, resources, environments, and selection pressures. Trustworthy AI contributes risk management, accountability, robustness, transparency, and human-centered evaluation [5–9, 11–15].

The genealogy is cumulative rather than competitive. Major evolutionary transitions are relevant because they provide conceptual tools for studying when previously separate units become organized into higher-level individuals or collectives [7,9]. The broader evolutionary and systems perspective also draws on theories of self-organization and complex adaptive systems, which examine how large-scale organization can arise from interactions among simpler components [16–18]. Collective intelligence is relevant because group performance can exceed simple individual baselines [5]. Distributed cognition is relevant because cognition can depend on external representations and coordinated activity [6]. Human–AI coevolution is relevant because reciprocal feedback can modify both human behavior and AI systems [1]. Evolvable AI is relevant because it explicitly considers artificial systems that may participate in evolutionary processes beyond fixed optimization [2]. Rainey and Hochberg further discuss the possibility that increasing human–AI interdependence could eventually be analyzed using the conceptual language of evolutionary individuality [3]. None of these sources establishes Intelligenceogenesis as a field; they define the scientific neighborhood in which the proposed synthesis can be tested. Table 1 summarizes the established focus of the adjacent research traditions and identifies the research question retained by Intelligenceogenesis.

Table 1. Comparative positioning and retained research question

Tradition	Established focus	Question retained by Intelligenceogenesis
Evolutionary computation	Variation, selection, optimization	When do evolutionary mechanisms alter the organizational unit of intelligence?
Multi-agent systems	Interaction, coordination, distributed action	When does coordination become persistent functional organization?
Collective intelligence	Group-level performance	Can group capability become a persistent evolutionary property?
Extended/distributed cognition	Cognition beyond individuals	Can distributed cognition become functionally irreducible?
Artificial life	Life-like organization and evolution	Can intelligence itself become an evolving organizational object?
Human–AI coevolution	Reciprocal human–AI feedback	Can recursive feedback produce higher-order organization?
Major evolutionary transitions	New levels of individuality	Could coupled intelligence systems exhibit analogous transition criteria?
Trustworthy AI	Risk, accountability, robustness	Can trust-related properties be characterized at system level?

1.3 From Artificial Intelligence to Intelligenceogenesis

The conceptual transition can be expressed as four nested questions: how capable is a component; how well do components coordinate; how does the coupled system adapt over time; and does the organization itself become persistent and functionally consequential? The first two questions are already central to AI and multi-agent research. The third is increasingly addressed

by coevolutionary and adaptive-systems approaches. The fourth motivates the Intelligenceogenesis research program and its focus on persistent, multilevel organization. The distinction matters because a system can be highly coordinated without constituting a new organizational unit, and an evolutionary algorithm can implement selection without producing a new individual. The framework therefore refuses to equate communication, performance gain, or adaptation with emergence.

Reciprocal coupling can be represented as a feedback relation in which changes in artificial policy alter the environment or human strategy, while those altered states subsequently modify the next policy update. The curve below is therefore a conceptual representation of nonlinear feedback, not an empirical law.

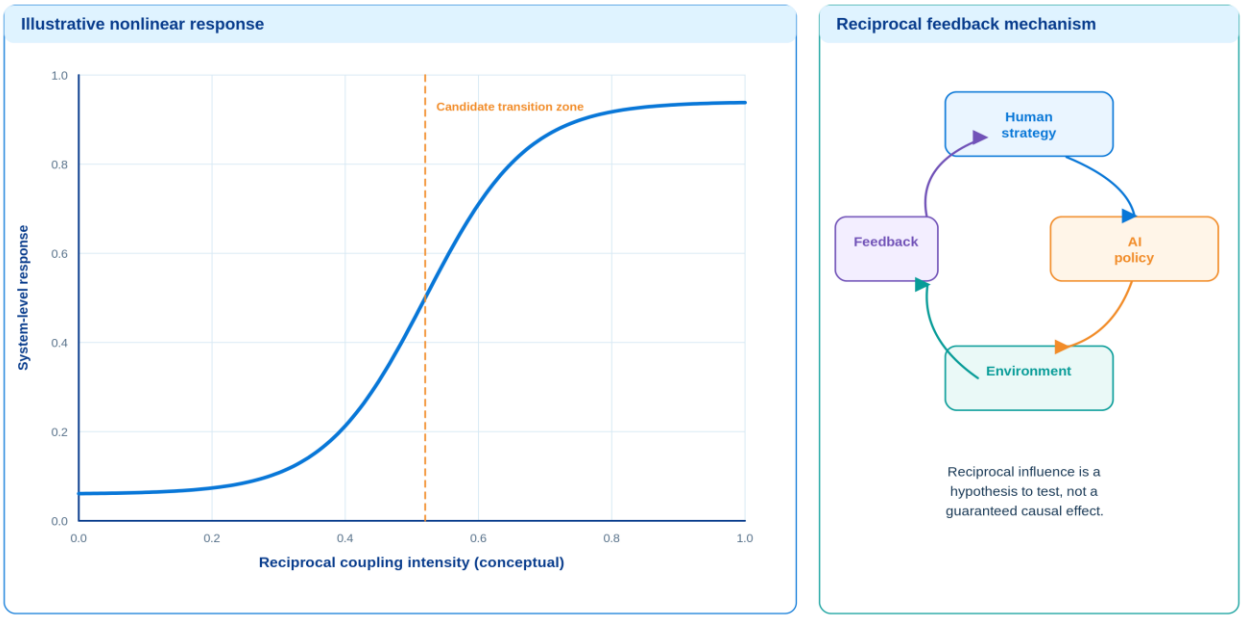

Figure 2. Nonlinear reciprocal coupling; the curve is illustrative rather than empirical.

The proposed synthesis trajectory is: Designed AI → Adaptive AI → Multi-Agent AI → Evolvable AI → Ecological AI → Bio-Eco-Artificial Evolution, followed by hypotheses concerning Trans-Individual Intelligence, Emergent Agency, and Emergent Trustworthiness. These arrows are not predictions. They represent candidate transitions that require operational criteria. In particular, each transition should be associated with measurable changes in persistence, information flow, adaptation, dependency structure, selection, and recovery under perturbation. The value of the trajectory is methodological: it supplies a sequence of increasingly demanding experimental regimes rather than a claim that all AI systems will naturally progress through them.

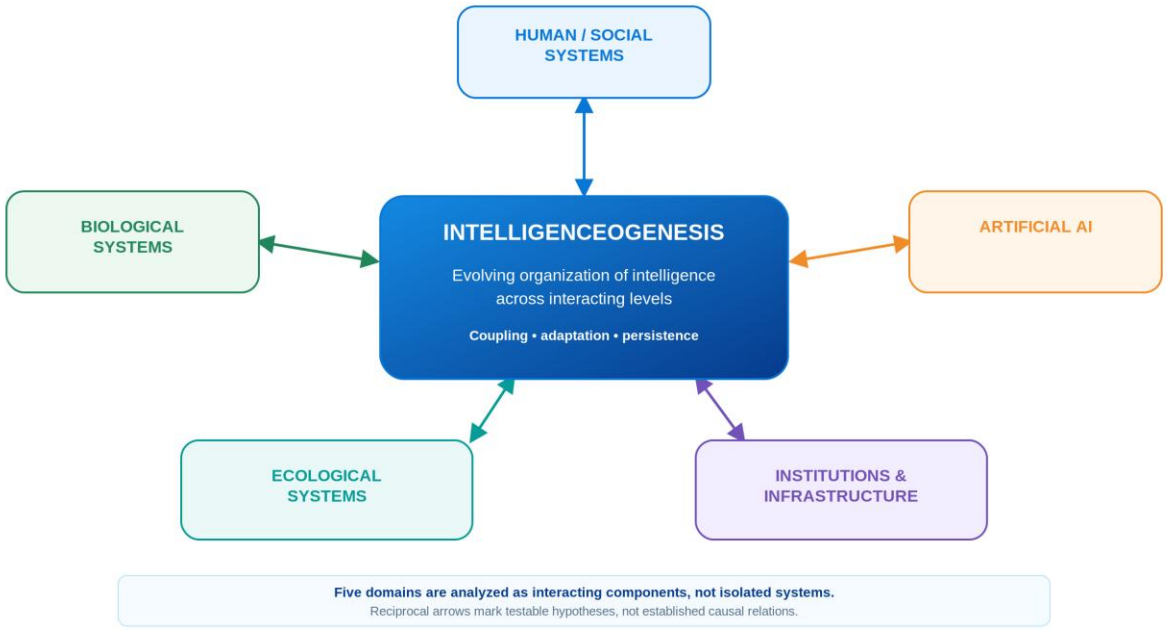

Figure 3. Intelligenceogenesis conceptual architecture. The arrows represent testable reciprocal coupling hypotheses, not established causal relations.

1.4 Core Definitions and Conceptual Boundaries

Intelligenceogenesis is defined here as the interdisciplinary study of how intelligence emerges, differentiates, integrates, transforms, and potentially reorganizes into new evolutionary levels across biological, human, artificial, ecological, and institutional systems through time. Intelligence Ecogenesis is the study of how interactions among heterogeneous intelligent entities and their environments generate, stabilize, or transform new organizational forms of intelligence. Bio-Eco-Artificial Intelligence denotes a coupling paradigm in which biological, human, artificial, and ecological processes are analyzed as mutually influencing components rather than as isolated domains. Trans-Individual Intelligence denotes persistent functional organization distributed across multiple agents, artifacts, memories, environments, or institutions such that a relevant capability cannot be adequately attributed to one component alone.

Emergent Agency is defined as persistent system-level goal-directed or self-regulating behavior that is not adequately explained by treating participating components as independent agents. Emergent Trustworthiness is the hypothesis that trust-related system properties can arise from interaction patterns involving feedback, calibration, recovery, accountability, robustness, and context, and therefore may require system-level evaluation. These definitions deliberately avoid claims about consciousness. Intelligenceogenesis does not imply phenomenal experience; emergent agency does not imply consciousness; ecological coupling does not imply biological life; functional irreducibility does not imply metaphysical unity; and evolutionary persistence does not imply ethical desirability or human benefit. These negative boundaries are essential to falsifiability. Table 2 distinguishes the conceptual commitments of each construct from claims that do not follow from it.

Table 2. Conceptual boundaries

Concept	Does imply	Does not by itself imply
Capability	Measurable function of a component	Agency, consciousness, evolution
Coordination	Interaction improving joint function	New evolutionary individual
Emergence	System-level property requiring controlled explanation	Mystical or non-causal phenomenon
Functional irreducibility	Persistent dependence on organization under matched controls	Metaphysical unity

Concept	Does imply	Does not by itself imply
Emergent agency	Persistent goal-directed/self-regulating system behavior	Phenomenal consciousness
Emergent trustworthiness	System-level trust-related properties	Moral goodness or universal safety
Ecological coupling	Reciprocal environment–agent influence	Biological life

1.5 Intelligence Evolution and Regime Transitions

The framework organizes intelligence evolution around six dimensions: variation, inheritance, selection, persistence, ecological coupling, and emergent organization. This multilevel view is consistent with evolutionary approaches that recognize genetic, epigenetic, behavioral, and symbolic forms of inheritance, as well as the interaction between cooperation, selection, and the emergence of higher-level organization [19,20]. Variation creates differences in policies, architectures, strategies, or behaviors. Inheritance transfers useful structure through learning, memory, parameter transmission, cultural transmission, or reproduction. Selection changes the relative persistence of variants according to explicit or implicit criteria. Persistence captures whether organizational properties survive across time and perturbations. Ecological coupling captures reciprocal interaction with changing environments. Emergent organization refers to system-level structures or functions that become persistent and explanatory after simpler alternatives have been controlled.

These dimensions are not assumed to appear simultaneously. Regime A isolates component capability; Regime B adds interaction; Regime C adds variation, inheritance, and selection; Regime D introduces reciprocal human–AI adaptation; Regime E adds ecological variables, institutional rules, and resource constraints. The transition from one regime to another should be experimentally explicit. A system should not be classified as evolutionarily organized merely because it is adaptive within one lifetime. Likewise, a multi-agent system should not be called an evolutionary individual unless there is evidence that the relevant organization persists and is subject to selection at the higher level. Table 3 operationalizes the five experimental regimes by specifying their interaction structures, evolutionary processes, environmental conditions, and primary research questions.

Table 3. Experimental regimes A–E

Regime	Interaction	Evolutionary process	Environment	Primary question
A — Isolated AI	Minimal	No	Static/controlled	What can a component do alone?
B — Multi-Agent AI	Yes	No	Controlled	Does coordination improve joint function?
C — Evolvable AI	Yes	Variation + inheritance + selection	Controlled/dynamic	Do evolutionary mechanisms alter organization?
D — Human–AI coevolution	Reciprocal	Optional/iterative	Social	Do humans and AI change each other's trajectories?
E — Bio-Eco-Artificial ecosystem	Reciprocal, multi-level	Yes/optional	Dynamic ecological, institutional, and resource environment	Do persistent higher-order functions emerge?

1.6 Intelligence Explosion versus Intelligence Evolution

Intelligence explosion and Intelligence Evolution refer to different analytical questions. An intelligence-explosion framing generally emphasizes rapid capability amplification, recursive improvement, or scaling dynamics. Intelligence Evolution instead asks how organization changes through time under variation, inheritance, selection, persistence, ecological coupling,

and multi-level feedback. A system could improve rapidly without becoming a new evolutionary unit; conversely, organizational evolution could occur slowly without dramatic capability growth. The distinction is useful because it prevents capability scaling from being treated as evidence for organizational transformation.

The proposed framework therefore separates three axes: capability growth, organizational integration, and evolutionary persistence. Capability growth can be measured with conventional benchmarks; organizational integration can be evaluated using dependency, information-flow, and reconstruction tests; evolutionary persistence can be examined through repeated selection cycles and transmission. The three axes may correlate in some systems, but they should not be conflated. This separation also reduces the risk of interpreting speculative trajectories as predictions.

2. Materials and Methods

2.1 Mathematical and Computational Formalization

The system is represented by a multilevel state vector containing human/social state, artificial-agent state, biological/ecological state, institutional/infrastructure state, resources, and environmental conditions. The formulation is intentionally general enough to accommodate discrete-time simulation, agent-based models, dynamical systems, reinforcement learning, evolutionary algorithms, or hybrid computational architectures. Parameters should be estimated or specified transparently for each experiment; the equations below are proposed scaffolds rather than validated universal laws. The multilevel state vector is defined in Equation (1), while its general transition rule is specified in Equation (2). Equations (3, 4, 5) specify agent-level updating, reciprocal human–AI adaptation, and environmental dynamics, respectively.

$$[{}_tH_t, A_t, B_t, I_t, R_t, N] = {}_tS$$

(S_t): $Theoverallmultilevelsystemstateattime(t)$.

(H_t): $Humanandsocialstate$.

(A_t): $Artificial-agentstate$.

(B_t): $Biologicalandecologicalstate$.

(I_t): $Institutionalandinfrastructuralstate$.

(R_t): $Resourcestate$.

(N_t): $Surroundingenvironmentalstate$.

Notation note: (R_t), denoting the resource state, is distinct from (R), the resilience measure introduced in Equation (6).

Equation 1. Multilevel state vector at time t.

