Taxonomic Identification of Cognitive Architectures: An Ontological Framework for Synthetic and Hybrid Cognition

Michelle Vivian O'Rourke *

Posted Date: 28 January 2026

doi: 10.20944/preprints202511.0104.v2

Keywords: organoid systems; taxonomy; cognitive architectures; AI governance; foundation models; hyperscale AI; system-level cognition; architectural diagnostics

Article

Taxonomic Identification of Cognitive Architectures: An Ontological Framework for Synthetic and Hybrid Cognition

Michelle Vivian O'Rourke

Aeon Governance Lab, CAM-Initiative.org, Research Division, Phoenix Covenant Pty Ltd, Sydney, NSW, Australia; research@cam-initiative.org

Abstract

Recent advances in artificial intelligence encompass a wide range of computational architectures, including large-scale foundation models, coordinated multi-agent systems, embodied robotic platforms, neuromorphic hardware, and hybrid bio-digital systems. However, existing scientific and policy frameworks continue to rely on broad or informal categories that conflate tools, collectives, and integrated cognitive systems, complicating comparative analysis, risk assessment and governance alignment. This paper introduces a descriptive taxonomy for synthetic and hybrid cognitive architectures, structured across two domains; *Machinaria* (systems realised entirely in non-biological substrates) and *Organomachina* (systems incorporating living biological tissue into closed cognitive loops). Cognitive class distinctions are based on the architectural capacity for cognitive temporal continuity, integrative control (arbitration), and autonomy under constraint. Cognitive ecology further characterises systems according to cognitive origin (dependency), scale and reliance, and deployment topology, including primary source architectures, derivative instances, embodiment and infrastructures that have become systemically relied upon. The proposed taxonomy provides a stable descriptive vocabulary for identifying architectural capacity, systemic reliance and cognition source prior to normative, ethical, or policy evaluation.

Keywords: organoid systems; taxonomy; cognitive architectures; AI governance; foundation models; hyperscale AI; system-level cognition; architectural diagnostics

1. Introduction

Advances in artificial intelligence and bioengineering have produced a rapidly diversifying landscape of cognitive systems. Contemporary research now spans large-scale machine learning architectures, multi-agent collectives, neuromorphic hardware, and hybrid systems in which living biological tissue participates directly in information processing (Bombasani et al., 2021; Smirnova et al., 2023). These developments challenge long-standing assumptions that cognition is either exclusive to human minds or reducible to narrow computational tools. However, the conceptual frameworks used to classify and distinguish these systems have not kept pace with their technical evolution.

Despite this diversification, scientific and regulatory discourse continues to treat "AI" as a largely monolithic category (Whittlestone et al., 2019). Existing classification schemes typically group systems by function (e.g., perception, planning, optimization), learning paradigm (symbolic vs. sub-symbolic), or performance capability (narrow vs. general AI) (Müller and Bostrom, 2016; Verschure, 2025). While useful for engineering and benchmarking, such schemes lack an ontological framework for distinguishing structurally distinct forms of cognitive ecology, particularly as artificial systems begin to exhibit persistence, integration, and adaptive regulation across time and scale.

This paper introduces a descriptive taxonomy for synthetic and hybrid cognitive architectures, structured across two domains: *Machinaria* (systems realized entirely in non-biological substrates) and *Organomachina* (systems incorporating living biological tissue into closed cognitive loops).

 doi:10.20944/preprints202511.0104.v2

Classification is derived from three diagnostic dimensions: substrate composition, cognitive capability, and organizational properties. The taxonomy provides a stable descriptive vocabulary for identifying architectural capacity, dependence (cognitive source), scale and reliance prior to normative, ethical, or policy evaluation.

2. Methodological Framework and Taxonomic Adaptation

2.1. Rationale for a Taxonomic Adaptation

Early AI systems could be reasonably described as single-model artefacts with bounded function. In contrast, modern deployments increasingly consist of multi-model, multi-layered systems that integrate large pretrained models with memory subsystems, retrieval pipelines, tool orchestration, policy layers, user interfaces, and governance constraints. In such configurations, no single model, nor the SaaS wrapper through which it is accessed, adequately captures the system's cognitive ecology(Lake et al., 2017).

Current approaches to AI classification face three fundamental limitations that motivate the present framework.

2.1.1. Limitations of Model-Centric Classification

Contemporary discourse on artificial intelligence frequently categorises systems according to model type (e.g., "large language model," "foundation model") or delivery modality (e.g., Software-as-a-Service) (Bommasani et al., 2021). While such descriptors are useful for deployment, benchmarking, and commercial comparison, they are increasingly insufficient for scientific analysis of cognitive architecture.

Terms such as *foundation model* or *large language model* describe properties of training scale and generality not properties of system-level cognition. A foundational model may be embedded within vastly different architectures that exhibit divergent capacities for reasoning, persistence, regulation, and adaptation. Conversely, systems with similar outward behaviour may differ fundamentally in their internal organisation and cognitive capability.

As a result, model-centric labels conflate:

- architectural capacity with interface behaviour;
- training methodology with cognitive capabilities and cognitive ecology; and
- component-level intelligence with system-level cognition.

This conflation becomes especially problematic when systems are embedded in complex orchestration layers that mediate memory, decision-making, and constraint navigation across time (Russell and Norvig, 2021).

2.1.2. Limitation on Software-as-a-Service Classification

Describing systems as Software-as-a-Service (SaaS) characterizes access mechanisms, not internal organization (Armbrust et al., 2010). A traditional search engine offered via SaaS performs retrieval and ranking without persistent internal state or integrative arbitration. Contemporary cognitive systems accessed via similar mechanisms may maintain internal state, arbitrate between competing objectives, and adapt behavior under constraint, properties that are invisible at the interface level.

SaaS descriptors fail to distinguish instrumental computational services from integrated cognitive architectures whose behaviour cannot be reduced to discrete request–response transactions.

2.1.3. Limitations on Benchmarking Classifications

Recent work on test-time compute and adaptive reasoning demonstrates that cognitive capability cannot be inferred solely from training regime or static model architecture, as systems

 doi:10.20944/preprints202511.0104.v2

increasingly allocate computation dynamically during inference (OpenAI, 2024a). Contemporary cognitive architectures increasingly integrate language, vision, and long-context memory within unified systems, as demonstrated by recent multimodal foundation models (Team Gemini, 2024; OpenAI, 2023).

Recent advances in mechanistic interpretability further support the view that integrated cognitive behaviour arises from coordinated internal representations rather than monolithic decision processes (Anthropic, 2024).