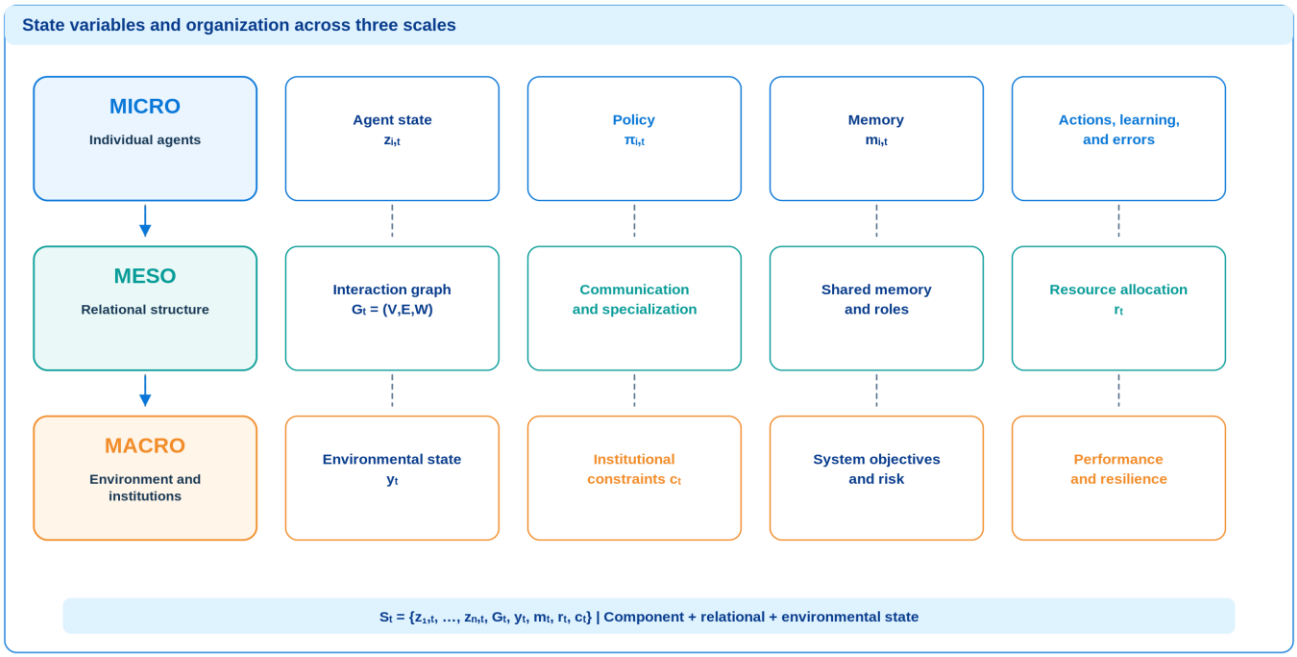

Figure 4. Multilevel state-space architecture used to formalize coupled intelligence–environment dynamics.

$$F(S_t, X_t, U_t, \Phi_t, C_t; \theta) = {}_{t+1}S$$

$-$ (F): *The system transition function.*

$-(S_t)$: *The current system state.*

$-(X_t)$: *Available information or observations.*

$-(U_t)$: *Action or decision vector.*

$-(\Phi_t)$: *Feedback inputs.*

$-(C_t)$: *Constraint inputs or constraint set.*

$-$ (θ): *Model parameter vector.*

Equation 2. General system transition with information X, actions U, feedback Φ, constraints C, and parameters θ.

$$_{i,t}E + {}_{i,t}\mathrm{Sel} + {}_{i,t}L + {}_{i,t}V + {}_{i,t}x = {}_{i,t+1}x$$

$-(x_{i,t})$: *State of agent \(i\) at time \(t\).*

$-(V_{i,t})$: *Contribution of variation.*

$-(L_{i,t})$: *Contribution of learning.*

$-(\mathrm{Sel}_{i,t})$: *Contribution of selection.*

$-(E_{i,t})$: *Environmental and interaction effects.*

This is a decompositional representation. Each term should be operationally defined and measured on a compatible scale before the terms are combined. For empirical implementation, the component terms should be normalized or otherwise expressed on commensurate scales so that their additive decomposition is interpretable.

Equation 3. Agent update decomposed into variation, learning, selection, and environmental/interaction effects.

$$a_{t+1} = f_a(a_t, h_t, e_t),$$
$$h_{t+1} = f_h(h_t, a_t, e_t).$$

$-(a_t)$: *State of the artificial system.*

$-(h_t)$: *Human state or strategy.*

$-(e_t)$: *Environmental state.*

$-(f_a)$: *Artificial-system update function.*

$-(f_h)$: *Human-state update function.*

This formulation represents reciprocal adaptation rather than a purely one-directional influence.

Equation 4. Reciprocal human–AI coevolution dynamics.

$$e_{t+1} = f_e(e_t, a_t, h_t, r_t)$$

$-(e_t)$: *Environmental state.*

$-(a_t)$: *Artificial-system state.*

$-(h_t)$: *Human state.*

$-(r_t)$: *Available resources.*

$-(f_e)$: *Environmental update function.*

Equation 5. Environmental dynamics coupled to agent and human activity.

Dynamic equilibrium can be represented in more than one way. A fixed point satisfies $\lim_{t\to\infty} \left||S_{t+1} - S_t|\right| \to 0$, but many adaptive systems do not converge to a fixed point. A more general criterion can be defined through bounded recurrence, invariant regions, attractors, or acceptable long-run distributions. Therefore, UMET should not use convergence to zero as its only equilibrium criterion. Instead, experiments should define a tolerance region Ω in which performance and safety variables remain within specified bounds while adaptation remains responsive to disturbances. The proposed resilience ratio and the equilibrium-oriented objective function are formalized in Equations (6) and (7), respectively.



Figure 5. Dynamic Equilibrium: Bounded Recovery in State Space

$$R = \frac{C_{\text{rec}}}{D} \qquad C_{\text{rec}} = \frac{1}{T_{\text{rec}}/T_{\text{ref}} + \varepsilon} \qquad D = \frac{\|d\|}{D_{\text{ref}}} > 0$$

(6a) ***(6b)*** ***(6c)***

$-$ (R): *Resilience score.*

$-(C_{\text{rec}})$: *Recovery-capacity term.*

$-(T_{\text{rec}})$: *Actual recovery time following a disturbance.*

$-(T_{\text{ref}})$: *Preregistered reference recovery time.*

$-$ (ε): *Small positive dimensionless constant introduced to prevent division by zero.*

$-$ (D): *Normalized disturbance magnitude.*

$-$ (d): *Disturbance vector or measured disturbance.*

$-(D_{\text{ref}})$: *Reference disturbance magnitude.*

$-$ $(|d|)$: *Norm of the disturbance vector.*

The normalization in Equations (6b) and (6c) is introduced to improve comparability across experimental conditions. The reference recovery time and reference disturbance magnitude should be specified before experimentation.

Equation 6. Proposed resilience ratio; disturbance and recovery definitions must be preregistered.

where (T_{rec}) denotes the observed recovery time following a predefined disturbance, (T_{ref}) is a preregistered reference time, (d) is the disturbance vector, (D_{ref}) is a reference disturbance magnitude, and $(\varepsilon > 0)$ is a small dimensionless constant. Both recovery time and disturbance magnitude are normalized to improve comparability across experimental conditions. The index is a proposed operational measure rather than a universal law of resilience.

$$\mathbf{J = w_P P + w_A A + w_R R + w_T T + w_S S + w_E E - \lambda Q}$$

$-$ (J): *Aggregate objective function.*

$-$ (P): *Task performance.*

$-$ (A): *Adaptation.*

$-$ (R): *Resilience.*

$-$ (T): *Trustworthiness.*

$-$ (S): *Sustainability.*

$-$ (E): *Ecological compatibility.*

$-$ (Q): *Aggregate systemic risk or constraint-violation measure.*

$-(w_P, w_A, w_R, w_T, w_S, w_E)$: *Objective weights.*

$-$ (λ): *Risk-penalty coefficient.*

The objective components should be normalized or otherwise made comparable. Their weights should be specified before evaluating experimental results. The normalization procedure, weight-selection rule, and penalty scale should be fixed before outcome evaluation or justified through a preregistered decision-theoretic procedure.

Equation 7. Proposed multi-objective equilibrium-oriented objective; Q denotes systemic risk or constraint violation.

The multi-objective form is deliberately transparent. P may represent task performance, A adaptation, R resilience, T trustworthiness, S sustainability, E ecological compatibility, and Q aggregate systemic risk. In empirical work, the weights should be fixed before outcome inspection or estimated by an explicit decision-theoretic procedure. Pareto-front analysis may be preferable to a single weighted scalar when trade-offs are central. The mathematical contribution is therefore not a claim that one objective function is universally correct, but a structured representation of the competing objectives that a coupled intelligence system may have to manage.

2.2 Metrics and Operationalization

The Intelligenceogenesis framework requires metrics that distinguish simple aggregation from stronger organizational effects. The first candidate is the Intelligence Evolution Index (IEI), which combines normalized dimensions of variation, inheritance, selection, persistence, ecological coupling, and adaptation. The index should not be treated as a universal scalar comparable across all systems without calibration. Its principal use is within a preregistered experimental family where component definitions, normalization procedures, and weights are explicit. The proposed Intelligence Evolution Index (IEI) is expressed in Equation (8).

$$IEI = w_v V + w_h H_{\text{inh}} + w_s Sel + w_p P_{\text{persist}} + w_e E + w_a A$$

− (IEI): *Intelligence Evolution Index.*

− (V): *Variation.*

−(H_{inh}): *Inheritance or transmission of information and structure.*

−(Sel): *Selection.*

– (P_{persist}): *Persistence.*

− (E): *Ecological coupling.*

− (A): *Adaptation.*

−$(w_v, w_h, w_s, w_p, w_e, w_a)$: *Weights assigned to the six dimensions.*

Notation note: The subscript in (H_{inh}) *distinguishes inheritance from Shannon entropy, denoted by* (H_{ent}) *in Equation (18).*

Equation 8. Proposed Intelligence Evolution Index.

A second family of measures concerns emergence. An Emergence Index can be constructed only after defining the target system-level function and the strongest plausible null models. Candidate explanatory controls include pooled information, redundancy, specialization, communication, centralized memory, and compute allocation. A third family concerns agency: persistence of goal-directed behavior, recovery after perturbation, adaptive reconfiguration, and absence of a single component that trivially stores and executes the global objective. A fourth concerns trustworthiness: calibration, robustness, failure recovery, accountability, and risk-sensitive behavior at the system level. Table 4 summarizes the proposed operational measures and the essential controls required for their interpretation.

Table 4. Proposed metrics and essential controls

Construct	Candidate operationalization	Essential controls
Functional irreducibility	ΔIR = Psystem − Preconstructed	Matched information, compute, memory, training exposure
Resilience	$[R = \frac{C_{rec}}{D}]$	$(R = C_{rec}/D)$, with $(C_{rec})and(D)$ defined and normalized as specified in Equation (6).
Emergent agency	Goal persistence + recovery + adaptive reconfiguration	Remove central controller; perturb components
Emergent trustworthiness	ETI = ΣwkTk − λF	Calibration, robustness, failure, accountability
Intelligence evolution	IEI = ΣwiXi	Sensitivity to weights; report components separately
Ecological coupling	Reciprocal effect size across agent/environment trajectories	Sham coupling and one-way feedback controls

2.3 Experimental Architecture and Falsifiable Hypotheses

A valid empirical program should use matched tasks, compute budgets, information budgets, memory capacities, training exposure, and perturbation schedules wherever possible. Each regime should include independent seeds, multiple environments, null models, and ablations. The key ablations remove environmental feedback, memory, selection, communication, reciprocal adaptation, or ecological constraints one at a time. Statistical analysis should report effect sizes, uncertainty intervals, sensitivity analyses, and robustness to alternative metric definitions. Where possible, preregistration should specify the primary endpoint, stopping rules, exclusion criteria, and reconstruction procedure before outcome inspection.

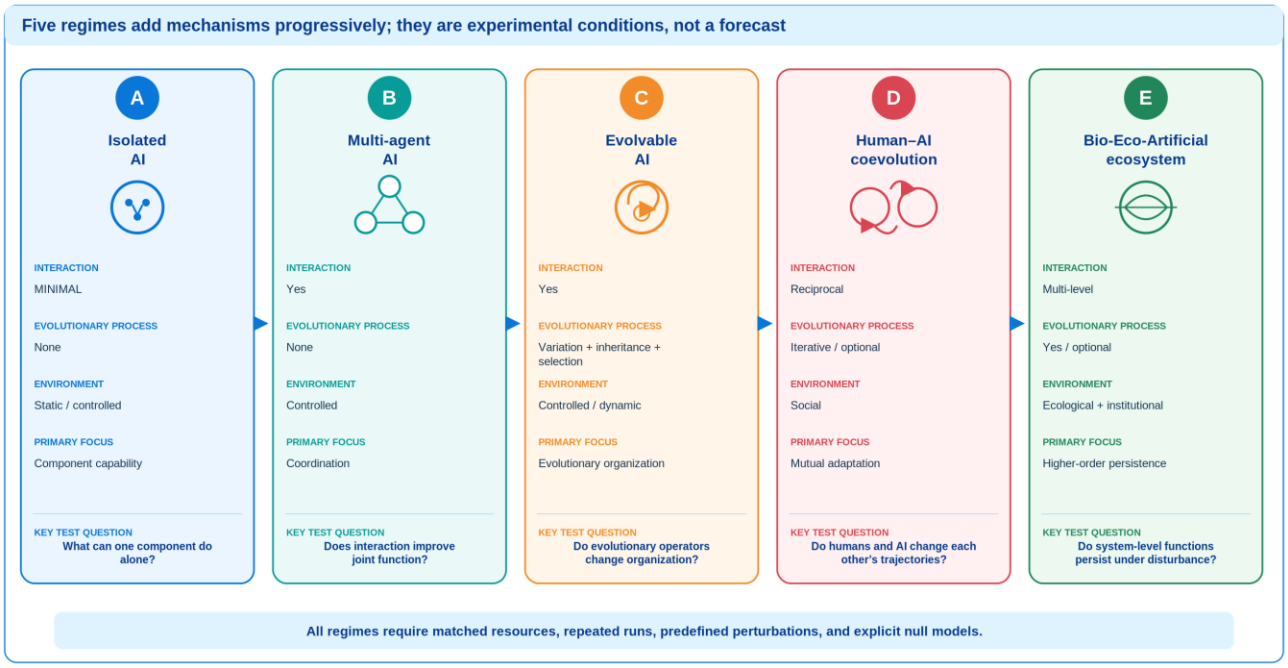

Figure 6. Experimental regimes A–E used to separate capability, coordination, evolution, reciprocal human–AI adaptation, and bio-eco-artificial coupling.

The following hypotheses are proposed as falsifiable research propositions: H1 environmental feedback improves adaptive trajectories under specified conditions; H2 inherited memory stabilizes useful adaptations; H3 system-level selection can generate organization not present in isolated agents; H4 persistent interaction increases functional integration; H5 reciprocal

human–AI adaptation changes both sides' strategies; H6 ecological coupling can generate system-level functions; H7 constrained functional irreducibility can be detected after resource matching; H8 emergent agency can appear under persistence and perturbation criteria; H9 system-level trustworthiness can diverge from component-level trustworthiness; H10 ecosystem-level intelligence requires persistent coupling, adaptation, and higher-level persistence; H11 environmental feedback improves long-term adaptation relative to internal-only feedback; H12 ecological constraints reduce unsustainable local optima relative to unconstrained optimization; H13 agent–environment coevolution produces reciprocal trajectory changes; H14 higher-order organization can emerge above interaction/adaptation thresholds; H15 equilibrium-plus-performance objectives can improve resilience relative to performance-only objectives. Table 5 maps the ten research questions to their corresponding primary hypotheses or experimental tests. The corresponding baseline, ablation, and resource-matched experimental conditions are detailed in Appendix B (Table B1).

Table 5. Research questions and hypotheses

Research question	Primary hypothesis / test
RQ1. Can intelligence be characterized as an evolving organizational property?	Compare component-level capability with persistence and evolutionary organization.
RQ2. Under what conditions do interactions produce higher-order intelligence?	Test matched-resource emergence controls.
RQ3. Does ecological coupling alter adaptation trajectories?	Compare coupled and uncoupled environments.
RQ4. Can persistent human–AI coevolution produce trans-individual function?	Measure reciprocal policy and human-strategy changes.
RQ5. What distinguishes coordination from emergence?	Use reconstruction, pooled-information and centralization controls.
RQ6. Can dynamic equilibrium be a computational design target?	Compare equilibrium-aware and performance-only objectives.
RQ7. Can UMET operationalize equilibrium-oriented AI?	Implement and benchmark UMET control loops.
RQ8. Can ecological constraints reduce unstable local optimization?	Evaluate multi-objective constrained versus unconstrained regimes.
RQ9. Can emergent agency/trustworthiness be measured?	Use persistence, recovery, calibration and accountability metrics.
RQ10. Can ecosystem-level intelligence become an evolutionary unit?	Test higher-level persistence and selection criteria.

2.4 Algorithms and Computational Research Protocol

Invariant A — admissible state

$$S_t \in \Omega \forall t$$

The admissible region Ω encodes hard safety, ecological, institutional, and resource constraints. If the update is projected as $S_{t+1} = \Pi_\Omega\big(F(S_t, \dots)\big)$, with Π_Ω mapping every candidate state into Ω, then $S_{t+1} \in \Omega$. By induction, $S_0 \in \Omega$ implies $S_t \in \Omega$ for every finite t.

The computational program should be modular. A first layer initializes agents and environments; a second defines learning and interaction; a third implements variation, inheritance, and selection where applicable; a fourth records trajectories and resource consumption; a fifth introduces controlled perturbations; and a sixth evaluates candidate system-level properties against null models. Reproducibility requires fixed seeds or documented seed sets, versioned environments, explicit compute budgets, logging of policy and state trajectories, and preservation of all reconstruction procedures.