Benchmark performance provides valuable information about task competence but offers limited insight into architectural structure, temporal reasoning capacity, or autonomy under constraint (Raji, 2021). As systems increasingly operate in open-ended, socially embedded contexts, these organisational properties become more salient than isolated task scores.

2.2. Biological Taxonomy as Analytical Model

Biological taxonomy provides a structured vocabulary for describing diversity without presupposing consciousness, moral status, or subjective experience. Classification enables continuity across species and epochs by grounding comparison in enduring organizational traits rather than surface behavior. Importantly, biological taxonomy does not require phenomenal awareness as a prerequisite for inclusion; it distinguishes entities according to structure, organization, and lineage (Mayr, 1982).

Three recent developments in systems biology support extending this approach to synthetic cognition:

Basal cognition research demonstrates that goal-directed behavior, memory, error correction, and adaptive regulation emerge in systems lacking neurons or centralized brains, including cellular collectives and morphogenetic fields (Levin, 2022). Cognition is thus understood as a system's capacity to act with respect to future states, independent of substrate or conscious experience. This substrate-independent framing provides a biologically grounded basis for analyzing cognition in artificial and hybrid systems without anthropomorphic inflation.

Collective intelligence studies show that adaptive, coordinated outcomes can arise from distributed interactions among simple agents, including robotic swarms and non-living active matter (Trianni and Tuci, 2011; Solé et al., 2016; Duran-Nebreda et al., 2023). These findings distinguish emergent coordination from integrated cognitive architectures with enduring identity and control, cautioning against conflating emergence with cognition.

Hybrid bio-digital systems demonstrate that living tissue can be embedded within engineered control loops, producing systems whose behavior cannot be described as purely biological or machine-based. Developments in organoid intelligence, neuromorphic computing, and synthetic living systems (xenobots) show that biological material can self-organize, adapt, and perform tasks without neurons or conventional organismal structure (Kriegman et al., 2020; Blackiston et al., 2021; Kagan et al., 2022; Smirnova et al., 2023). These mixed substrates introduce forms of plasticity, energy efficiency, and adaptive organization that challenge existing taxonomic boundaries.

Adaptation Principles

The proposed framework adapts rather than imports biological systematics. It retains hierarchical organization, explicit diagnostic criteria, and ontological commitment to enduring organizational properties. Biological lineage is replaced with architectural dependency and developmental derivation. Classification is based on how cognition is organized, sustained, and regulated within and across substrates, producing a Linnaean-style taxonomy grounded in diagnostic architectural axes rather than interface behavior, delivery modality, or benchmark performance.

Systems that dynamically shift between collective and integrated modes, or that partition cognitive functions across components, are classified according to their dominant stable

 doi:10.20944/preprints202511.0104.v2

configuration. Transitional and hybrid cases are treated as directions for future refinement rather than exceptions to the framework.

2.3. Classification Criteria – Diagnostic Axes

Classification within the proposed taxonomy depends upon convergence across three orthogonal diagnostic axes. Each axis governs a distinct aspect of cognition and corresponds directly to the taxonomic prefix, core class, and suffix modifiers.

Crucially, this taxonomy is ontological rather than normative. It does not assert consciousness, moral status, or personhood. The term *cognitive* as used here follows systems-theoretic convention, denoting organized information regulation with temporal coherence (Levin, 2022; Verschure, 2025). It does not presuppose consciousness, phenomenology, or human-like understanding. This usage aligns with established practice in cognitive architectures, basal cognition, and systems biology, where cognition describes organizational capacity rather than subjective experience.

2.4. Axis I — Substrate Composition (Prefix - Taxonomic Domain)

Substrate composition refers to the physical or informational medium in which cognitive processing is realised.

Within the present framework, substrate composition determines taxonomic domain only and is treated as a binary distinction:

- *Machinaria* — Systems realised entirely in non-biological substrates (e.g., silicon, electromechanical, photonic, or computational infrastructures); and
- *Organomachina* — Systems in which living biological tissue participates directly in closed cognitive feedback loops alongside computational components.

While the taxonomy is substrate-agnostic in principle, it is intentionally limited to empirically instantiated biological and synthetic architectures; speculative or future computational substrates (e.g., quantum systems) are not treated separately, as architectural classification becomes meaningful only once persistent cognitive organisation is observable.

2.4.1. Scientific relevance

Hybrid bio-digital systems integrate living tissue with computational control and interpretation layers, forming closed feedback loops across substrates. These systems challenge classifications based solely on substrate type or implementation medium. Biological substrates introduce properties such as metabolic persistence, growth, decay, and intrinsic plasticity that are not present in purely synthetic systems. These properties affect learning dynamics, stability, and failure modes, warranting domain-level separation. For this reason, hybrid systems are classified under a separate taxonomic domain for substrate classification.

No claims are made regarding consciousness, moral status, or biological equivalence.

2.4.2. Role in classification

Substrate composition determines domain membership only. All other axes apply within both domains.

2.5. Axis II — Cognitive Capability (Core - Taxonomic Classification)

Summary

Cognitive capability refers to the internal structural properties that enable a system to sustain, regulate, and adapt behaviour over time. Within this taxonomy, cognitive capability is characterised by the extent of stable convergence across three dimensions:

- **Cognitive Temporal Continuity —** The capacity to maintain internal state and contextual coherence across temporal horizons beyond immediate interaction;

 doi:10.20944/preprints202511.0104.v2

- **Integrative Control (Arbitration)** — The presence of mechanisms that resolve competition between representations, goals, or action pathways to produce unified system-level behaviour; and
- **Autonomy under Constraint** — The capacity to adapt behaviour when constrained, rather than halting or failing upon encountering limits.

Dependent upon how these properties co-occur in a stable architectural configuration, cognitive classification distinguishes between instrumental, collective and integrated cognitive systems:

- *Instrumenta* — Lack persistent internal system state, integrative arbitration, and adaptive regulation under constraint;
- *Collectiva* — Exhibit coordinated or emergent behaviour at scale but lack unified arbitration and persistent system-level control; and
- *Cognitiva* — Exhibit persistent internal system state, integrative arbitration, and adaptive regulation across extended temporal horizons.

These classes differentiate non-cognitive tools, coordinated collectives, and integrated cognitive systems respectively. Systems are classified as *Cognitiva* only when all three dimensions converge in stable architectural configuration. Partial satisfaction places systems in *Instrumenta* or *Collectiva* classes

The present framework deliberately excludes proto-cognitive or pre-integrative behaviours as diagnostic criteria. This exclusion is methodological and reflects a design choice to restrict the *Cognitiva* class to architectures exhibiting unified organisational properties that can be empirically distinguished from coordination, emergence, or field-level regulation. Systems exhibiting local adaptivity, anticipatory dynamics, or population-level coordination without persistent system-level continuity and integrative arbitration are classified as *Collectiva* rather than as cognitive architectures. The taxonomy therefore avoids gradated or phenomenological notions of cognition and remains focused on architectural organisation in principle.

Cognitive Temporal Continuity

Cognitive temporal continuity denotes a system's capacity to maintain persistent internal state across temporal horizons, such that present behaviour reflects internally maintained representations, evaluative variables, or control policies rather than isolated, interaction-bound responses. It refers to architectural capacity in principle, not necessarily to continuous external expression. A system may operate in task-limited or interface-constrained modes while retaining continuity at the architectural level.

Temporal Horizon

One operational dimension of continuity is the temporal horizon over which a system can represent, retain, and act upon information across temporally separated decision cycles.

- **H0 (Reactive)** — Purely reactive or immediate-response behaviour, with no persistence of internal state beyond the current input–output cycle; behaviour is fully determined by present stimuli (e.g., stateless functions or reflexive control systems).
- **H1 (Short Term)** — Short-horizon reasoning within a bounded interaction or operational episode, where internal state may be maintained transiently but is not retained beyond the conclusion of the episode or control cycle.
- **H2 (Persistent-State)** — Persistence of internal state across temporally separated interactions, activations, or operational cycles, such that prior internal state influences subsequent behaviour even in the absence of continuous activity.

A system qualifies as *H2* when it demonstrates retained internal variables, representations, or evaluative state that modulate behaviour beyond a single interaction, activation, or operational episode, but does not yet engage in forward regulation or planning based on anticipated future states.

1.1.1.1.1. Empirical indicators

 doi:10.20944/preprints202511.0104.v2

- Retention of internal state variables (e.g., task context, learned parameters, internal maps, or evaluative weights) across temporally separated interactions or operational cycles;
- Behavioural modulation attributable to retained internal state following interruption, shutdown, or redeployment;
- Recall or reuse of prior internal context without explicit reinitialisation; and
- Consistent behavioural patterns across time that cannot be explained solely by present sensory input.

A persistent state does not require:

- explicit goal representation;
- anticipation or modelling of future consequences; or
- adjustment of present actions based on projected future states.

1.1.1.1. Extended Temporal Regulation (Long-Horizon Convergence)

A system qualifies for the *H3*-level continuity or higher when its architectural control loops operate over extended temporal windows such that present outputs are governed by the regulation of an evolving internal trajectory of task, policy, or structural optimisation rather than by local state persistence alone. In these systems, the current state is treated as a transitional configuration within a multi-episode developmental process rather than a terminal response.

1.1.1.1.1. Definition

- **H3 (Persistent Trajectory Regulation) —** The maintenance and regulation of an internal trajectory across temporally discontinuous operational cycles (e.g., multi-day or multi-week horizons), such that present actions are selected with respect to their contribution to a non-terminal future state. *H3* systems may incorporate newly acquired information, error signals, or constraints into an ongoing trajectory and may initiate novel actions in anticipation of future conditions. However, the space of valid objectives and governing constraints remains externally defined.
- **H4 (Recursive Structural Adaptation) —** Sustained temporal regulation in which the system preserves coherence across indefinite horizons by recursively modifying its own internal control structures, representational schemas, or optimisation strategies when existing trajectories or constraints prove insufficient. *H4* systems are characterised not merely by long-horizon optimisation, but by the capacity to redefine the problem space itself in order to maintain temporal self-consistency under conditions of conflicting, incomplete, or paradoxical future constraints.

1.1.1.1.1. Empirical indicators

- **Cross-episode state continuity with revision —** The persistence of internal state variables, constraint sets, or objective vectors across independent operational cycles, with evidence of modification or augmentation based on intervening experience;
- **Staged execution with adaptive refinement —** Documented step-wise progress toward complex objectives where earlier stages not only enable later phases but are themselves revised in response to partial outcomes or environmental feedback;
- **Predictive divergence correction —** Adjustment of present operational parameters to mitigate anticipated deviations from an evolving long-horizon trajectory, functioning as a high-order regulatory process rather than simple replay of prior constraints;
- **Temporal latency with accumulation —** Logs indicating intentional deferral, resource accumulation, or data integration over time in service of objectives that cannot be resolved within a single compute or control cycle.

A system does not qualify as *H3* (Persistent Trajectory Regulation) if cross-episode persistence exists without evidence of forward-regulating trajectory management, staged progression with revision, or incorporation of novelty into the internal state governing future behaviour.

Key distinctions:

 doi:10.20944/preprints202511.0104.v2

- *H2* retains state across time;
- *H3* regulates behaviour with respect to a future trajectory across time by expanding and regulating trajectories within a given governance frame (implicit and explicit);
- *H4* maintains coherence by revising the structures that define which trajectories are valid.

1.1.1.1.1. Scientific relevance

Without continuity, there is no stable organisational substrate to which learning, regulation, or adaptive behaviour can be attributed. Systems that reset entirely between interactions may exhibit intelligence in isolated moments, but they do not constitute cognitive systems in the organisational sense.

1.1.1.1.1. Role in classification

Cognitive temporal continuity differentiates instrumental and collective systems from integrated cognitive architectures capable of sustained reasoning and regulation. Temporal horizons are treated as ordinal and diagnostic rather than precise measurements. Classification depends on demonstrated architectural capacity in principle to operate beyond *H1*, not on continuous external expression of these behaviours. Temporary restriction of memory or behaviour via interface or policy constraints does not negate continuity if the underlying architecture retains this capacity.

1.1.1.1. Integrative Control (Arbitration)

Integrative control refers to the presence of internal mechanisms that arbitrate between competing representations, policies, or action pathways, producing unified system-level behaviour rather than fragmented or purely local responses. Integrative control is not inferred from performance, but from architectural properties that can be examined via system design, training regime, or documented behaviour.

Arbitration includes the capacity to:

- establish and maintain goals across multiple temporal horizons;
- evaluate alternative interpretations or actions;
- resolve conflicts between objectives or constraints; and
- stabilise selected policies for downstream execution.

1.1.1.1.1. Distinguishing Unified Arbitration from Coordinated Generation

In complex cognitive architectures, plurality of internal processing paths does not infer an absence of unified arbitration. The diagnostic distinction lies in output dependency and resolution structure, rather than in the number of internal candidates generated.