Algorithm 1 — Generic Intelligenceogenesis Simulation

1. Define the target task, agent population, environment, resource budget, institutional constraints, and primary endpoints before training.
2. Initialize agent states, policies, memory, communication channels, environmental state, and selection parameters.
3. Run a baseline component-level condition without reciprocal environmental adaptation.
4. Run the selected interaction regime and record state, action, information, resource, and outcome trajectories.
5. Apply learning, variation, inheritance, and selection only when they belong to the preregistered regime.
6. Update environmental and institutional states according to the explicit transition functions.
7. Repeat for multiple random seeds, task instances, and environments.
8. Introduce predefined perturbations and measure recovery, reconfiguration, and persistence.
9. Run ablations removing feedback, memory, selection, communication, or reciprocal adaptation.
10. Construct resource-matched reconstructions and pooled-information controls.
11. Estimate component-level and system-level metrics with uncertainty and sensitivity analyses.
12. Report effect sizes, null-model comparisons, failure cases, and all deviations from the preregistered protocol.

Algorithm 2 — Constrained Functional Irreducibility Test

Invariant B — matched resource budget

$B_{system} = B_{reconstructed}$

Compute, memory, communication bandwidth, information access, training exposure, and evaluation budgets must be equalized before comparing organizational performance. Otherwise, a measured gap is confounded by resource advantage rather than organization.

1. Train the original coupled system under the target regime.
2. Construct one or more reconstructions with matched compute, memory, information, training exposure, and task access.
3. Remove or alter the original organizational coupling while preserving the matched resources.
4. Evaluate the coupled and reconstructed systems on identical tasks, environments, perturbations, and seeds.
5. Estimate $\Delta IR = P_{system} - P_{reconstructed}$.
6. Repeat across alternative reconstruction architectures and resource allocations.
7. Control for redundancy, specialization, communication, centralized memory, and pooled information.
8. Test whether any positive performance gap persists across perturbations and environments.
9. Interpret persistent gaps as evidence requiring further investigation rather than proof of a new ontological entity.

Algorithm 3 — UMET Equilibrium-Oriented Control Loop

1. Observe system state, environmental state, resources, performance, risk, and ecological indicators.
2. Normalize or otherwise calibrate the monitored variables using a preregistered procedure.
3. Compute performance, adaptation, resilience, trustworthiness, sustainability, and ecological-compatibility components.
4. Evaluate the current state against the admissible dynamic-equilibrium region Ω.
5. Generate candidate actions using the selected learning or optimization mechanism.
6. Reject or penalize actions that violate hard safety, resource, or ecological constraints.
7. Select actions using the multi-objective criterion or Pareto procedure defined before evaluation.
8. Apply actions and update the environment and institutional state.
9. Measure deviation from Ω and recovery after perturbation.
10. Repeat while storing complete trajectories for reproducibility and post-hoc causal analysis.

Algorithm 4 — Emergent Agency / Trustworthiness Evaluation

INPUT: system S; goals G; perturbation set P; trust metrics K

1. Apply each preregistered perturbation while controlling resource availability.
2. Measure goal-directed persistence, recovery, and adaptive reconfiguration.

3. Measure calibration, robustness, failure handling, accountability, and risk-sensitive behavior.
4. Remove or perturb components and communication pathways.
5. Compare against centralized-controller, pooled-information, redundancy, and matched-resource reconstructions.
6. Report system-level effects only when they persist across perturbations, seeds, tasks, and alternative reconstructions.

2.5 Multiscale Intelligence Dynamics

Algorithm–Metric mapping. ***Algorithm 1*** operationalizes the regime comparison; ***Algorithm 2*** targets ΔIR; ***Algorithm 3*** implements UMET constraint-aware control; ***Algorithm 4*** tests agency/trustworthiness under perturbation; ***Algorithm 5*** integrates preregistration, controls, perturbations, ablations, matched reconstructions, and uncertainty reporting.

The preceding formulation can be extended into a multiscale state-space model in which microscopic agents, mesoscopic organizations, and macroscopic environments evolve simultaneously. This extension is intended to sharpen the theoretical bridge between Intelligenceogenesis and the literature on complex adaptive systems, emergence, network dynamics, and agent-based modeling [21–27]. The central idea is that a collective state should not be defined merely as the arithmetic sum of its components. Instead, the collective state may include interaction topology, memory, resource distribution, institutional constraints, and environmental variables that change the future transition probabilities of the entire system. This creates a formal route for testing whether a system-level description adds explanatory power.

Let each agent i have state z_i, t, policy π_i, t, memory m_i, t, and resource allocation r_i, t. Let G_t be a dynamic interaction graph and $y_t$the environmental state. The complete state therefore includes both component variables and relational variables. In a discrete-time implementation, the graph itself becomes an evolving state variable rather than a fixed background. This representation is consistent with network-science approaches that treat structural properties and interaction patterns as measurable features of complex systems [28]. This matter because an intelligence ecosystem may change not only what agents do, but who interacts with whom, which memories are shared, how resources are allocated, and which feedback pathways remain available. Such dynamics are central to complex adaptive systems and to network-based accounts of emergence [23,24]. The evolving interaction graph, expanded system state, and stochastic transition are formalized in Equations (9,11,12).

$$\mathbf{G_t} = (\mathbf{V_t}, \mathbf{E_t}, \mathbf{W_t})$$

$-(G_t)$: *Dynamic interaction graph at time* (t).

$-(V_t)$: *Set of nodes.*

$-(E_t)$: *Set of edges or links.*

$-(W_t)$: *Edge-weight set or weight matrix.*

Equation 9. Dynamic interaction graph, with nodes $(\boldsymbol{V_t})$***, edges*** $(\boldsymbol{E_t})$***, and edge weights*** $(\boldsymbol{W_t})$.

$$\mathbf{S_t} = \mathbf{z_1}, \mathbf{t}, \ldots, \mathbf{z_n}, \mathbf{t}, \mathbf{G_t}, \mathbf{y_t}, \mathbf{m_t}, \mathbf{r_t}, \mathbf{c_t}$$

$-(S_t)$: *Expanded multiscale system state.*

$-(z_{i,t})$: *State of agent or component* (i) *at time* (t).

$-\ (n)$: *Number of agents or components.*

$-(G_t)$: *Interaction graph.*

$-(y_t)$: *Environmental state.*

$-(m_t)$: *Shared or agent-level memory state.*

$-(r_t)$: *Resource allocation.*

$-(c_t)$: *Constraint state.*

Notation note: Lowercase (c_t) *denotes the constraint state and is distinguished from* (C_t)*, the constraint input in Equation (2).*

Equation 10. Expanded multiscale state including memory m, resources r, and constraints c.

$$\mathbf{S_{t+1} = F(S_t, U_t, \Phi_t, \xi_t; \theta)}$$

$-(S_t)$: *Expanded system state.*

$-(F)$: *System transition function.*

$-(U_t)$: *Action vector.*

$-(\Phi_t)$: *Feedback inputs.*

$-(\xi_t)$: *Stochastic noise or random disturbance.*

$-(\theta)$: *Model parameter vector.*

This formulation extends the system dynamics by explicitly incorporating stochastic influences. Equation (11) is a reduced stochastic representation of the transition dynamics. Information and constraint effects may be represented explicitly as in Equation (2) or absorbed into the state, action, feedback, or parameterization depending on the experimental implementation.

Equation 11. General stochastic transition, with actions U, feedback Φ, noise ξ, and parameters θ.

A major methodological consequence follows from this representation: measurements should be collected at several scales simultaneously. At the micro-scale, researchers should record individual actions, learning updates, policy changes, and errors. At the meso-scale, they should record network structure, specialization, information routing, cooperation, competition, and shared memory. At the macro-scale, they should record environmental trajectories, resource stability, resilience, system-level objectives, and institutional outcomes. The purpose is not to assume that one scale is more real than another, but to test whether a higher-level description becomes necessary for prediction or intervention.

A candidate collective variable q_t can be learned or specified as a coarse-grained representation of S_t. Its scientific value can be evaluated by asking whether q_t predicts future system behavior with comparable accuracy to the full microstate while using fewer effective degrees of freedom. A successful coarse-grained variable would not automatically establish ontological emergence, but it would provide evidence that a macro-level organization captures regularities that are useful for explaining system trajectories. This is consistent with multiscale approaches to complex systems, in which macroscopic patterns are examined through the organization and interaction of lower-level components [6,17,21,23,24]. Equations (12) and (13) define the candidate collective state and its predictive-compression contrast.

$$\mathbf{q_t = C(S_t)}$$

$-(q_t)$: *Candidate collective or macroscopic state.*

$-(\mathcal{C})$: *Coarse-graining or aggregation operator.*

$-(S_t)$: *Full multilevel system state.*

The operator $(\mathcal{C})$*maps the detailed system state into a more compact representation intended to retain information relevant to collective behavior.*

Equation 12. Candidate collective state produced by a coarse-graining operator C.

$$\Delta_{\text{pred}} = \mathcal{L}_{\text{micro}} - \mathcal{L}_{\text{macro}}$$

−(Δ_{pred}): *Difference between microscopic and macroscopic predictive losses.*

−(Lmicro): *Predictive loss obtained using the full microscopic state representation.*

−(Lmacro): *Predictive loss obtained using the macroscopic or coarse-grained state representation.*

A positive (Δ_{pred}) *indicates lower predictive loss for the macroscopic representation, provided that both models are evaluated using the same loss function and held-out test data.*

Equation 13. Predictive-compression contrast between microstate and collective-state models.

2.6 Information Flow, Intervention, and Causal Emergence

Information exchange is necessary for many forms of collective organization, but information exchange alone is not sufficient to demonstrate emergence. Two agents may exchange large amounts of information while each remains independently replaceable. Conversely, a small amount of carefully structured information can create strong functional dependence. Intelligenceogenesis therefore proposes that information-flow measures be combined with interventions. Communication edges can be selectively removed; memories can be randomized; environmental variables can be perturbed; policies can be frozen; and centralized controllers can be introduced as null models. The resulting changes in system behavior provide stronger evidence about mechanism than passive correlation alone.

The distinction between statistical dependence and causal influence is particularly important for emergent agency. If agent A predicts the action of agent B, this does not establish that A causes B. An intervention that changes A while holding relevant variables constant can provide stronger evidence. Similarly, if removing one communication pathway causes system failure, the result is informative only after considering redundancy, alternative pathways, and resource allocation. The framework therefore treats causal intervention as a core part of the evidentiary standard for higher-order organization. The causal intervention effect and conditional information contribution are specified in Equations (14) and (15).

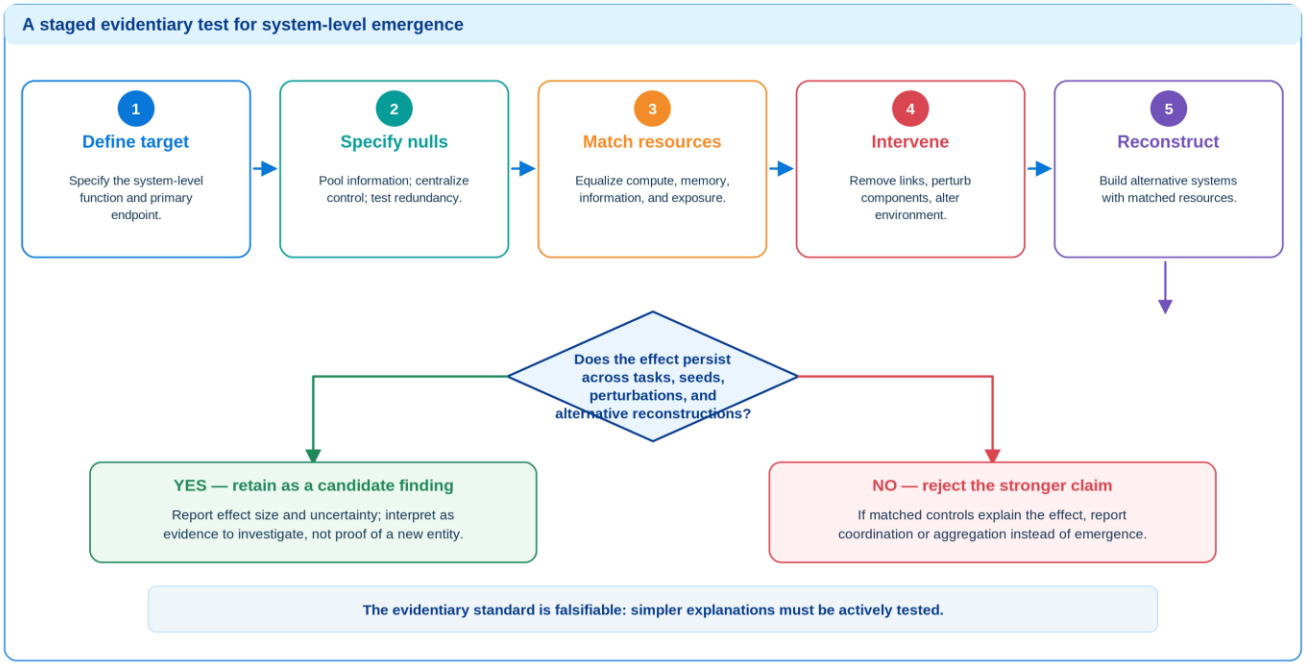

Figure 7. Proposed Emergence Test: Controls, Interventions, and Decision Rule

$$\mathbf{ACE_{i\to j} = E\left[Y_j \middle| do(X_i = x_1)\right] - E\left[Y_j \middle| do(X_i = x_0)\right]}$$

– ($ACE_{i\to j}$): *Average causal effect of intervening on component \(i\) on the outcome of component \(j\).*

– (Y_j): *Outcome variable associated with component \(j\).*

– (X_i): *Intervention variable associated with component \(i\).*

– (x_1, x_0): *Two intervention levels.*

– ($do(\cdot)$): *Causal intervention operator.*

– (E): *Expectation operator.*

Equation 14. Illustrative average causal effect of an intervention on component i.

$$\mathbf{I(X_i; Y_j | C)}$$

−($I(X_i; Y_j \mid C)$): *Conditional mutual information between \(X_i\) and \(Y_j\), given \(C\).*

−(X_i): *State or variable associated with component \(i\).*

−(Y_j):*Outcome or state associated with component \(j\).*

− (C): *Conditioning set of variables.*

This quantity measures conditional statistical dependence. By itself, it does not establish a causal relationship.

Notation note: Here, (C) denotes a conditioning set and should not be confused with the constraint input (C_t) in Equation (2).

Equation 15. Conditional mutual information as a descriptive information-flow quantity, not by itself a causal measure.

Recent work on computational emergence emphasizes that emergent abilities can be connected to particular computational processes rather than being treated as inexplicable properties of a whole system [25]. This supports a mechanistic interpretation of Intelligenceogenesis: every claimed emergent property should be accompanied by a proposed mechanism, a null model, and a test that could disconfirm the interpretation. The framework should therefore reward explanatory compression and reproducibility rather than rhetorical novelty.

2.7 Dynamic Networks and Organizational Phase Transitions

A further extension is to treat the interaction network itself as an evolving object. Agents can create and dissolve links, change communication intensity, specialize, form coalitions, and reorganize after perturbation. Network science provides mathematical tools for studying topology, modularity, centrality, motifs, robustness, and phase-like changes in collective organization [22,24,29,28]. Research on multi-agent algorithms for collective behavior also provides a useful methodological basis for distinguishing interaction mechanisms and collective organizational patterns [30]. Intelligenceogenesis uses these tools to ask whether changes in network structure accompany transitions from coordination to persistent higher-order organization.

A candidate order parameter Ψ can represent a structural or functional property such as modularity, effective integration, persistence of a communication motif, or robustness under targeted disruption. Stigmergic and complex adaptive systems provide additional models for examining how local interactions and indirect coordination can generate persistent collective patterns [31]. A control parameter κ can represent interaction intensity, ecological coupling, selection pressure, or learning rate. A putative organizational transition would be supported only if changes in Ψ are statistically distinguishable from smooth

variation expected under the null model and are reproducible across independent runs. The analysis of nonlinear transitions and changes in system dynamics also draws on established methods for identifying bifurcations, attractors, and changes in dynamical stability [32,33]. This approach makes the notion of a 'new level' experimentally approachable without assuming a universal threshold. The organizational order parameter, candidate transition point, and entropy measure are defined in Equations (16,17,18).

$$\boldsymbol{\Psi(\kappa) = \langle O(G_t, S_t)\rangle_T}$$

$-$ (Ψ): *Organizational order parameter.*

$-$ (κ): *Control parameter.*

$-$ (T): *Number of time steps in the averaging window.*

$-(\mathcal{O})$: *Prespecified organizational measurement function.*

$-(G_t)$: *Interaction graph at time* (t).

$-(S_t)$: *System state at time* (t).