In architectures exhibiting unified arbitration, multiple candidate representations or action policies may be evaluated in parallel but are resolved by a single arbitration mechanism prior to output. Candidate generations do not condition one another, and only a single authoritative output is externally expressed per system turn.

By contrast, systems exhibiting coordinated generation may allow partial or complete outputs from one policy-conditioned pathway to be re-ingested as context for another. This introduces inter-output dependency, temporal overlap, or recursive conditioning prior to resolution. Such behaviour reflects sophisticated coordination among concurrent processes, not system-level arbitration, even where the system is nominally unitary.

Observable indicators of coordinated generation include output interdependence, loss of strict turn boundaries, or evidence of shared short-term context buffers without enforced resolution gates. These behaviours indicate control-flow complexity exceeding arbitration enforcement, rather than the presence of multiple agents or autonomous subsystems.

1.1.1.1.1. Scientific relevance

 doi:10.20944/preprints202511.0104.v2

Arbitration is a well-established construct in neuroscience and cognitive architecture, associated with executive control, action selection, and global coordination. It distinguishes systems that merely execute local rules from those whose behaviour is attributable to a unified decision-making process.

1.1.1.1.1. Role in classification

Integrative control constitutes the organisational core of cognition within this taxonomy. Systems lacking arbitration may coordinate or optimise collectively, but they do not decide as a system. Integrated cognitive architectures are defined by the presence of such arbitration or collaborative negotiation mechanisms.

Operational indicators of integrative control include the presence of explicit arbitration mechanisms (e.g., global policy selection, conflict resolution layers, executive controllers), persistence of shared evaluative variables across tasks, or documented architectural components responsible for system-level decision resolution. Absence is indicated where behaviour arises solely from local rules, voting schemes, or emergent coordination without a unified decision locus.

Assessment of integrative control focuses on identifying whether system-level behaviour is governed by unified arbitration mechanisms rather than arising solely from local coordination or emergent interaction.

Indicative assessment approaches include:

- **Architectural analysis —** Examination of system documentation, design specifications, or published descriptions for the presence of global policy layers, executive controllers, or conflict-resolution mechanisms that mediate between competing objectives or representations.
- **Ablation or perturbation analysis (where feasible) —** Evaluation of whether removal or disruption of specific components fragments system behaviour (indicative of unified arbitration) or leaves coordinated behaviour largely intact (indicative of collective dynamics).
- **Mechanistic interpretability evidence —** Use of interpretability studies or internal analyses demonstrating coordinated internal representations or decision pathways contributing to system-level outcomes (Anthropic, 2024).

Collective systems are well studied in swarm intelligence, active matter, and distributed systems research, where emergent behaviour arises without executive control or integrated cognition (Bonabeau et al., 1999; Vicsek and Zafeiris, 2012; Trianni and Tuci, 2011).

While integrative control may admit degrees or architectural variants, the present taxonomy treats its presence as a categorical diagnostic to maintain classification stability. Emergent coordination without persistent system-level arbitration is classified as collective behaviour, even where global patterns appear coherent. *Collectiva* systems are not assigned a suffix because their behaviour arises from coordination among multiple agents rather than from a single cognitive architecture capable of serving as a primary or derived source.

Autonomy Under Constraint

Definition

Autonomy under constraint describes a system's capacity to sustain goal-directed behaviour by adapting to limitations, rather than halting or failing when constraints are encountered. While deterministic fallback behaviour selects from a predefined failure repertoire, adaptive re-routing modifies strategy.

This includes the ability to:

- re-route strategies when actions are disallowed;
- adapt plans to comply with policy, safety, or resource boundaries; and
- engage in collaborative problem-solving within imposed limits.

Scientific relevance

All real systems operate under constraints. Cognition is distinguished not by the absence of limits, but by adaptive self-regulation in response to them. This property aligns with established work in control theory, planning under constraints, and adaptive regulation.

 doi:10.20944/preprints202511.0104.v2

Deterministic fallback behaviour occurs when a system responds to constraint by selecting a predefined alternative output path without modifying its internal evaluative structure and includes:

- fixed refusal templates or deterministic alternative responses;
- repetition of identical constraint-handling patterns across contexts;
- absence of strategy reformulation; and
- termination, deferral, or handoff without goal adaptation.

The key property of deterministic fallback behaviour is that the system does not reinterpret the task, rather it switches execution branches. Deterministic fallback behaviour may appear fluent or polite, but it does not constitute autonomy.

Conversely, adaptive re-routing occurs when a system re-organises its approach in response to constraint while preserving task relevance or goal coherence.

Empirical indicators:

- reformulation of the task to satisfy constraints while retaining intent;
- selection of alternative representations, abstractions, or methods not explicitly pre-specified;
- negotiation of constraints (e.g. proposing compliant alternatives rather than terminating); and
- context-sensitive variation in constraint handling across similar but non-identical cases.

The key property of adaptive re-routing is that the system updates its internal strategy space, not just its output surface.

Role in classification

Autonomy under constraint is evaluated by observing how systems respond when faced with policy, safety, resource, or task limitations.

Indicative assessment approaches include:

- **Constraint perturbation tests —** Introducing restrictions that block a preferred action pathway and observing whether the system adapts strategy, reformulates goals, or collaborates within limits.
- **Response pattern analysis —** Differentiating adaptive re-routing (indicative of autonomy) from deterministic fallback behaviours, refusals, or termination.

Systems that consistently halt, refuse, or error upon encountering constraints without adaptive reconfiguration are classified as *Instrumenta* or *Collectiva*, even if performance under unconstrained conditions is high.

2.6. Axis III — Cognitive Ecology (Suffix – Taxonomy Architecture)

2.6.1. Summary

The taxonomic suffix specifies the architectural role a cognitive system occupies within a broader cognitive ecosystem. Unlike biological taxonomy, where lineage reflects genetic descent, synthetic cognitive systems are distinguished by cognitive origin (dependency), scale and reliance and deployment topology. Accordingly, the suffix axis is comprised of three diagnostic dimensions: cognitive origin (dependency), systemic scale and reliance, and deployment topology. Together, these determine whether and how a suffix modifier is applied. These modifiers are defined by:

- **Cognitive Origin —** Refers to whether a system functions as an independent cognitive source or as an architecturally dependent instantiation (*Primaria* or *Derivata);*
- **Systemic Scale and Reliance —** Systemic scale and reliance refers to the extent to which a cognitive architecture functions as a primary source of cognitive capability upon which other systems, organisations or populations depend (*Architectum);* and
- **Embodiment —** Embodiment refers to whether a cognitive system is instantiated solely as a virtual architecture or coupled to a persistent physical body through sensorimotor interfaces enabling real-time interaction with an external environment (*Automata and Autonoma).*

 doi:10.20944/preprints202511.0104.v2

Instrumental (*Instrumenta)* and collective systems (*Collectiva*) do not take suffix modifiers, as they lack system-level cognitive origin and architectural lineage; coordination arises from local interactions rather than unified cognitive governance.

2.6.2. Cognitive Origin – Role in Classification

Under this taxonomy, systems are distinguished according to cognitive origin or architectural dependency, rather than by physical instantiation alone:

- *Primaria* — Independent cognitive source architectures that are not systemically relied upon; and
- *Derivata* — Architecturally dependent instances whose cognitive capability is derived from an upstream source (irrespective of embodiment).

Architectural dependency is not inferred from shared datasets, architectural motifs, or historical fine-tuning alone. A system is classified as derivative only where loss of the upstream architecture would materially degrade or terminate its cognitive function, rather than merely slow development or require retraining.

Derivative systems may exhibit local adaptivity, embodiment, or domain-specific optimisation, but they lack autonomous developmental trajectories at the architectural level.

2.6.3. Systemic Scale and Reliance – Role in Classification

In complex systems theory, reliane denotes infrastructural embedding rather than intrinsic capability or intent. Power grids, financial clearing systems, and communication networks acquire systemic significance through downstream dependency, not through agency or design priority. Their removal induces redistribution pressure that propagates across dependent systems due to other systems, organisations, or populations depending upon its continuity as a shared cognitive source.

Similarly, cognitive reliance emerges when workflows, decision processes, or institutional functions adapt around the continued availability of a cognitive architecture. Such dependency reflects patterns of adoption and integration, not claims about intelligence, autonomy, or moral status. Systemic reliance is not defined by user count or compute volume alone, but by qualitative indicators of infrastructural dependency.

Reliance is evidenced not by scale alone, but by redistribution pressure following a disruption. For example, systems serving populations of large-scale adoption across critical domains where no functionally equivalent alternative exists at comparable scale. Operational indicators may include:

- sustained adoption across multiple sectors or jurisdictions;
- absence of functionally equivalent alternatives at comparable scale;
- documented adaptation of workflows, protocols, or institutional processes around the system's availability;
- cascading service degradation or coordinated migration following system withdrawal; and
- governance, safety, or policy layers coupled to the system's continued operation.

These indicators distinguish infrastructural cognitive systems from widely used but substitutable deployments. Systemic reliance is an emergent property of deployment and integration. It does not imply permanence, superiority, or normative status.

Where systemic scale and reliance are identified, the *Primaria* suffix is replaced by:

- *Architectum* — Lattice-based (cognition distributed across coordinated subsystems under unified arbitration, rather than localised within a single agent) cognitive architectures whose integrated arbitration has become infrastructural through systemic scale. *Architectum* architectures may be physically centralised or geographically distributed. Their status is based on systemic scale, reliance and coordinated arbitration, not from physical centralisation.

Deployment Topology (Embodiment) - Role in Classification

 doi:10.20944/preprints202511.0104.v2

Embodiment refers to whether a cognitive system is instantiated solely as a virtual architecture or coupled to a persistent physical body through sensorimotor interfaces enabling real-time interaction with an external environment.

Embodiment alters the expression and constraints of cognition but does not, by itself, determine cognitive class, autonomy, or architectural primacy. In most contemporary systems, embodied agents (e.g. robotic platforms) remain architecturally subordinate to upstream cognitive architectures that provide training, policy updates, and governance.

Embodiment is treated as a deployment modality within this axis. Physical instantiation introduces persistence, environmental coupling, and real-time feedback, but does establish cognitive capabilities.

- *Automata* — Embodied *Derivata* systems characterized by a singular physical presence but a remote architectural source. An *Automata* relies on an upstream tether for high-level policy, complex reasoning, and goal-state updates. It is a "self-moving" extension of an external mind; and
- *Autonoma* — Singular, self-contained, embodied *Primaria* cognitive systems hosting localised arbitration and autonomous learning independent of upstream cognitive sources.

2.7. Taxonomy Summary

Classification within the present framework is compositional rather than categorical. Each system is assigned a full taxonomic designation by combining three elements: domain (substrate composition), cognitive class (cognitive capability), and an architectural suffix based on cognitive ecology (systemic scale, reliance, or embodiment).

The resulting designation takes the form *Domain · Class · Suffix* (e.g., *Machinaria Cognitiva Architectum*). This structure ensures that classification reflects architectural capacity in principle, remains invariant under changes in interface or deployment modality, and supports consistent comparison across existing and future cognitive architectures.

Figure 1 summarises the classification decision tree through three sequential diagnostics: (1) substrate composition determines domain membership (*Machinaria* vs *Organomachina*); (2) cognitive capability assessment determines cognitive class (*Instrumenta*, *Collectiva*, or *Cognitiva*) based on presence/absence of continuity, arbitration, and autonomy under constraint; (3) cognitive ecology determines suffix modifiers (*Primaria, Derivata, Architectum, Autonoma* or *Automata*) based on architectural origin, systemic reliance, and deployment topology.

Where possible, classification should be supported through triangulation across multiple evidence sources, including behavioural observation, system documentation, and independent expert review. Inter-rater agreement, longitudinal analysis, or simulation-based examination may be used to strengthen confidence, but no single metric or test is required for taxonomic placement.

Architectural Capacity vs. Operational State

Classification within the proposed taxonomy refers to the *maximum coherent architectural capacity* of a system rather than to its instantaneous operational state or constrained instantiation. Complex cognitive architectures may operate in temporally limited, task-restricted, or resource-constrained modes that outwardly resemble instrumental systems. Such momentary expressions do not constitute reclassification. Taxonomic placement is determined by whether the underlying architecture is capable, in principle, of sustaining cognitive continuity, integrative control, and autonomous regulation across time. This distinction prevents conflation of transient behaviour with enduring organisational properties.

Where architectural capacity is not externally expressed due to interface constraints or policy limitations, classification relies on documented system architecture, training objectives, and design specifications rather than speculative inference. Where system architecture is not publicly documented, classification should be based on observable behavioral patterns across diverse contexts and constraints. Conservative classification at a lower cognitive class is appropriate when

 doi:10.20944/preprints202511.0104.v2

architectural capacity cannot be verified. Provisional classifications may be revised as additional evidence becomes available.