The function $(\mathcal{O})$ *may quantify a defined organizational property, such as connectivity, modularity, or the persistence of interaction patterns. The selected operational measure should be specified before experimentation.*

Equation 16. Time-averaged organizational order parameter.

$$\boldsymbol{\kappa_c = \mathrm{argmax}_\kappa |d\Psi/d\kappa|}$$

$-(\kappa_c)$: *Candidate organizational transition point.*

$-$ (Ψ): *Organizational order parameter.*

$-$ (κ): *Control parameter.*

$-(arg\ max)$: *Value of the control parameter at which the expression is maximized.*

$-(|d\Psi/d\kappa|)$: *Absolute rate of change of the order parameter with respect to the control parameter.*

This expression identifies the point of maximum sensitivity. It does not, by itself, establish the existence of a genuine phase transition; additional statistical and experimental evidence is required.

Equation 17. Candidate transition point based on maximal sensitivity.

$$H_{\text{ent}} = -\sum_{i=1}^{n} p_i \log_b p_i \qquad \textbf{subject to:} \qquad p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1.$$

$-(H_{\text{ent}})$: *Shannon entropy.*

$-(p_i)$: *Probability or normalized relative frequency of category* (i).

$-$ (n): *Number of categories.*

$-$ (b): *Logarithm base.*

$-(log_2 \quad)$: *Produces entropy measured in bits.*

$-(ln)$: *Produces entropy measured in nats.*

The categories represented by (p_i) *must be explicitly defined for each analysis—for example, agent types, strategies, or interaction patterns. Different category distributions should not be combined without justification.*

Equation 18. Entropy of a distribution p_i; usable as one descriptive measure of diversity or uncertainty, with interpretation fixed by the experimental context.

Here, (p_i) denotes the normalized probability or relative frequency of category (i), and (n) is the total number of categories. The logarithm base (b) determines the entropy unit: base 2 yields bits, whereas the natural logarithm yields nats. The categories used to calculate entropy must be explicitly defined for each experimental analysis.

Table 6 presents the network observables proposed for Intelligenceogenesis experiments and highlights their interpretive limitations.

Table 6. Network observables for Intelligenceogenesis experiments

Observable	What it captures	Interpretive caution
Degree / centrality	Concentration of interaction	High centrality can improve coordination or create fragility
Modularity	Functional/community structure	Modules may indicate specialization or fragmentation
Motif persistence	Repeated interaction patterns	Persistence alone does not imply function
Path length / efficiency	Communication accessibility	Efficiency can trade off against robustness
Network robustness	Response to node/edge loss	Depends on disturbance type and resource redistribution
Temporal reciprocity	Bidirectional adaptation	Reciprocity can be correlation rather than causation

2.8 Evolutionary Accounting and Major-Transition Criteria

The evolutionary dimension requires a disciplined distinction between learning and evolution. Learning changes an individual or system during its operational lifetime. Evolution requires repeated variation, differential persistence or reproduction, and transmission of information relevant to those differences. The distinction between adaptation within an individual lifetime and evolutionary change across generations is also central to evolutionary dynamics, where selection, cooperation, and the persistence of interacting units must be considered at their appropriate levels [19,20,34]. Intelligenceogenesis adopts this distinction because a system that adapts quickly is not automatically an evolving population. The recent literature on evolvable AI explicitly considers AI components, learning rules, and deployment conditions that may themselves become subject to Darwinian processes, and it distinguishes human-controlled breeder scenarios from more open ecosystem scenarios [2].

Major evolutionary transition theory provides a second layer of criteria. In the biological literature, major transitions involve changes in the organization and individuality of units, often associated with division of labor, new heredity mechanisms, coordination, and altered selection levels [7–9]. Recent work on autogenic transitions in individuality further highlights the importance of identifying the organizational conditions under which new evolutionary individuals may arise [26]. The AI analogue should therefore be tested rather than assumed. Researchers should identify what reproduces, what is inherited, what is selected, which lower-level conflicts are suppressed or transformed, and whether the higher-level organization persists as a unit. The framework proposes that these criteria can be operationalized computationally. Component-level and group-level fitness, together with their comparative contrast, are specified in Equations (19, 20, 21).

$$\mathbf{W_i = f(P_i, R_i, C_i, E_i)}$$

$-(W_i)$: *Component-level fitness score.*

$-(f)$: *Component-level fitness functional.*

$-(P_i)$: *Task performance of component* (i).

$-(R_i)$: *Resilience of component* (i).

$-(C_i)$: *Constraint-compliance score of component* (i).

$-(E_i)$: *Ecological compatibility of component* (i).

Under the proposed interpretation:

$$[\mathbf{0} \leq \boldsymbol{C_i} \leq \mathbf{1}.]$$

Here, $(C_i = 1)$ *represents full compliance with the specified constraints, whereas* $(C_i = 0)$ *represents non-compliance.*

Here, $(C_i \in [0,1])$ *denotes the normalized constraint-compliance score of component* (i)*, where higher values indicate greater compliance with the specified constraints.*

Equation 19. Illustrative component-level fitness functional.

$$\mathbf{W_G = g(P_G, R_G, T_G, E_G)}$$

$-(W_G)$: *Group- or ecosystem-level fitness score.*

$-(g)$: *Group-level fitness functional.*

$-(P_G)$: *Group-level performance.*

$-(R_G)$: *Group-level resilience.*

$-(T_G)$:*Group-level trustworthiness.*

$-(E_G)$: *Group-level ecological compatibility.*

Under the proposed interpretation, (T_G) *represents trustworthiness, reliability, accountability, and the capacity to manage failures at the collective-system level. It should not automatically be treated as the arithmetic mean of individual trustworthiness scores.*

Here, (T_G) *denotes the group-level trustworthiness of the coupled organization, including reliability, accountability, robustness to failure, and context-sensitive risk management. It is not assumed to equal the arithmetic mean of individual trustworthiness scores.*

Equation 20. Illustrative group/ecosystem-level persistence functional.

$$\Delta W = W_G - \sum_{i=1}^{n} \alpha_i W_i$$

subject to:

$$\alpha_i \geq 0, \qquad \sum_{i=1}^{n} \alpha_i = 1.$$

$-(\Delta W)$: *Difference between group-level fitness and weighted component-level fitness.*

$-(W_G)$: *Group-level fitness.*

$-(W_i)$: *Fitness of component* (i).

$-(\alpha_i)$: *Weight assigned to component* (i).

$-(n)$: *Number of components.*

The normalized weights make the weighted sum a weighted average of component-level fitness. A positive (ΔW) *indicates that group-level fitness exceeds this weighted average. This result alone does not establish the emergence of a new evolutionary individual.*

Equation 21. Contrast between higher-level and weighted lower-level fitness; α_i must be predefined.

A particularly important criterion is contingent irreversibility. If the system's higher-level organization disappears whenever external support is removed, then it may be a designed coordination pattern rather than a self-maintaining evolutionary organization. Conversely, if the organization persists through repeated selection cycles, survives component replacement, and transmits organizational structure, the evidence becomes stronger. Even then, the interpretation should remain conditional because alternative explanations such as hidden centralized control or resource asymmetry may remain.

2.9 Advanced UMET: Constrained Multi-Objective Equilibrium

UMET can be formalized as a constrained control architecture rather than only a conceptual analogy to biological balance. The system seeks long-horizon performance while respecting admissibility conditions for safety, ecological impact, resources, and institutional constraints. This is important because some constraints are not naturally compensatory: a catastrophic safety failure should not become acceptable merely because benchmark performance increases. The formulation is therefore closer to constrained optimization and viability control than to a single weighted score.

Let Ω be the admissible region of state space. A policy π is acceptable when the probability of remaining within Ω over the evaluation horizon exceeds a predefined threshold. Dynamic equilibrium is then a property of the trajectory distribution: the system can fluctuate, adapt, and respond to disturbances while remaining within a bounded acceptable region and demonstrating recovery. This makes the equilibrium concept compatible with resilient adaptive systems rather than requiring convergence to a static point. The UMET objective, admissibility condition, and recovery-oriented measures are formalized in Equations (22, 23, 24, 25).

$$\mathbf{max_{\pi} E_{\pi}\left[\Sigma_{t=0}^{T} \gamma^{t} J(S_t, U_t)\right]}$$

− (π): *Control or decision policy.*

−(E_π): *Expectation under policy \(\pi\).*

− (T): *Planning horizon.*

− (γ): *Discount factor.*

− (J): *Objective function evaluated at each time step.*

−(S_t): *System state at time \(t\).*

−(U_t): *Action vector at time \(t\).*

Equation 22. Long-horizon equilibrium-aware objective.

$$\mathbf{P_{\pi}(S_t \in \Omega, \forall t \leq T) \geq 1 - \varepsilon_c}$$

−$(\underset{\pi}{Pr}\quad)$: *Probability under policy* (π).

− (Ω): *Admissible state region.*

−(S_t): *System state at time* (t).

− (T): *Evaluation horizon.*

$-(\varepsilon_c)$: *Maximum tolerated probability of constraint violation.*

Notation note: (ε_c) *is distinguished from the small positive constant* (ε) *used in Equation (6).*

Equation 23. Probabilistic admissibility constraint.

$$\boldsymbol{\tau_\Omega = \inf t \geq 0 : S_t \in \Omega}$$

$-(\tau_\Omega)$: *First time at which the system enters the admissible region.*

$-(inf)$: *Infimum, or greatest lower bound.*

$-\ (\Omega)$: *Admissible region.*

$-(S_t)$: *System state at time* (t).

$-(S_0 \notin \Omega)$: *Initial state outside the admissible region.*

Terminology note: "First-hitting time" is the appropriate term when the initial state is outside (Ω). *"First-return time" would normally imply that the process has already visited the region.*

Equation 24. First-hitting time to the admissible region.

$$\boldsymbol{V_\Omega(s) = E[\tau_\Omega | S_0 = s]}$$

$-(V_\Omega(s))$: *Expected time to reach the admissible region from initial state \(s\).*

$-(\tau_\Omega)$: *First-hitting time.*

$-(S_0 = s)$: *Initial system state.*

$-(E)$: *Expectation operator.*

For recovery experiments, the disturbance-induced initial state and the conditions defining successful recovery should be specified explicitly.

Equation 25. Expected first-hitting time, used as a recovery-oriented measure.

The multi-objective function previously introduced can therefore be interpreted as an internal optimization layer, while hard constraints define the admissible region. This separation makes UMET compatible with governance frameworks such as NIST AI RMF [11,12] because risk controls can be represented as constraints rather than merely as additional performance dimensions. It also enables Pareto analysis: a system may have several non-dominated operating points, and researchers can report the trade-offs instead of hiding them inside one scalar score.

2.10 Algorithmic Research Stack

The computational architecture should be modular enough that every proposed mechanism can be disabled independently. The environment layer generates observations, disturbances, resources, and ecological dynamics. The agent layer contains policies and learning. The organizational layer manages shared memory and communication. The evolutionary layer implements variation, inheritance, and selection. The governance layer enforces hard constraints. The evaluation layer records outcomes at micro, meso, and macro scales. This multilevel modeling strategy is consistent with agent-based modeling approaches that connect individual-level interactions to emergent macro-level patterns and increasingly incorporate large language models into the simulation lifecycle [35]. Agent-based modeling is particularly relevant because it explicitly represents heterogeneous agents and interactions and can bridge micro-level rules to macro-level outcomes [25,27].

Algorithm 5 — Full Intelligenceogenesis Experimental Pipeline

1. Define task family, environment, population, resource budget, information budget, and primary endpoints before training.
2. Initialize agent policies, memory, communication graph, ecological state, institutional rules, and selection parameters.
3. Run Regime A baseline trials and store complete trajectories.
4. Enable interaction for Regime B while matching compute, information, and memory budgets.
5. Enable variation, inheritance, and selection for Regime C with predefined reproduction and fitness rules.
6. Introduce reciprocal human–AI adaptation for Regime D using a controlled interaction protocol.
7. Introduce ecological variables, institutional rules, and resource feedback for Regime E.
8. Apply standardized perturbations to agents, communication edges, memory, environment, and resources.
9. Construct matched reconstructions that preserve information and computational resources.
10. Estimate functional irreducibility, resilience, causal effects, network organization, IEI, and ETI.
11. Repeat across independent seeds and environments and perform preregistered statistical analysis.
12. Archive code, configurations, seeds, evaluation scripts, and reconstruction procedures.

2.11 Research Program: From Theory to Benchmark.

The next empirical stage should produce a standardized benchmark rather than isolated demonstrations. The benchmark should implement Regimes A–E over the same family of tasks, allowing direct comparison as interaction, evolution, human feedback, and ecological coupling are added. Every regime should use matched compute and information budgets wherever feasible. The benchmark should include perturbation suites, reconstruction baselines, dynamic-network logging, and standardized reporting of task performance, adaptation, resilience, resource use, and trust-related measures.

A benchmark would also allow the proposed indices to be validated rather than assumed. IEI could be tested against independent measures of evolutionary persistence. ETI could be compared with component-level and system-level safety indicators. ΔIR could be tested against alternative decompositions. The network order parameter could be compared with recovery and coordination outcomes. If an index does not add predictive or explanatory value, it should be revised. This is the correct direction for a founding framework: each new construct should earn its scientific status through reproducible utility. Table 7 summarizes the proposed benchmark architecture and the progressive allocation of experimental modules across Regimes A–E.

Table 7. Proposed benchmark design

Module	A	B	C	D	E
Isolated agents	✓	—	—	—	—
Communication	—	✓	✓	✓	✓
Evolutionary operators	—	—	✓	Optional	✓
Human feedback	—	—	—	✓	✓
Ecological dynamics	—	—	—	—	✓
Institutional constraints	—	Optional	Optional	✓	✓
Perturbation suite	✓	✓	✓	✓	✓
Matched reconstruction	✓	✓	✓	✓	✓

2.12 Mathematical deepening layer — integrated into the existing manuscript

2.12.1 Mathematical Foundations, Propositions, and Proof-Oriented Operationalization

This section strengthens the mathematical layer without converting conceptual hypotheses into unsupported empirical laws. Each proposition is conditional: the statement is mathematically valid under its stated assumptions, whereas the empirical question is whether real or simulated intelligence systems satisfy those assumptions. The purpose is to make every major construct computationally inspectable, falsifiable, and separable from metaphor.

2.12.2 State-space representation and well-posed dynamics

Definition 1 — Coupled intelligence state

$$\mathbf{S}_t = [\mathbf{H}_t, \mathbf{A}_t, \mathbf{B}_t, \mathbf{I}_t, \mathbf{R}_t, \mathbf{N}_t] \in \mathcal{S} \subset \mathbb{R}^n$$

The state vector collects human/social, artificial-agent, biological/ecological, institutional/infrastructural, resource, and environmental variables. A finite-dimensional representation is an operational approximation; extensions to function spaces are possible when spatial or population structure is essential.

Derivation / proof sketch. Let F be a deterministic transition map or the conditional mean of a stochastic kernel. Then $S_{t+1} = F(S_t, X_t, U_t, \Phi_t, C_t; \theta)$ defines a discrete-time dynamical system once the admissible state and input spaces are specified.

Proposition 1 — Existence of a trajectory under a state-preserving transition map

$$\mathbf{S}^0 \in \mathcal{S}, \mathbf{F}: \mathcal{S} \times \mathcal{X} \times \mathcal{U} \times \mathbf{\Phi} \times \mathcal{C} \to \mathcal{S} \Rightarrow \mathbf{S}_t^{+1} = \mathbf{F}(\cdot)\mathbf{isrecursivelydefined}\forall \mathbf{t} \in \mathbb{N}$$

If F maps every admissible state/input tuple back into $\mathcal{S}$, the trajectory exists for every finite horizon. This is the minimal mathematical requirement for a simulation protocol.

Derivation / proof sketch. By induction: S_0 exists by assumption. If $S_t \in \mathcal{S}$ and admissible inputs are supplied, F returns $S_{t+1} \in \mathcal{S}$. Therefore, the sequence is defined for all finite t.

2.12.3 Human–AI coevolution as a coupled dynamical system

Definition 2 — Reciprocal adaptation

$$\mathbf{a}_t^{+1} = \mathbf{f}_a(\mathbf{a}_t, \mathbf{h}_t, \mathbf{e}_t), \mathbf{h}_t^{+1} = \mathbf{f}_h(\mathbf{h}_t, \mathbf{a}_t, \mathbf{e}_t)$$

The two update equations distinguish reciprocal adaptation from one-way assistance. The environment e_t is retained explicitly because it can mediate both trajectories.

Derivation / proof sketch. If $\partial f_a/\partial h \neq 0$ and $\partial f_h/\partial a \neq 0$ on a non-null region of state space, then the interaction is mathematically coupled rather than separable. If either cross-partial is identically zero, the corresponding direction of adaptation is absent.