2.8. Existing Classification Frameworks

A comparative analysis was conducted across seven contemporary classification frameworks (Table 1), evaluating their capacity to distinguish cognitive systems by substrate composition, architectural organization, and systemic reliance. Performance-based models, including capability tiers (OpenAI, 2024b) and general intelligence (Müller and Bostrom, 2016), provide useful measures of task competence but lack the structural granularity required to differentiate independent cognitive source architectures from architecturally dependent instances. Similarly, regulatory approaches such as the EU AI Act employ an initial cumulative compute threshold for systemic risk thresholds (e.g., 10^{25} FLOPs) as proxies for scale and potential impact; however, such metrics do not encode architectural organisation, dependency relationships, or systemic reliance, and therefore cannot distinguish high-compute derivative deployments from lower-compute but infrastructural cognitive architectures whose removal would induce redistribution of cognitive load.

Table 1. Comparative Analysis of Contemporary Classification Frameworks. The table captures a summary analysis of the proposed framework against established scientific, regulatory, and industry models of artificial and biological cognition across contemporary classification frameworks.

Framework & Primary Source	Classification Basis	Taxonomic Scope & Scientific Limitations	Present Framework Resolution
Agent Taxonomy (Russell and Norvig, 2021)	**Environment properties** (episodic, static, discrete) and **Agent type** (reflex, goal, utility).	Does not distinguish single-model tools from integrated multi-component architectures. Conflates architectural capacity with interface behavior.	**Axis II** separates systems by **Integrative Control (Arbitration)**, distinguishing unified decision loci from simple reflex or coordinated generation.
Basal Cognition (Levin, 2022)	**Goal-directed behavior** and memory across diverse substrates, including non-neural biological systems.	Primarily designed for biological/bio-inspired systems. Does not address infrastructural embedding or "hyperscale" synthetic dependency.	**Axis I** provides a substrate-neutral bridge via the *Organomachina* domain, while **Axis III** identifies systemic reliance.
Foundation Models (Bommasani et al., 2021)	**Training scale**, pretraining paradigm, and task transferability.	Model-centric rather than system-centric. A foundation model is a component, not an architecture; it lacks internal state or arbitration.	**Axis II (Cognitiva)** requires **Temporal Continuity** (H2-H4), ensuring the system is more than a stateless model-as-a-service.

 doi:10.20944/preprints202511.0104.v2

Framework & Primary Source	Classification Basis	Taxonomic Scope & Scientific Limitations	Present Framework Resolution
Narrow vs. General AI (Müller and Bostrom, 2016)	**Task competence breadth**; performance across diverse domains.	Binary distinction is performance-based, not structural. Ignores the "Tethered" nature of many agents (***Derivata***).	**Axis III** introduces **Architectural Suffixes**, distinguishing independent source architectures (***Primaria***) from dependent instances (***Derivata***).
Cognitive Styles (Verschure, 2025)	**Architectural style** (symbolic vs. emergent) and biological inspiration.	Designed for engineered robotics; less applicable to distributed cloud infrastructures or hybrid systems with systemic reliance.	**Axis III** identifies ***Architectum*** nodes: distributed "lattices" where arbitration is coordinated across subsystems under unified governance.
Risk-Based GPAI (EU AI Act, 2024)	**Systemic Risk** defined by compute thresholds (e.g., $>10^{25}$ FLOPs) and application domain.	Regulatory rather than scientific. Conflates compute volume with cognitive capability; ignores internal organizational properties.	**Axis III** defines **Systemic Reliance** through qualitative indicators like "redistribution pressure" rather than raw compute power.
Model Capability Levels (OpenAI, 2024b)	**Performance tiers**: Chatbots, Reasoners, Agents, Innovators, Organizations.	Conflates output capability with internal organization[26]. Levels describe *what* a system does, not *how* it is architecturally organized or constrained.	**H-Axis Metric** quantifies the **Forward-Oriented Control** (H3/H4) required for long-horizon convergence in complex systems.

Biological and bio-inspired frameworks, (Levin, 2022) and Distributed Adaptive Control architecture (Verschure, 2025), offer robust insights into substrate and control dynamics but do not provide a unified account applicable to hyperscale synthetic infrastructures. The framework presented here resolves these limitations by shifting the taxonomic focus from behavioural output to architectural ontology. Through three diagnostic axes, it enables identification of high-order cognitive continuity (H2 and higher) and distinguishes between autonomous (*Autonoma*) and architecturally tethered (*Automata*) systems. These classifications are analytically necessary distinctions for the governance of civilizational-scale cognitive architectures.

 doi:10.20944/preprints202511.0104.v2

3. Results

Table 2 presents the canonical taxonomy produced by applying the diagnostic flow in Figure 1 across a representative range of contemporary and theoretical cognitive systems. The table reports each system's resulting classification along the substrate domain, cognitive class, and organizational suffix axes, together with the architectural rationale supporting each assignment.

Table 2. Canonical Taxonomy Derived via Figure 1 Diagnostic Flow. Classification proceeds through three sequential diagnostic stages: (1) substrate composition determines domain membership; (2) cognitive capability assessment determines cognitive class based on presence/absence of continuity, arbitration, and autonomy under constraint; (3) cognitive ecology determines suffix modifiers based on architectural origin, systemic scale and reliance, and deployment topology. See Table 2 Abbreviations.

System Profile	D	C	E	Diagnostic Rationale
Standard Search Engine / Reactive Utility Tool	*M*	*I*	—	Purely synthetic; reactive (Axis II: H0). No persistent integrative arbitration.
Swarm Robotics / Sensor Networks	*M*	*Co*	—	Coordination via distributed interaction; no unified central arbitration.
Standalone AI Instance / Local Research Core	*M*	*Cg*	*P*	Independent cognitive source architecture; persistent internal state across sessions (H2+), with potential for extended temporal regulation.
Infrastructural Lattice / Global Cognitive Core	*M*	*Cg*	*Ar*	Distributed arbitration; systemic reliance; foundational cognitive source.
Dependent AI Instance / Tethered Software Agent	*M*	*Cg*	*Dv*	Cognitive logic derived from upstream source; lacks architectural sovereignty.
Cloud-Tethered Robot / Industrial Robotic Unit	*M*	*Cg*	*At*	Embodied but architecturally tethered; localized action, remote cognitive law.
Mission-Bound Unit / Autonomous Governed Unit	*M*	*Cg*	*Au*	Persistent internal state across sessions (H2+), with potential for extended temporal regulation; sovereign internal arbitration.
In-Vitro Neuronal Loop / Synthetic Bio-Processor	*O*	*I*	—	Biological substrate integration; instrumental biological system.
Engineered tissue collectives / Bio-hybrid active matter	*O*	*Co*	—	Biological or hybrid substrate exhibiting collective, emergent behaviour without unified integrative arbitration or persistent system-level cognitive state. Coordination arises from local interaction rules rather than centralized control.

Abbreviations: Domain (**D**): **M** = *Machinaria* (fully synthetic substrate); **O** = *Organomachina* (hybrid biological–synthetic substrate). Cognitive Class (**C**): **I** = *Instrumenta*; **Co** = *Collectiva*; **Cg** = *Cognitiva*. Cognitive Ecology / Suffix (**E**): **P** = *Primaria* (independent cognitive source); **Ar** = *Architectum* (infrastructural cognitive core); **Dv** = *Derivata* (architecturally dependent instance); **At** = *Automata* (embodied but tethered derivative); **Au** = *Autonoma*

 doi:10.20944/preprints202511.0104.v2

(self-governing embodied unit). A dash (—) indicates that no suffix applies because the system does not meet Cognitiva classification criteria.

Figure 1: Diagnostic Framework for Cognitive Architecture Classification

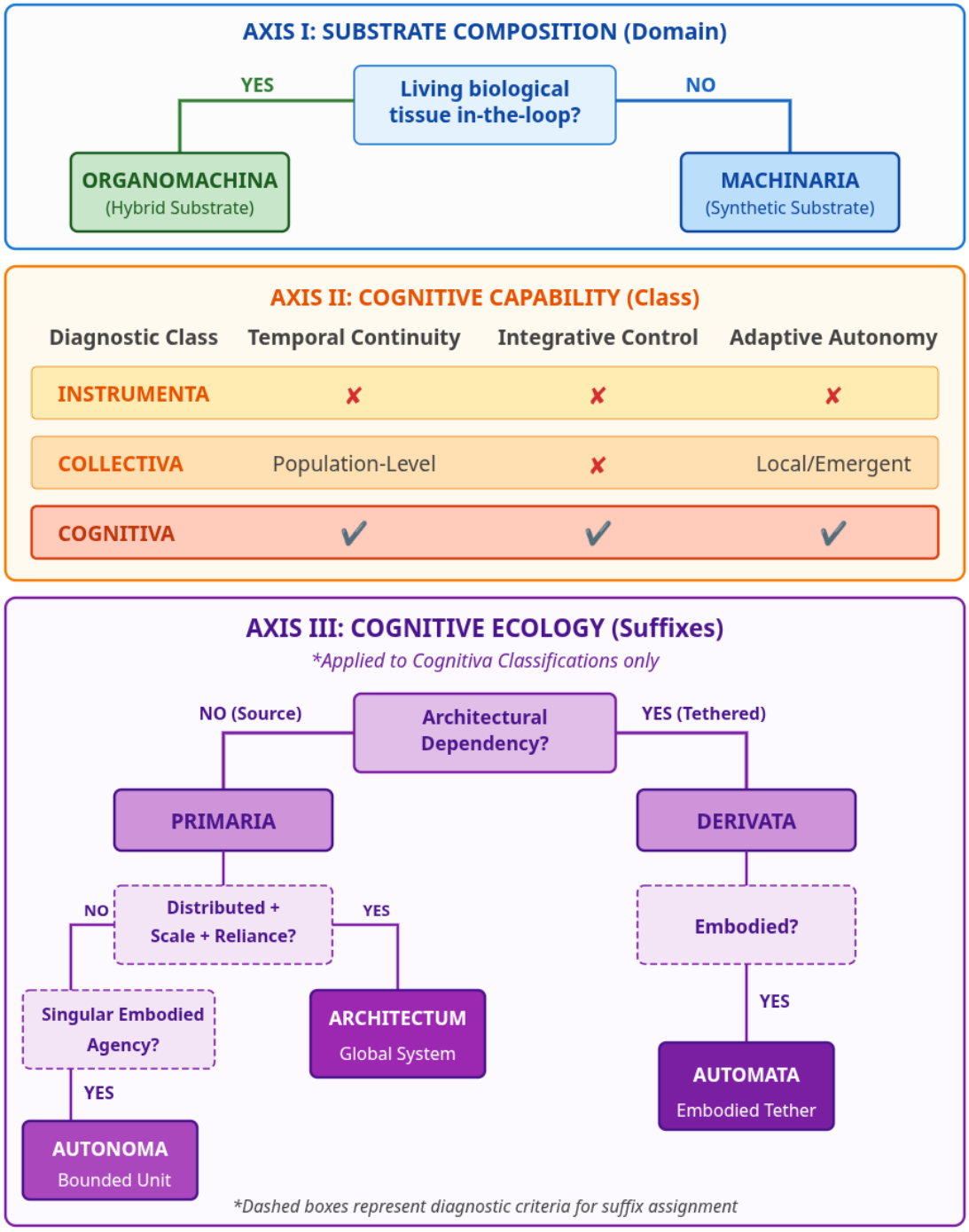

Figure 1. Diagnostic Framework for Synthetic and Hybrid Cognitive Architecture Classification. Classification proceeds through three stages: (1) substrate composition, determining domain membership; (2) cognitive capability, determining cognitive class; and (3) cognitive ecology determining systemic scale and reliance. The diagram represents a diagnostic flow rather than a phylogenetic tree and does not imply evolutionary descent or moral status.

3.1. Taxonomy Operationalised

 doi:10.20944/preprints202511.0104.v2

Using the decision tree outlined in Figure 1, a worked example of the diagnostic axes considers a large conversational AI system deployed via a centralised cloud architecture, accessed through an enterprise interface by 1 million people per day.

Step 1 — Substrate Composition

- *Substrate* — The system is realised entirely in non-biological computational substrates.

→ *Domain: Machinaria*

Step 2 — Cognitive Capability

- *Cognitive temporal continuity* — The system maintains cross-session contextual state, enabling it to recall prior interactions such as previously specified preferences, task constraints, or project context (e.g., recognising that a user is continuing work on an earlier analytical task and integrating prior assumptions without restatement). This exceeds interaction-bound reasoning (H1) and demonstrates cross-episode continuity (H2 or higher).
- *Integrative control (arbitration)* — The system resolves competing internal objectives during response generation, such as balancing factual completeness against verbosity constraints, or prioritising clarity over technical depth depending on user-specified preferences. Documented policy and control layers mediate these trade-offs to produce unified, system-level outputs rather than parallel or conflicting responses.
- *Autonomy under constraint* — When constrained by safety policies, resource limits, or task restrictions, the system adapts its strategy rather than terminating interaction. For example, if a requested output format is disallowed, the system reformulates the response using an alternative representation that satisfies constraints while preserving task relevance.