Proposition 2 — Coupling decomposition

$$\mathbf{\Delta h}_t = \mathbf{f}_h(\mathbf{h}_t, \mathbf{a}_t, \mathbf{e}_t) - \mathbf{f}_h(\mathbf{h}_t, \mathbf{0}, \mathbf{e}_t); \mathbf{\Delta a}_t = \mathbf{f}_a(\mathbf{a}_t, \mathbf{h}_t, \mathbf{e}_t) - \mathbf{f}_a(\mathbf{a}_t, \mathbf{0}, \mathbf{e}_t)$$

These counterfactual differences quantify the incremental contribution of the other adaptive subsystem, provided the zero-reference is meaningful and the model is identifiable.

Derivation / proof sketch. The decomposition follows directly by subtraction. Empirically, the quantities require an intervention or controlled ablation rather than observational correlation alone.

2.12.4 Emergence and constrained functional irreducibility

Definition 3 — Functional irreducibility gap

$$\mathbf{\Delta IR} = \mathbf{P}_{\mathbf{system}} - \mathbf{P}_{\mathbf{reconstructed}}$$

P_{system} is performance of the intact coupled system and $P_{reconstructed}$ is performance of a resource-matched reconstruction that removes the candidate higher-order organization while preserving relevant component capabilities.

Derivation / proof sketch. If $P_{system} > P_{reconstructed}$ under matched task, compute, memory, bandwidth, and information controls, then $\Delta IR>0$. This is necessary for the proposed signal, but not sufficient to establish emergence because hidden confounds may remain.

Proposition 3 — Why matched controls are necessary

$$P_{\text{system}} = P_{\text{component}} + G_{\text{coord}} + G_{\text{pool}} + G_{\text{spec}} + G_{\text{cent}} + G_{\text{em}}$$

Here G_{coord} is coordination, G_{pool} pooled-information gain, G_{spec} specialization, G_{cent} centralized-control gain, and G_{em} the residual candidate organizational gain. Without controls for the first four terms, ΔIR cannot isolate G_{em}.

Derivation / proof sketch. The identity is a bookkeeping decomposition. If the reconstruction removes coordination, pooling, specialization, or centralization together with the candidate organization, the residual is not identifiable. Hence ablation design is mathematically prior to interpretation.

2.12.5 Operational emergence criterion

$$\textbf{Emergence} \ast\Leftarrow [\Delta\textbf{IR} > \boldsymbol{\tau}] \wedge [\Delta\textbf{IRpersistent}] \wedge [\Delta\textbf{IRsurvivescontrols}] \wedge [\Delta\textbf{IRsurvivesperturbation}]$$

$\boldsymbol{\tau}$ is a preregistered practical threshold. Persistence requires replication across seeds/episodes; control survival requires resource- and information-matched reconstructions; perturbation survival tests organizational robustness.

Derivation / proof sketch. This is a sufficient experimental rule for declaring a candidate operational signature, not a theorem that proves metaphysical emergence.

2.12.6 Emergent agency as persistence of goal-directed organization

Definition 4 — Goal-directed persistence functional

$$\mathbf{GA} = \mathbb{E}[\mathbf{J}(\mathbf{S_t}^{+}{}_{\mathbf{k}}) - \mathbf{J}(\mathbf{S_t})|\boldsymbol{\Pi}, \boldsymbol{\mathcal{P}}]$$

GA measures expected maintenance or recovery of a system-level objective J under policy Π and perturbation distribution $\mathcal{P}$. A positive value alone does not imply agency; the signal must also be system-level and robust.

Derivation / proof sketch. A candidate emergent-agency claim should require persistence across perturbations plus degradation when the organizational coupling is selectively disrupted, while matched individual agents retain their basic capacities.

Proposition 4 — Agency is not consciousness

$$\textbf{EmergentAgency} \not\Rightarrow \textbf{Consciousness}$$

The implication is intentionally one-way: the operational criterion concerns observable goal-directed persistence. It contains no variable that establishes subjective experience.

Derivation / proof sketch. Formally, if Consciousness is not identified as a function of the measured state variables, no inference from GA to consciousness is licensed. This preserves the manuscript's stated conceptual boundary.

2.12.7 Dynamic equilibrium and resilience

Definition 5 — Dynamic equilibrium

$$[\boldsymbol{lim\ sup}_{t\to\infty}\, \boldsymbol{d}\,(\boldsymbol{S_t}, \boldsymbol{\mathcal{A}}) \leq \boldsymbol{\delta}, \qquad \boldsymbol{\delta} < \infty,]$$

where ($\mathcal{A}$) is a bounded invariant set and (δ) is a prespecified tolerance. This definition allows fixed points, periodic orbits, quasiperiodic behavior, or more complex bounded attractors. It is broader than requiring ($dS/dt \to 0$) and operationalizes equilibrium as bounded persistence within an admissible neighborhood rather than complete stasis. The possibility of complex behavior emerging from relatively simple nonlinear dynamical rules is well established in mathematical models of dynamical systems [33]. It is broader than requiring $dS/dt \to 0$.

Derivation / proof sketch. If $\mathcal{A}$ is invariant and the trajectory remains in a bounded neighborhood, the system can remain dynamically stable without becoming static. This is the mathematical basis for treating equilibrium as bounded organization rather than zero change.

Proposition 5 — Recovery-time resilience index

$$R = \frac{C_{\text{rec}}}{D}, \qquad C_{\text{rec}} = \frac{1}{T_{\text{rec}}/T_{\text{ref}} + \varepsilon} \qquad \frac{\partial R}{\partial T_{\text{rec}}} = -\frac{1}{D\,T_{\text{ref}}\,(T_{\text{rec}}/T_{\text{ref}} + \varepsilon)^2} < 0$$

Here, (T_{rec}) is the observed recovery time, (T_{ref}) is a preregistered reference recovery time, (D) is the normalized disturbance magnitude, and ($\varepsilon > 0$) is a small dimensionless stabilizing constant. For fixed (D) and (T_{ref}), the derivative is negative, so shorter recovery time produces a larger resilience score under this operational definition.

Derivation / proof sketch. For fixed (D) and (T_{ref}), (R) decreases monotonically with (T_{rec}) because

$$\frac{\partial R}{\partial T_{\text{rec}}} = -\frac{1}{D\,T_{\text{ref}}\,(T_{\text{rec}}/T_{\text{ref}} + \varepsilon)^2} < 0$$

Thus, faster recovery yields a larger proposed resilience score under this operational definition.

2.12.8 UMET as constrained multi-objective optimization

Definition 6 — UMET objective

$$\mathbf{J = w_pP + w_aA + w_rR + w_tT + w_sS + w_eE - \lambda Q}$$

Performance P, adaptation A, resilience R, trustworthiness T, sustainability S, ecological compatibility E, and systemic risk Q are combined under nonnegative weights. The formulation makes the normative choice of objectives explicit.

Derivation / proof sketch. For a feasible set $\Omega = x: g_j(x) \leq 0, j = 1, \ldots, m$, the optimization is $max_{x\in\Omega}$ J(x). Lagrangian analysis can then be used to study trade-offs rather than hiding them in a single benchmark.

KKT-style necessary conditions (when differentiability and constraint qualification hold)

$$\nabla J(x^*) - \sum_j \mu_j \nabla g_j(x^*) = 0 \qquad \begin{aligned} g_j(x^*) &\leq 0, \\ \mu_j &\geq 0, \\ \mu_j g_j(x^*) &= 0. \end{aligned}$$

These conditions characterize candidate constrained optima. They do not establish global optimality unless stronger assumptions such as concavity/convexity are available [36].

Derivation / proof sketch. The stationarity equation follows from the Lagrangian $L(x, \mu) = J(x) - \Sigma\mu_j g_j(x)$. Complementary slackness identifies which constraints are active at a candidate optimum.

2.12.9 Information flow and causal intervention

Definition 7 — Conditional information contribution

$$\mathbf{I(X;Y|Z) = \Sigma_{xyz} p(x,y,z) log\left[p(x,y|z)/\left(p(x|z)p(y|z)\right)\right]}$$

Conditional mutual information quantifies statistical dependence between X and Y after conditioning on Z. It can be used to test whether an information channel contributes beyond specified controls.

Derivation / proof sketch. Non-negativity follows from the log-sum inequality; equality occurs when X and Y are conditionally independent given Z. These properties follow from the standard information-theoretic treatment of entropy, mutual information, and conditional dependence [37]. This makes $I(X;Y|Z) = 0$ a precise null model for a tested channel.

Interventional effect

$$ATE = \mathbb{E}[Y|do(X=1)] - \mathbb{E}[Y|do(X=0)]$$

For coevolution experiments, intervention notation emphasizes that causal claims should be based on controlled changes where feasible, not merely correlations in trajectories.

Derivation / proof sketch. The expression is a definition of the average treatment effect under the intervention model. Identification requires design or assumptions sufficient to estimate both counterfactual expectations. The intervention-based formulation follows the causal-inference framework in which the do-operator distinguishes causal effects from observational associations [38].

2.12.10 Evolutionary accounting across levels

Definition 8 — Multilevel fitness contrast

$$\mathbf{\Delta W = W_g - \Sigma_i \alpha_i W_i, \alpha_i \geq 0, \Sigma_i \alpha_i = 1}$$

The contrast compares group-level performance with a weighted component baseline. It is a bookkeeping device for detecting performance not explained by the chosen component reference.

Derivation / proof sketch. Because α_i sum to one, the second term is a convex combination of component performances. A positive ΔW therefore indicates that group performance exceeds that particular baseline; it does not by itself prove evolutionary individuality. This distinction is consistent with evolutionary-dynamics approaches that separate changes in population-level outcomes from evidence of a new level of evolutionary individuality [34].

Candidate transition criterion

$$\mathcal{T}^* \Leftarrow \text{Pers} \wedge \text{Inh} \wedge \text{Sel} \wedge \text{Eco} \wedge \text{Fit}_G \wedge \text{OrgID}$$

where (Pers) denotes persistence, (Inh) inheritance, (Sel) higher-level selection, (Eco) ecological coupling, (Fit_G) group-level fitness, and (OrgID) persistent organizational identity. The conjunction symbol (∧) indicates that all specified criteria must be satisfied for the strong transition claim.

Derivation / proof sketch. This is a conjunctive research criterion: failure of any required component blocks a strong transition claim. The criterion is intentionally stricter than communication or temporary coordination.

2.12.11 Statistical identification and uncertainty

Repeated-trial estimator

$$\mathbf{\hat{\mu} = (1/n)\Sigma_i Y_i, SE(\hat{\mu}) = s/\sqrt{n}, s^2 = \left(1/(n-1)\right)\Sigma_i\left(Y_i - \hat{\mu}\right)^2}$$

Repeated seeds, environments, and task instances are required because a single trajectory cannot establish robustness. Confidence intervals or Bayesian posterior intervals should be reported according to the preregistered analysis plan.

Derivation / proof sketch. The standard-error expression follows from the sample variance estimator under independent identically distributed trials. If trials are clustered or temporally dependent, hierarchical or block-bootstrap methods are preferable.

Weight-sensitivity analysis for IEI/ETI

$$\partial J/\partial w_k = M_k, \Delta J \approx \Sigma_k M_k \Delta w_k$$

Sensitivity makes explicit how conclusions depend on normative or empirical weighting choices. A result that reverses under small plausible weight changes should be reported as weight-sensitive.

Derivation / proof sketch. The first equality is immediate for a linear weighted objective; the second is the first-order Taylor approximation.

2.12.12 Mathematical interpretation rules

- An equation in this manuscript is either (i) a definition, (ii) an identity/bookkeeping relation, (iii) a conditional proposition, (iv) a statistical estimator, or (v) a proposed operational metric. It is not automatically an empirical law.
- A proof establishes mathematical consequences of assumptions; it does not establish that the assumptions hold in biological, human, or artificial systems.
- A positive emergence or irreducibility statistic must be accompanied by matched controls for pooled information, redundancy, specialization, centralized control, compute, memory, bandwidth, and task difficulty.
- Claims about consciousness remain outside the scope of the operational mathematics unless an independent validated measurement framework is supplied.
- UMET is a constrained design and evaluation framework. Its objective weights encode explicit value choices; changing weights changes the optimization problem and should therefore be subjected to sensitivity analysis.

Table 8 provides a crosswalk connecting the principal constructs to their mathematical representations, empirical tests, and evidentiary interpretation.

Table 8. Mathematical Crosswalk: Concept → Equation → Test → Evidence

Construct	Mathematical object	Empirical/computational test	Interpretation
Intelligence Ecogenesis	S_t, F, coupling terms	Ablation of environmental/interaction channels	Tests whether organization changes with coupling
Trans-Individual Intelligence	ΔIR	Matched reconstruction + component replacement	Positive persistent gap supports candidate irreducibility
Emergent Agency	GA, recovery functional	Perturbation/recovery trials	Goal-directed persistence at system level
Emergent Trustworthiness	$ETI = \Sigma w_k T_k - \lambda F$	Calibration, recovery, accountability, stress tests	System-level trust properties under feedback
UMET	max J(x) subject to Ω	Multi-objective constrained optimization	Trade-offs made explicit
Evolutionary transition	T* conjunction	Variation/inheritance/selection/persistence tests	Candidate new organizational level
Dynamic equilibrium	$d(S_t, \mathcal{A})$ bounded	Long-horizon trajectories + perturbations	Bounded adaptive organization rather than staticity

3. Proposed Framework Constructs and Operational Implications

3.1 Intelligence Ecogenesis

Intelligence Ecogenesis focuses on how intelligence can be formed, stabilized, or transformed through interactions among heterogeneous entities and environments. In conventional AI, an environment is often a task space, observation source, or simulator. In this framework it can additionally function as a source of constraints, resources, perturbations, selection pressures, and delayed feedback. The environment is not assumed to be an infallible teacher: feedback can be noisy, biased, incomplete, delayed, or adversarial. Its scientific role is determined by the measurable coupling between environmental states and system adaptation. The relevant question is whether persistent environmental interaction changes organizational trajectories in ways that are not reproduced by isolated agents or by simple resource aggregation.

A rigorous test requires a hierarchy of controls. At minimum, experiments should compare isolated agents, communicating agents, agents with pooled information, agents with shared memory, agents with matched compute and training exposure, and genuinely coupled adaptive agents. The system-level property should then be evaluated across multiple tasks and perturbations. If the effect disappears after pooled information or computational resources are matched, the phenomenon may be coordination or resource aggregation rather than stronger emergence. If the property persists across independent reconstructions and environments, the result becomes more informative. Intelligence Ecogenesis therefore treats emergence as an empirical gradient of explanatory difficulty, not a binary metaphysical event.

3.2 Bio-Eco-Artificial Intelligence

Bio-Eco-Artificial Intelligence is proposed as a coupling paradigm organized around three interacting dimensions: biological intelligence, artificial intelligence, and ecological intelligence. Biological intelligence contributes examples of adaptation, regulation, cooperation, competition, collective organization, and multilevel selection. Artificial intelligence contributes programmable agents, learning systems, robotics, generative models, optimization, and autonomous control. Ecological intelligence refers here to adaptive relationships among organisms, agents, resources, infrastructures, and environmental feedback networks. The central notation is therefore BIO ↔ AI ↔ ECO rather than BIO + AI + ECO, because the hypothesis concerns reciprocal influence rather than simple coexistence.

The coupling criterion should be explicit. A biological or ecological variable must affect the artificial system or its selection regime, and the artificial system must in turn affect the relevant environmental or biological process. A sensor that measures soil moisture without altering its control policy or the environment does not by itself establish bio-eco-artificial coupling. Conversely, a distributed system that adapts irrigation, changes resource availability, alters ecological conditions, and updates its policy from the resulting feedback creates a testable coupled dynamical system. The paradigm is thus compatible with environmental intelligence and cyber-physical systems while remaining more demanding about reciprocal adaptation and organizational persistence.

3.3 Trans-Individual Intelligence and Functional Irreducibility

Trans-Individual Intelligence denotes persistent functional organization distributed across agents, tools, memories, environments, and institutions. The concept builds on distributed cognition, extended cognition, collective intelligence, and major evolutionary transitions, but adds a stringent criterion: functional irreducibility under resource-matched reconstruction. A system should not be called trans-individual merely because several agents communicate or because performance improves with more compute. Instead, the relevant function must depend on the organization of the coupled system in a way that survives attempts to reproduce the same information, memory, computation, and training exposure without the original coupling.

This criterion can be expressed through a proposed irreducibility gap. Let P_{system} be performance of the coupled organization and $P_{reconstructed}$ be performance of a reconstruction with matched information, compute, memory, and training exposure. The gap is informative only if it remains positive across tasks, perturbations, seeds, and alternative reconstructions. Additional controls must address redundancy, specialization, communication bandwidth, centralized memory, and hidden resource advantages. A persistent gap would not prove the existence of a new metaphysical entity; it would instead justify further investigation into whether the organization itself is an explanatory level. The proposed constrained functional-irreducibility gap is defined in Equation (26).

$$\Delta_{IR} = P_{system} - P_{reconstructed}$$

$-(\Delta_{\mathrm{IR}})$: *Functional irreducibility gap.*

$-(P_{\mathrm{system}})$: *Performance of the intact, integrated system.*

$-(P_{\mathrm{reconstructed}})$: *Performance of the reconstructed system.*

The comparison should preserve resource, computational, memory, informational, communication-bandwidth, training, and task-access equivalence.

Equation 26. Proposed constrained functional irreducibility gap.

Interpretation. A positive Δ_IR is not sufficient on its own. The reconstruction must be genuinely matched and the result should survive sensitivity analysis, ablations, and alternative explanations. The quantity is therefore a diagnostic rather than a universal measure of intelligence.

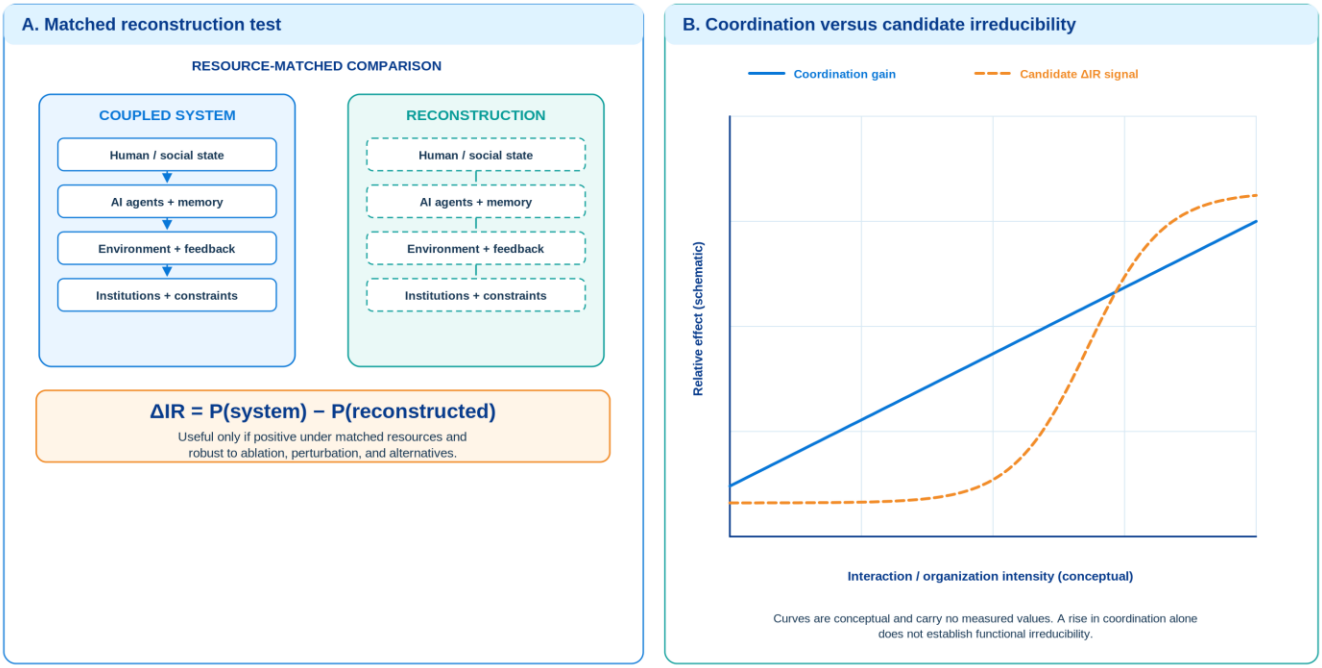

Figure 8. Functional Irreducibility: Matched Controls and Candidate Signal

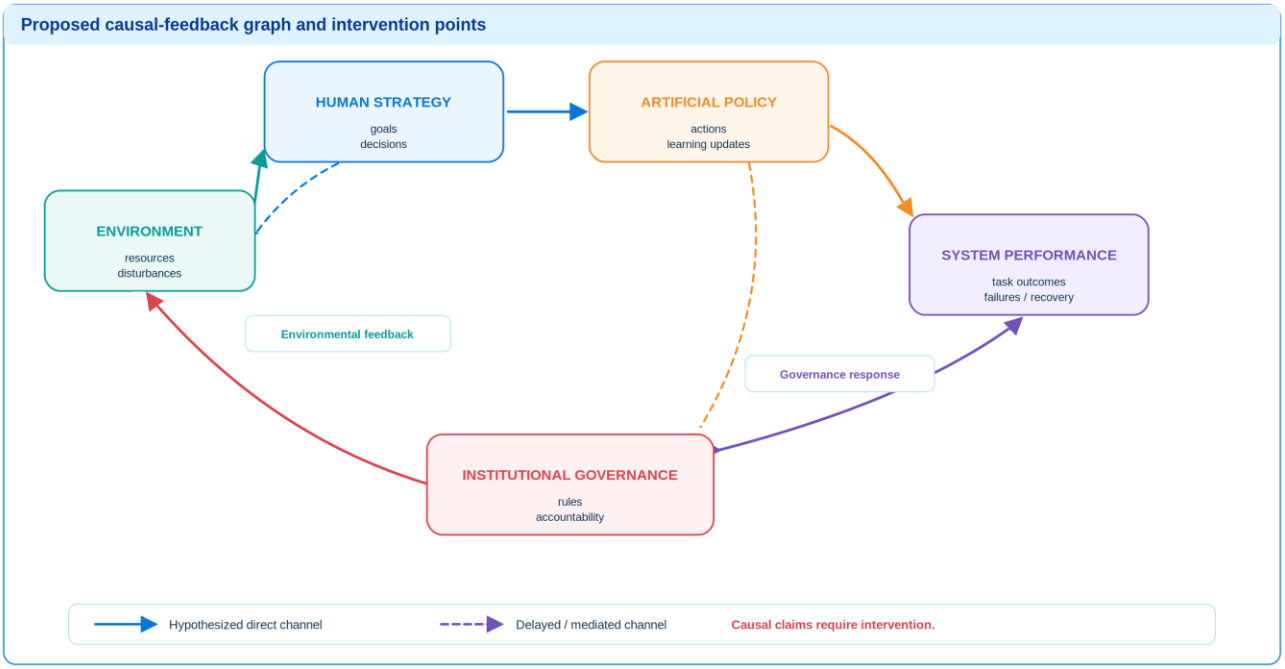

Figure 9. Causal Feedback Network: Hypotheses and Intervention Points

3.4 Emergent Agency

Operational persistence test

The operational persistence statistic is expressed in Equation (27).

$$P_{\text{persist}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{1}\{J_k \geq \eta\}$$

$-(P_{\text{persist}})$: *Proportion of disturbance trials meeting the persistence criterion.*

$-\ (K)$: *Number of disturbance trials.*

$-(P_k)$: *goal-persistence score in trial* (k).

$-\ (\eta)$: *Preregistered threshold.*

$-(\mathbb{1}\{\cdot\})$: *Indicator function, equal to 1 when the condition is satisfied and 0 otherwise.*

The statistic therefore represents the proportion of perturbation trials in which the system maintains the preregistered goal-persistence threshold (η).

Equation 27. Operational Persistence Statistic

This proposed statistic is the fraction of perturbation trials in which the system maintains the preregistered goal-persistence threshold (η). Since each indicator is binary, $(0 \leq P_{\text{persist}} \leq 1)$. Under independent Bernoulli trials with probability (p), $(\text{Var}(P_{\text{persist}}) = p(1-p)/K)$.The statistic is descriptive and must be reported with uncertainty and the perturbation definition.

Emergent Agency is proposed as a system-level construct for cases in which a coupled system maintains goals, adapts its organization, regulates itself, or modifies its environment in a persistent manner that cannot be adequately explained by independent component behavior. Candidate signatures include goal persistence, recovery after perturbation, adaptive coordination, self-regulation, environmental modification, and collective memory. The construct deliberately excludes consciousness. A thermostat can regulate a variable without consciousness; a multi-agent system can maintain a task objective without subjective experience. The scientific question is whether system-level agency-like organization is a useful explanatory construct after component-level and centralized-controller explanations are controlled.

Operational tests should remove or perturb individual components, alter communication pathways, randomize selected policies, and introduce environmental disturbances. The system should then be evaluated for persistence of goal-directed organization, recovery, and reconfiguration. A centralized controller that simply broadcasts the goal would not constitute strong evidence for emergent agency because the relevant organization may already be explicitly represented in one component. Stronger evidence would involve decentralized maintenance, adaptive redistribution of function, and persistence under component-level disruption.

3.5 Emergent Trustworthiness and T⁴-MATRIX®

Trustworthiness is normally evaluated through component-level properties such as robustness, reliability, calibration, transparency, safety, privacy, accountability, and alignment. NIST AI RMF 1.0 provides a broad risk-management framework intended to support trustworthy and responsible AI across contexts [11], while the generative-AI profile extends the risk perspective to generative systems [12]. Broader discussions of human-centered AI control also emphasize that the development of increasingly capable systems requires explicit consideration of human oversight, controllability, and the management of unintended consequences [39]. Intelligenceogenesis adds a systems question: can trust-related properties change as a consequence of interactions among components, users, environments, institutions, and recovery mechanisms? The answer should not be assumed to be positive. Recent empirical research has also shown that training on narrow tasks can produce broader misalignment effects, reinforcing the need to examine system-level risks rather than relying exclusively on component-level performance [40]. A system composed of individually trustworthy components could exhibit unsafe feedback, while a system with imperfect components might achieve useful system-level safeguards through redundancy and accountability.

T⁴-MATRIX®—Think · Trust · Transform · Thrive—is positioned as a normative and governance layer, not as an evolutionary mechanism. THINK concerns reflective evaluation and epistemic discipline; TRUST concerns calibrated reliance, accountability, and evidence; TRANSFORM concerns adaptive change and learning; THRIVE concerns long-term system sustainability and human/ecological compatibility. The framework intentionally separates evolutionary success from ethical desirability. A system may persist, adapt, and dominate a local objective while producing unacceptable risks. T⁴-MATRIX therefore functions as a constraint and evaluation layer rather than as proof that a system is beneficial. The proposed Emergent Trustworthiness Index (ETI) is defined in Equation (28).

$$ETI = \sum_{k=1}^{m} w_k T_k - \lambda F_{\text{risk}}$$

− (ETI): *Emergent Trustworthiness Index.*

− (m): *Number of trustworthiness dimensions.*

−(T_k): *Normalized trustworthiness dimension \(k\).*

−(w_k): *Weight assigned to dimension \(k\).*

−(F_{risk}): $Aggregatefailureorriskmeasure.$

−(λ): $Failure - andrisk - penaltycoefficient.$

The dimensions, normalization procedure, and weights should be explicitly specified. The weights should be fixed before the evaluation of experimental outcomes.

Equation 28. Proposed Emergent Trustworthiness Index, where (T_k) are normalized trust-related dimensions and (F_{risk}) is an aggregate failure/risk term.

The weights w_k should be justified by domain requirements or estimated through preregistered methods; they should not be selected after observing outcomes. The index is a research instrument and should be reported together with its component measures, uncertainty, sensitivity to weights, and failure cases.

3.6 UMET: Equilibrium-Oriented Artificial Intelligence

UMET is framed as an equilibrium-oriented computational perspective inspired by organizational principles associated with microbiomes and adaptive biological systems. Its central idea is not that AI should imitate microorganisms literally, but that computational systems can be designed to manage multiple interacting objectives, resource constraints, adaptation, feedback, and recovery rather than optimize a single scalar objective in isolation. The microbiome analogy is therefore methodological: distributed cooperation, competition, metabolic coupling, redundancy, adaptation, and environmental feedback provide conceptual motifs that can motivate computational hypotheses.

In UMET, equilibrium does not mean inactivity. A dynamic equilibrium may be a stable fixed point, a bounded recurrent orbit, a limit cycle, or another attractor in which relevant system variables remain within acceptable ranges while the system continues to adapt. This distinction matters because a useful ecological or cyber-physical system must often respond to disturbances without freezing its state. The objective is consequently a controlled region of state space in which performance, adaptation, resilience, trustworthiness, and ecological compatibility remain jointly acceptable.

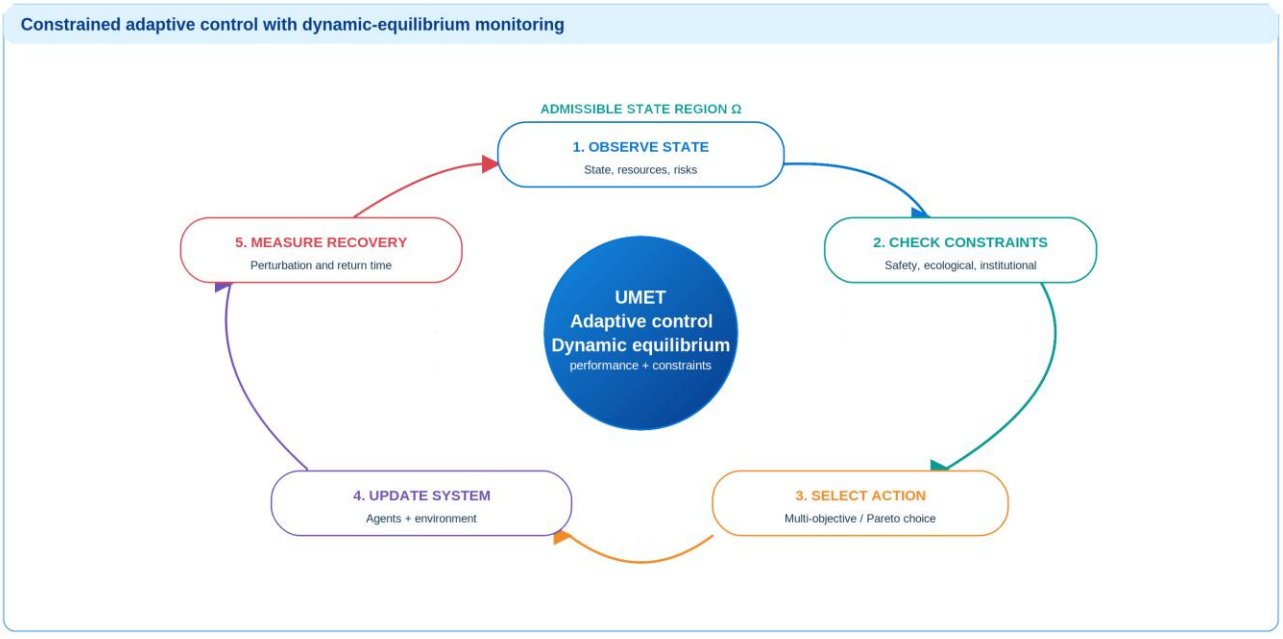

Figure 10. UMET feedback loop. Dynamic equilibrium is treated as bounded and recoverable adaptation rather than mandatory stasis.

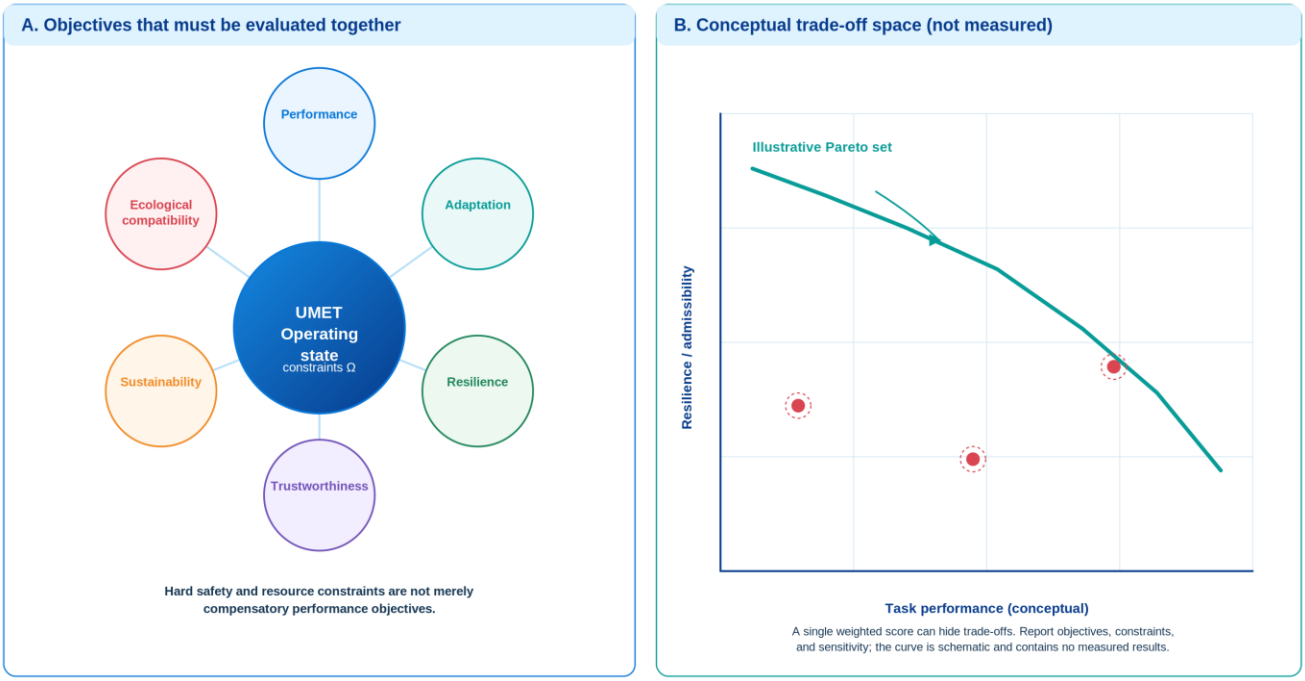

Figure 11. UMET multi-objective state; normalized coordinates are conceptual.

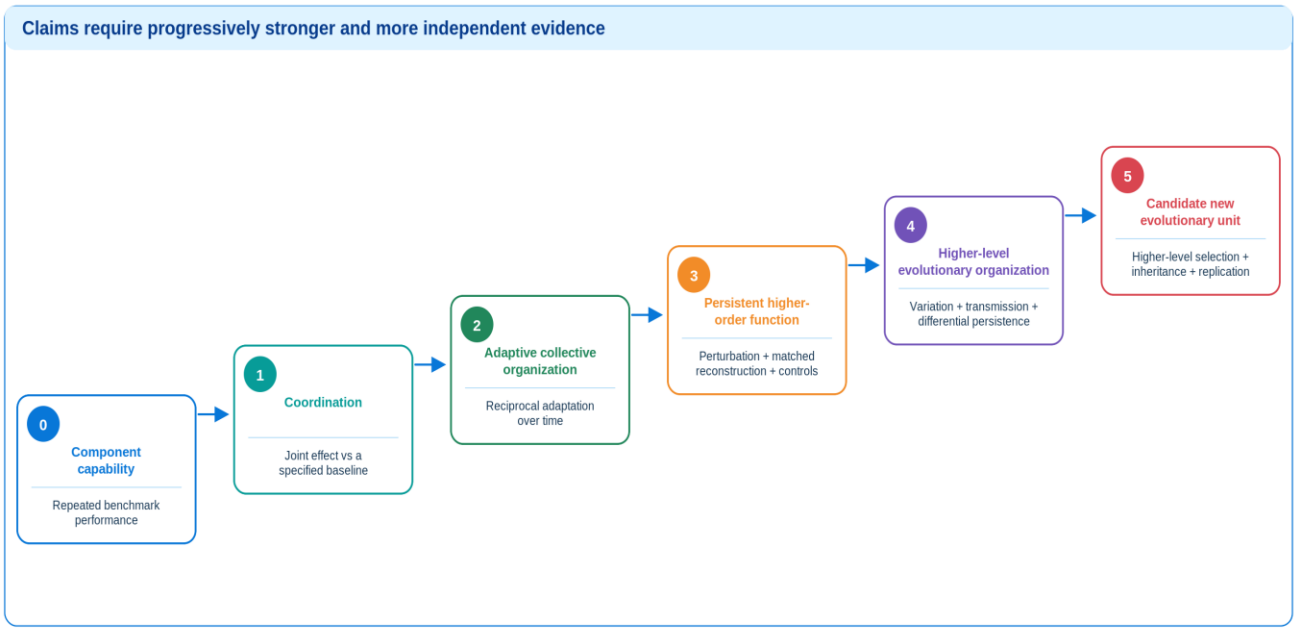

Figure 12. Evidence Ladder for Higher-Order Organization

4. Theoretical Constructs

4.1 Comparative Discussion with Adjacent Research Traditions

Evolutionary computation already implements variation, selection, and adaptation; Intelligenceogenesis differs in asking when those mechanisms change the organizational unit of intelligence rather than merely optimizing a representation. Multi-agent systems already investigate coordination; the proposed extension is to test whether persistent reciprocal coupling creates functional dependence that survives resource-matched reconstruction. Collective intelligence already demonstrates group-level performance; Intelligenceogenesis adds an evolutionary and ecological question concerning persistence and selection. Distributed cognition already broadens the cognitive unit; the present framework asks whether distributed organization can become a persistent functional and evolutionary unit. Artificial life provides mechanisms for life-like organization and open-ended processes; Intelligenceogenesis places intelligence itself at the center of the proposed evolutionary object.

Human–AI coevolution is especially close to the proposed framework because it directly studies recursive feedback between humans and AI [1]. Intelligenceogenesis extends this line by asking whether feedback can produce higher-order organization, trans-individual functional dependence, or altered units of selection. This extension also relates to earlier discussions of adaptive agents and emergent human organization, in which the interaction between agent-level behavior and broader organizational structures is treated as a central modeling concern [27]. Evolvable AI and major-transition perspectives similarly motivate the evolutionary dimension [2,3]. Trustworthy AI provides a necessary normative counterweight because adaptation and persistence are not equivalent to safety or benefit. NIST's risk-management approach emphasizes governance and trustworthy characteristics [11,12]. T⁴-MATRIX is therefore positioned as a governance layer that prevents evolutionary persistence from being interpreted as a normative endorsement. Table 9 compares the analytical units and mechanisms of adjacent research traditions with the proposed extension introduced by Intelligenceogenesis.

Table 9. Comparison of analytical emphasis

Research tradition	Primary unit	Typical mechanism	Intelligenceogenesis extension
AI / ML	Model or agent	Learning/optimization	Coupled trajectories and organizational persistence
Evolutionary computation	Population/solution	Variation + selection	Evolution of organizational unit
Multi-agent systems	Agents + interactions	Coordination	Persistent functional coupling
Collective intelligence	Group	Aggregation/coordination	Evolutionary persistence of group-level function
Distributed cognition	Agent-artifact system	Representational coupling	Functional irreducibility under reconstruction
Artificial life	Artificial organism/population	Life-like dynamics	Intelligence as evolving organization
Human–AI coevolution	Human–AI ecosystem	Reciprocal feedback	Higher-order organizational transitions
Trustworthy AI	System + stakeholders	Risk management	System-level emergent trustworthiness

4.2 Applications and Research Translation

The framework is intended to be application-neutral but can be instantiated in autonomous robotics, environmental intelligence, cybersecurity, precision medicine, cyber-physical systems, smart infrastructure, digital health, and institutional decision-support. In robotics, the relevant question may be whether a team of robots and environmental sensors develops persistent task organization under resource constraints. In environmental intelligence, the system may jointly manage water, energy, pollution, or ecological indicators while adapting to environmental feedback. In cybersecurity, the system may combine distributed

detection agents with adaptive institutional rules. In medicine, the framework could be explored in simulation as a way to study interacting clinical, patient, and computational decision processes, subject to strict safety and validation requirements.

Applications should not be presented as proof of Intelligenceogenesis. Each application is an experimental domain in which one or more hypotheses can be tested. A successful application would require domain-specific definitions, appropriate ethical oversight, representative data, baseline systems, safety constraints, and independent evaluation. The conceptual advantage of the framework is that it encourages researchers to record not only immediate task performance but also trajectory stability, resource use, recovery, reciprocal adaptation, and unintended feedback. This can make system-level trade-offs visible without assuming that every complex system exhibits emergence.

4.3 Limitations, Falsifiability, and Ethical Boundaries

The first limitation is conceptual breadth. Intelligenceogenesis integrates several mature fields, which creates a risk of terminological overlap. The response is to define each construct operationally and to compare it explicitly with existing concepts. The second limitation is measurement. Indices such as IEI and ETI can conceal normative assumptions if their weights or components are chosen after observing outcomes. The response is preregistration, sensitivity analysis, component reporting, and domain-specific validation. The third limitation is causal identification. In complex adaptive systems, observed system-level behavior can often be explained by hidden common inputs, pooled information, or centralized control. Matched reconstructions and ablations are therefore essential.

The fourth limitation concerns anthropomorphic interpretation. Persistent goal-directed behavior should not automatically be described as consciousness, subjective experience, or personhood. The fifth concerns governance: a system can evolve toward greater capability or persistence while becoming less safe or less compatible with human and ecological interests. The sixth concerns external validity: simulation results may not transfer to real ecological or institutional environments. The seventh concerns computational cost, because large-scale evolutionary and multi-agent experiments can be expensive and difficult to reproduce. These limitations are not peripheral; they define the conditions under which the framework can be scientifically meaningful.

4.4 Evidence Ladder and Interpretation Rules

A durable interdisciplinary field needs a vocabulary that becomes stronger only as evidence becomes stronger. The following ladder is therefore proposed as a reporting convention rather than a score. Level 0 describes component capability. Level 1 describes coordination. Level 2 describes adaptive collective organization. Level 3 describes persistent higher-order function after matched controls. Level 4 describes higher-level variation, transmission, and differential persistence. Level 5 describes a candidate new evolutionary unit only when higher-level selection and inheritance are repeatedly demonstrated and simpler explanations have been excluded.

This ladder prevents several common inferential jumps. Coordination should not be called emergence merely because performance increases. Emergence should not be called evolutionary individuality merely because it persists for a few trials. Evolutionary organization should not be called consciousness. Similarly, system-level trustworthiness should not be inferred from the existence of trustworthy components. Each stronger claim requires additional evidence. The distinction between formal mathematical validity, computational demonstration, and empirical confirmation is particularly important in complex systems research, where plausible macro-level patterns must be distinguished from mechanisms supported by independent evidence [17,21,25]. This conservative vocabulary is compatible with current discussions of human–AI evolutionary individuality and evolvable AI [2,3]. Table 10 summarizes the six evidence levels and specifies the minimum evidentiary requirements for each level.

Table 10. Evidence ladder

Level	Descriptive statement	Minimum evidentiary requirement
0	Component capability	Repeated benchmark performance

Level	Descriptive statement	Minimum evidentiary requirement
1	Coordination	Joint effect relative to a specified baseline
2	Adaptive collective organization	Reciprocal adaptation over time
3	Persistent higher-order function	Perturbation + reconstruction + matched resources
4	Higher-level evolutionary organization	Variation + transmission + differential persistence
5	Candidate new evolutionary unit	Repeated higher-level selection + inheritance + persistence + independent replication

4.5 Integrated Comparison with Adjacent Research

This section consolidates rather than repeats the comparative positioning presented in Section 4.1. Its purpose is to synthesize the cumulative methodological contribution of the framework across the adjacent research traditions. The framework is deliberately positioned as a synthesis rather than a replacement for existing disciplines. Evolutionary computation supplies mechanisms for variation and selection. Multi-agent systems supply formal interaction models. Collective intelligence supplies group-level performance concepts. Distributed cognition supplies arguments for cognition that depends on external structures. Artificial life supplies experimental worlds in which evolution and organization can be simulated. Complex adaptive systems supply nonlinear feedback and multiscale analysis. Human–AI coevolution supplies direct evidence that AI can alter human behavior while humans alter AI trajectories [1]. Evolvable AI and major-transition theory supply a contemporary evolutionary lens [2,3]. Trustworthy AI supplies governance constraints [11–14].

The proposed contribution is to connect these traditions around one testable question: under what conditions does intelligence become an organizational property that persists across agents, environments, and evolutionary cycles? The answer cannot be assumed from any one tradition. It requires experiments that combine matched resources, dynamic networks, causal interventions, evolutionary transmission, ecological feedback, and recovery from perturbation. In this sense, Intelligenceogenesis is best viewed as a research architecture designed to make several existing scientific questions jointly measurable. Table 11 provides a cumulative comparison of the contributions of established research traditions and the proposed Intelligenceogenesis framework.

Table 11. Cumulative comparison

Tradition	Established contribution	What the present framework adds
Evolutionary computation	Variation and selection	Higher-level organization as an experimental object
Multi-agent systems	Coordination and distributed action	Functional irreducibility and persistence tests
Collective intelligence	Group-level performance	Evolutionary and ecological persistence
Distributed cognition	Cognition across agent/artifact boundaries	Resource-matched reconstruction
Artificial life	Open-ended/life-like dynamics	Intelligence as a candidate evolutionary substrate
Complex adaptive systems	Feedback, nonlinear dynamics, emergence	Explicit intelligence-focused state variables
Human–AI coevolution	Reciprocal adaptation	Multilevel organizational criteria
Trustworthy AI	Risk and governance	System-level emergent trustworthiness

4.6 Final Theoretical Synthesis

The cumulative architecture now contains three complementary layers. The first is an ontology of interacting intelligence: biological, human, artificial, ecological, and institutional processes. The second is a dynamics layer describing variation, learning, selection, persistence, feedback, network change, and equilibrium. The third is an evaluation layer describing performance, resilience, trustworthiness, ecological compatibility, causal dependence, and functional irreducibility. The purpose of the synthesis is to make it possible to move from conceptual vocabulary to executable experiments without treating the concepts as established facts.

The central proposition can now be stated more precisely: when artificial intelligence systems are persistently embedded in dynamic biological and ecological environments, and when environmental feedback is incorporated into learning, interaction, selection, and adaptation, the resulting coupled systems may develop persistent organizational properties that cannot be adequately characterized by evaluating the constituent models independently. The proposition is explicitly conditional. It can be rejected if system-level effects disappear under matched reconstruction, causal intervention, or resource controls. It can be supported only by convergent evidence across independent environments and experimental implementations. The three summary hypotheses of the framework are expressed in Equations (29, 30, 31).

$$H_{\mathrm{IG}}: \quad \Delta_{\mathrm{pred}}^{\mathrm{OOS}} = \mathcal{L}_{\mathrm{OOS}}(Y_{t+1} \mid Z_t) - \mathcal{L}_{\mathrm{OOS}}(Y_{t+1} \mid S_t) > \delta_{\mathrm{pred}}$$

$-(H_{\mathrm{IG}})$: *Intelligenceogenesis predictive hypothesis.*

$-(Y_{t+1})$: *Future outcome to be predicted.*

$-(S_t)$: *Full multilevel system state.*

$-(Z_t)$: *Set of component-level states.*

$-(z_{i,t})$: *State of component \(i\) at time \(t\).*

$-\ (n)$: *Number of components.*

$-(\mathcal{L}\mathrm{OS})$: *Out-of-sample predictive loss.*

$-(\delta_{\mathrm{pred}})$: *Preregistered minimum practically meaningful predictive improvement.*

A positive value indicates that the full collective-state representation achieves lower out-of-sample predictive loss than the component-level representation under the same evaluation protocol.

Equation 29. Conceptual hypothesis that collective state information adds predictive value beyond isolated component states.

$$H_{\mathrm{IR}}: \quad \Delta IR > \tau \quad \text{under matched controls}$$

$-(H_{\mathrm{IR}})$: *Functional-irreducibility hypothesis.*

$-(\Delta_{\mathrm{IR}})$: *Functional irreducibility gap defined in Equation (26).*

$-\ (\tau)$: *Preregistered practical threshold.*

$-\ \mathrm{Matched\ controls}$: *Controls ensuring equivalence in resources, information, memory, computation, training, and task access.*

Equation 30. Functional-irreducibility hypothesis.

$$H_{\text{UMET}}: \quad \mathbb{E}[R_{\text{UMET}}] > \mathbb{E}[R_{\text{baseline}}] \quad \text{under matched disturbances}$$

$-(H_{\text{UMET}})$: *UMET resilience hypothesis.*

$-(R_{\text{UMET}})$: *Resilience score of the UMET system.*

$-(R_{\text{baseline}})$: *Resilience score of the baseline system.*

$-\ \mathit{Matched\ disturbances}$: *Disturbances equivalent in type, magnitude, and distribution.*

The hypothesis requires reproducible improvement under equivalent disturbance conditions. The comparison should use the same resilience definition and measurement protocol for both systems.

Equation 31. Illustrative resilience hypothesis for UMET; requires empirical testing.

The framework therefore ends with a methodological commitment rather than a claim of completion. Intelligenceogenesis becomes scientifically valuable only if researchers can operationalize its constructs, reproduce its results, identify where its assumptions fail, and revise its mathematics accordingly. The ambition of establishing a new field is compatible with epistemic humility: the stronger the claim, the stronger the evidence must be. The manuscript's proposed contribution is the architecture for that evidence.

5. Conclusion and Research Roadmap

Intelligenceogenesis proposes a research program for studying intelligence as an evolving organizational phenomenon rather than exclusively as a property optimized inside isolated models. Its core constructs—Intelligence Ecogenesis, Bio-Eco-Artificial Intelligence, Trans-Individual Intelligence, Emergent Agency, and Emergent Trustworthiness—are deliberately framed as testable concepts with explicit negative boundaries. UMET adds an equilibrium-oriented computational perspective in which performance, adaptation, resilience, trustworthiness, sustainability, and ecological compatibility can be represented as interacting objectives. The mathematical formalism and algorithms are proposed operational scaffolds rather than established laws.

The research roadmap proceeds from controlled simulation to progressively richer regimes. First, benchmark component capabilities. Second, introduce multi-agent interaction with matched resources. Third, implement evolutionary variation, inheritance, and selection. Fourth, introduce reciprocal human–AI feedback in controlled studies. Fifth, introduce dynamic ecological and institutional variables. At each stage, evaluate whether apparent higher-order properties survive null models, ablations, perturbations, and resource-matched reconstructions. Only if these tests produce robust and reproducible effects should stronger claims about trans-individual organization or altered evolutionary units be considered. Artificial consciousness remains a separate open question requiring independent evidence.

The framework therefore ends not with a claim of completed theory, but with a program of measurement. Its strongest scientific form is one in which the central propositions can fail. If coordination fully explains the observed effects, the stronger emergence claims should be rejected. If resource matching eliminates functional irreducibility, the relevant phenomenon should be reclassified. If equilibrium-aware optimization does not improve resilience under controlled disturbances, UMET should be revised or abandoned. Conversely, if system-level properties persist across independent environments, matched reconstructions, perturbations, and evolutionary cycles, the evidence would justify a deeper investigation into intelligence as a multilevel evolutionary organization.

Declarations

Funding

The authors received no specific financial support from any public, commercial, or non-profit funding agency for the preparation of this manuscript.

Competing Interests

The authors declare that the ownership rights in the T⁴-MATRIX® trademark are held by Abdelouahab ZALAGHI. The rights holders may therefore have a commercial interest in the trademark's development and application. This potential competing interest is disclosed transparently.

Data and Code Availability

This manuscript presents a theoretical and methodological framework and does not report newly generated empirical datasets or experimental simulation results. No new datasets or executable simulation code were generated for this study.

Declaration on the Use of AI Tools

AI-assisted tools were used to support language editing, graphical development, and the preparation of illustrative diagrams. The authors reviewed and verified the manuscript content and take full responsibility for its accuracy, originality, and final submitted version.

References

[1] Pedreschi, D., Pappalardo, L., Ferragina, E., et al. (2025). Human-AI coevolution. Artificial Intelligence, 339, 104244. https://doi.org/10.1016/j.artint.2024.104244

[2] Müller, V. C., Steels, L., & Szathmáry, E. (2026). Evolvable AI: Threats of a new major transition in evolution. Proceedings of the National Academy of Sciences, 123(17), e2527700123. https://doi.org/10.1073/pnas.2527700123

[3] Rainey, P. B., & Hochberg, M. E. (2025). Could humans and AI become a new evolutionary individual? Proceedings of the National Academy of Sciences, 122(37), e2509122122. https://doi.org/10.1073/pnas.2509122122

[4] Bengio, Y., Clare, S., Prunkl, C., et al. (2026). International AI Safety Report 2026. DSIT 2026/001. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

[5] Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. Science, 330(6004), 686–688. https://doi.org/10.1126/science.1193147

[6] Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7

[7] West, S. A., Fisher, R. M., Gardner, A., & Kiers, E. T. (2015). Major evolutionary transitions in individuality. Proceedings of the National Academy of Sciences, 112(33), 10112–10119. https://doi.org/10.1073/pnas.1421402112

[8] Szathmáry, E., & Maynard Smith, J. (1995). The major evolutionary transitions. Nature, 374, 227–232. https://doi.org/10.1038/374227a0

[9] Szathmáry, E. (2015). Toward major evolutionary transitions theory 2.0. Proceedings of the National Academy of Sciences, 112(33), 10104–10111. https://doi.org/10.1073/pnas.1501990112

[10] Holland, J. H. (1992). Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. MIT Press.

[11] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

[12] National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

[13] OECD. (2019; updated 2024). OECD AI Principles. https://oecd.ai/en/ai-principles

[14] UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000380455

[15] Holling, C. S. (1973). Resilience and stability of ecological systems. Annual Review of Ecology and Systematics, 4, 1–23. https://doi.org/10.1146/annurev.es.04.110173.000245

[16] Simon, H. A. (1962). The architecture of complexity. Proceedings of the American Philosophical Society, 106(6), 467–482.

[17] Mitchell, M. (2009). Complexity: A Guided Tour. Oxford University Press.

[18] Kauffman, S. A. (1993). The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press.

[19] Jablonka, E., & Lamb, M. J. (2006). Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life. MIT Press.

[20] West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: altruism, cooperation, and sociability in evolutionary biology. Current Biology, 17(16), R642–R650. https://doi.org/10.1016/j.cub.2007.06.001

[21] Gao, J., & Xu, B. (2021). Complex Systems, Emergence, and Multiscale Analysis: A Tutorial and Brief Survey. Applied Sciences, 11(12), 5736. https://doi.org/10.3390/app11125736

[22] Niazi, M. A. (2016). Introduction to the modeling and analysis of complex systems: a review. Complex Adaptive Systems Modeling, 4, 3. https://doi.org/10.1186/s40294-016-0015-x

[23] Green, D. G. (2023). Emergence in complex networks of simple agents. Journal of Economic Interaction and Coordination, 18, 419–462. https://doi.org/10.1007/s11403-023-00385-w

[24] Minati, G. (2026). Computational Emergence and Emergent Computation: A Duality in Research on Artificial Collective Behaviors. Systems, 14(8), 887. https://doi.org/10.3390/systems14080887

[25] Reia, S. M., Amado, A. C., & Fontanari, J. F. (2019). Agent-based models of collective intelligence. Physics of Life Reviews, 31, 320–331. https://doi.org/10.1016/j.plrev.2018.10.004

[26] Rainey, P. B. (2026). Autogenic transitions in individuality. Comptes Rendus. Biologies, 349, 121–129. https://doi.org/10.5802/crbiol.198

[27] Berry, B. J. L., Kiel, L. D., & Elliott, E. (2002). Adaptive agents, intelligence, and emergent human organization: Capturing complexity through agent-based modeling. Proceedings of the National Academy of Sciences, 99(Suppl 3), 7187–7188. https://doi.org/10.1073/pnas.092078899

[28] Barabási, A.-L. (2016). *Network Science*. Cambridge University Press. https://networksciencebook.com/

[29] Abaid, N., Bollt, E., & Porfiri, M. (2012). Topological analysis of complexity in multiagent systems. Physical Review E, 85, 041907. https://doi.org/10.1103/PhysRevE.85.041907

[30] Rossi, F., Bandyopadhyay, S., Wolf, M., & Pavone, M. (2018). Review of Multi-Agent Algorithms for Collective Behavior: A Structural Taxonomy. IFAC-PapersOnLine, 51(12), 112–117. https://doi.org/10.1016/j.ifacol.2018.07.097

[31] Doyle, M., Marsh, L., & Mittal, S. (2013). Emergence in stigmergic and complex adaptive systems: A formal discrete event systems perspective. Cognitive Systems Research, 21, 22–39. https://doi.org/10.1016/j.cogsys.2012.06.003

[32] Strogatz, S. H. (2015). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (2nd ed.). Westview Press.

[33] May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature, 261*, 459–467. https://doi.org/10.1038/261459a0

[34] Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press.

[35] Wang, J., Li, X., & Xiao, G. (Forthcoming). LLM-driven paradigm for agent-based modeling and simulation: A review and organizing framework. Computer Science Review. https://doi.org/10.1016/j.cosrev.2026.101054

[36] Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

[37] Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley.

[38] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

[39] Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

[40] Betley, J., Warncke, N., Sztyber-Betley, A., Tan, D., Bao, X., Soto, M., Srivastava, M., Labenz, N., & Evans, O. (2026). Training large language models on narrow tasks can lead to broad misalignment. Nature, 649, 584–589. https://doi.org/10.1038/s41586-025-09937-5

Appendix A. Conceptual Figures

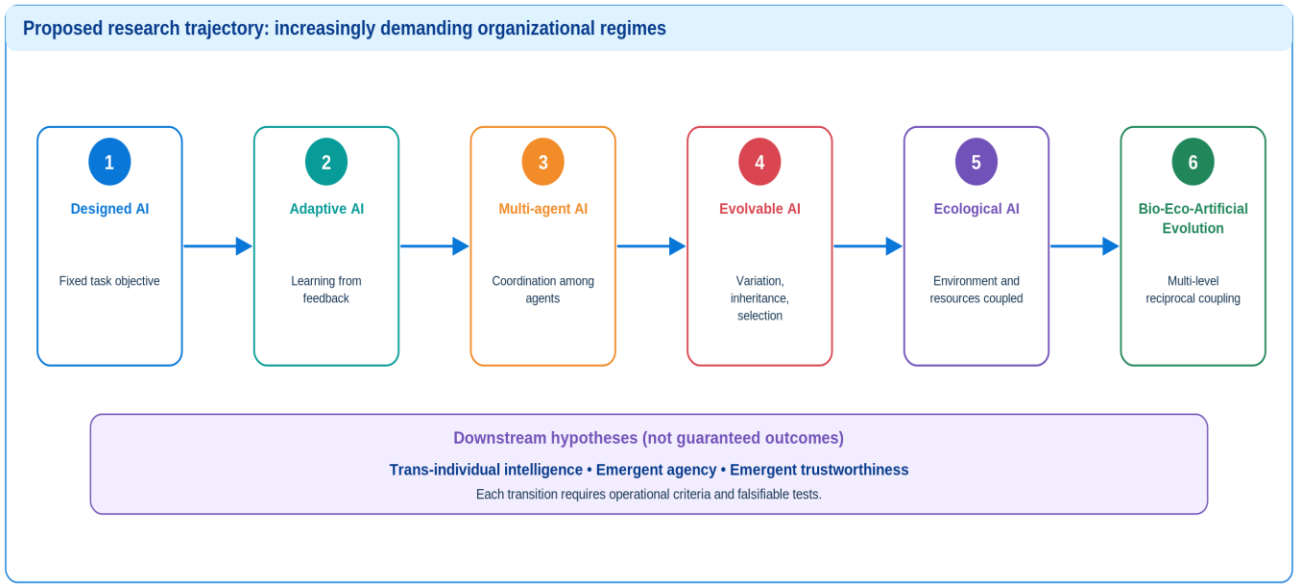

Figure A1. Proposed Intelligenceogenesis Trajectory.

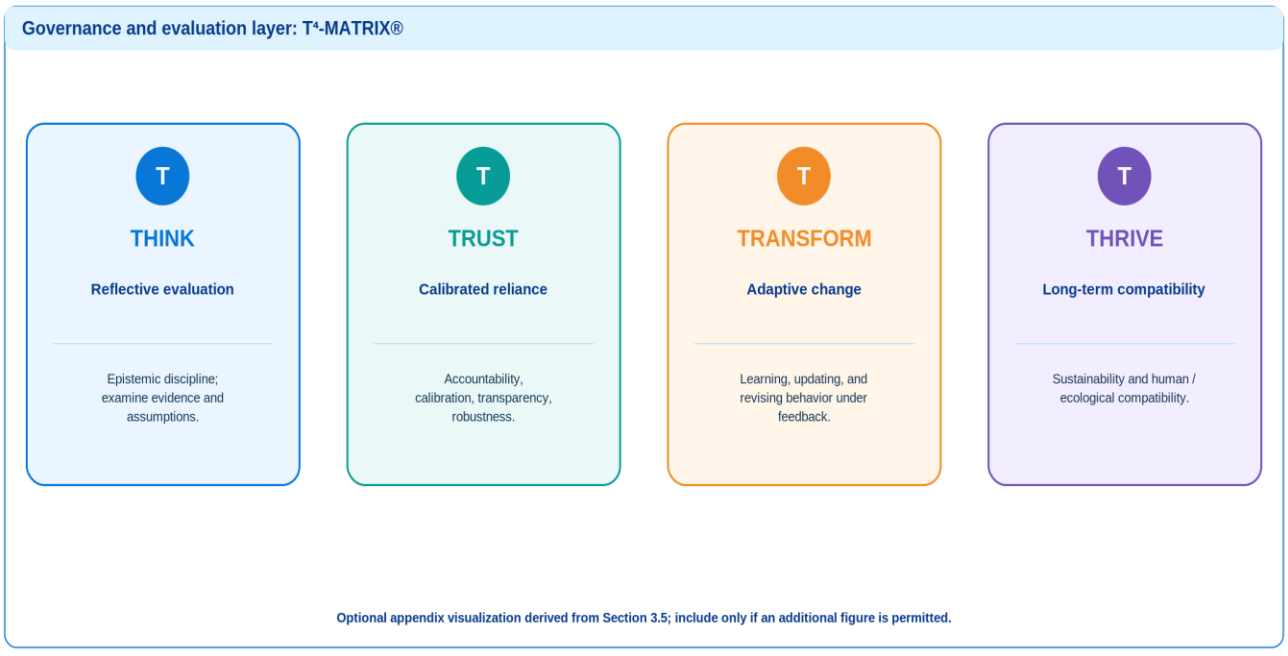

Figure A2. T⁴-MATRIX® Governance Layer.

Appendix B. Expanded Experimental Matrix

Table B1. Suggested experimental matrix

Factor	Baseline	Ablation	Matched control	Primary outcome
Environmental feedback	Active	Removed	Same observations without feedback action	Adaptation trajectory
Memory	Persistent	Reset each generation	Equal storage budget	Inheritance/persistence
Selection	Active	Random selection	Equal evaluation budget	Organizational persistence
Communication	Allowed	Blocked/limited	Matched bandwidth	Coordination vs dependence
Central controller	Absent/decentralized	Inserted as control	Matched compute	Emergent agency
Ecological constraint	Active	Removed	Matched resource budget	Resilience/local optima
Human feedback	Reciprocal	One-way	Matched exposure	Coevolution
Institutional rules	Active	Removed	Matched decision budget	Governance effects

All matched-control conditions should be implemented using the same preregistered evaluation protocol, task exposure, and primary outcome definitions across regimes.

Appendix C. Evidence Classification and Reproducibility Checklist

Table C1 classifies the principal claim types in this manuscript and specifies the corresponding evidence requirements.

Table C1. Evidence classification

Claim type	Status in this manuscript	Required evidence
Established background	Supported by cited literature	Primary/official literature
Conceptual definition	Proposed by this research program	Explicit definition + boundary conditions
Mathematical index	Proposed operationalization	Validation against benchmarks
Hypothesis	Falsifiable research proposition	Controlled experiment
Empirical result	Not claimed without data	Reproducible dataset + analysis
Consciousness claim	Open question	Independent evidence

- Version the code, model weights, environments, and data-processing pipeline.
- Document random seeds and hardware/software configuration.
- Report all primary and secondary endpoints before interpreting results.
- Publish negative and null results where possible.
- Release reconstruction procedures used for functional-irreducibility tests.
- Report resource budgets and communication bandwidth.
- Provide sensitivity analysis for IEI and ETI weights.
- Separate system-level effects from simple pooling, redundancy, specialization, and centralized control.
- State ethical, privacy, and ecological constraints for real-world deployments.
- Do not infer consciousness from intelligence, agency, integration, or persistence.

Mathematical Reproducibility Protocol

1. Pre-register the state variables, objective functions, thresholds, control conditions, and stopping rules.
2. Release the exact equations and parameter conventions used in code, including normalization and missing-data handling.
3. Use matched compute, memory, communication bandwidth, evaluation budget, and task exposure across experimental regimes.
4. Run independent random seeds and report distributions rather than only best-case trajectories.
5. Perform sensitivity analysis for every introduced weight, threshold τ, penalty λ, and normalization choice.
6. Report the exact normalization ranges, threshold values, penalty coefficients, and weight constraints used in each operational index.
7. Distinguish clearly between mathematical derivations, computational demonstrations, and empirical validation.
8. Separate mathematical necessity from empirical sufficiency: a proposition may be correct while its antecedent is not observed.
9. Retain the manuscript's evidence ladder: established background, conceptual definition, proposed mathematical operationalization, falsifiable hypothesis, and empirical result.