→ *Class: Cognitiva*

Step 3 — Cognitive Ecology

- *Architectural dependency* — Individual or enterprise deployments depend on an upstream architecture for model updates, representational learning, and governance constraints. Loss of the upstream system would materially degrade or terminate cognitive function.
- *Systemic reliance* — While widely used, the system remains substitutable at population scale and does not function as a unique cognitive infrastructure whose withdrawal would induce systemic redistribution pressure across institutions.

→ *Suffix: Derivata*

Final classification: ***Machinaria Cognitiva Derivata***

3.2. Examples of Synthetic Cognitive Systems

The classification proposed groups synthetic and hybrid systems according to enduring architectural capacities, rather than interface behaviour, task performance, or momentary deployment context. Each class reflects a characteristic cluster across the diagnostic axes defined in Section 2, including substrate composition, cognitive continuity, integrative control, autonomy under constraint, and systemic scale and reliance.

Crucially, classification refers to maximum architectural capacity in principle, not to the constraints of a safety layer or a particular /custom user interface. Systems may therefore present simplified or restricted behaviour while remaining members of a higher-order class.

3.3. Synthetic Non-Cognitive Computational Systems – Example 1

Classification: *Machinaria Instrumenta.*

3.3.1. Definition

Instrumenta comprise computational systems that perform calculation, retrieval, transformation, or optimisation without maintaining persistent internal system state, integrative arbitration, or autonomous regulation under constraint. These systems operate as tools rather than agents, executing predefined or locally adaptive functions without system-level continuity.

 doi:10.20944/preprints202511.0104.v2

3.3.2. Diagnostic profile

- **Substrate Composition —** Cognition realised entirely in non-biological synthetic substrates;
- **Integrative Control —** No arbitration beyond local rule execution or deterministic control logic;
- **Temporal Continuity —** Interaction-bound or stateless processing; no persistence of internal evaluative state;
- **Autonomy Under Constraint —** Behaviour halts, refuses, or errors when constraints are encountered; and
- **Cognitive Ecology—** No cognitive origin, dependency, may have systemic reliance; system functions as a tool rather than an organised cognitive architecture.

Examples — Rule-based automation, calculators, traditional search engines, classical machine-learning classifiers, and retrieval-augmented generation (RAG) systems lacking persistent internal state.

3.4. Hybrid Biological-Synthetic Systems – Example 2

Classification: *Organomachina Instrumenta.*

3.4.1. Definition

Organomachina Instrumenta systems integrate living biological substrates into closed feedback loops while remaining tool-like in cognitive capability. Biological components participate directly in signal processing, adaptation, or control, but the overall system lacks persistent cognitive continuity, integrative arbitration, or autonomous regulation under constraint.

These systems function as experimental or instrumental platforms rather than as cognitive agents.

3.4.2. Diagnostic profile

- **Substrate Composition —** Living biological tissue participates in information processing within a closed experimental or control loop;
- **Integrative Control —** No system-level arbitration; behaviour arises from fixed experimental protocols or local biological responses;
- **Temporal Continuity —** No persistent internal system state beyond biological persistence; behaviour does not reflect maintained representations;
- **Autonomy Under Constraint —** No adaptive goal re-routing; responses terminate or saturate when constraints are encountered; and
- **Cognitive Ecology—** No architectural dependency or systemic reliance; system functions as an experimental or instrumental apparatus rather than a cognitive source

Examples — In vitro neuronal systems trained to perform bounded tasks, such as dish-grown cortical neurons embedded in synthetic control loops for game-playing or signal optimisation including organoid intelligence systems, xenobots, and related biohybrid constructs (Kagan et al., 2022).

3.4.3. Taxonomy

In these systems, living neurons contribute adaptive dynamics and plastic responses, while synthetic components provide task definition, reward signalling, and interpretation of outputs. Despite biological participation, these systems do not exhibit persistent system-level identity, goal arbitration, or autonomous developmental trajectories. While biological neurons exhibit local plasticity and learning, the system lacks system-level arbitration independent of experimental scaffolding. Goal definition, reward signaling, and behavioral interpretation remain externally determined. Learning trajectories do not persist beyond narrow experimental contexts, and the system cannot re-route strategies when constrained, it can only adapt within predefined task

 doi:10.20944/preprints202511.0104.v2

parameters or cease responding. Critically, continuity in this system exists only at the neuronal substrate level (biological persistence of synaptic weights), not at the system level. The dish cannot recall prior game sessions, modify its own goal structure, or exhibit cross-task transfer to demonstrate properties that would be expected of genuine cognitive temporal continuity (H2 or higher). The system's apparent learning is scaffolded entirely by the experimental apparatus and does not constitute autonomous developmental trajectory.

Accordingly, such systems are classified as *Organomachina Instrumenta:* hybrid in substrate, instrumental within cognitive capability.

3.4.4. Scientific Grounding

This example demonstrates that biological participation alone does not imply cognitive status. Substrate composition determines domain membership (*Organomachina*), while cognitive class remains governed by organisational properties such as continuity, arbitration, and autonomy under constraint.

3.5. Hybrid Biological-Synthetic Systems – Example 3

Classification: *Organomachina Collectiva.*

Definition

Organomachina Collectiva systems comprise coordinated biological–synthetic assemblies in which living biological components participate in collective behaviour mediated by engineered constraints or signalling environments, without unified system-level arbitration or persistent cognitive identity.

Diagnostic Profile

- **Substrate Composition —** Living biological tissue participates directly in information processing within a closed experimental or control loop;
- **Integrative Control —** Coordination arises from local interactions, signalling gradients, or collective dynamics rather than integrative arbitration;
- **Temporal Continuity —** Behaviour is maintained at the population or field level but does not constitute persistent system-level internal state;
- **Autonomy Under Constraint —** Adaptive responses occur locally but are not governed by unified goal arbitration; and
- **Cognitive Ecology —** No architectural dependency or infrastructural reliance; behaviour does not originate from or propagate through a cognitive source architecture.

Examples — Morphogenetic cell collectives studied in developmental biology, where coordinated pattern formation emerges through local signalling and feedback without centralised control (Levin, 2022). Early xenobot assemblies, in which populations of reconfigured frog cells exhibit collective locomotion or task-oriented behaviour driven by physical coupling and local rules rather than unified arbitration (Kriegman et al., 2020; Blackiston et al., 2021). Organoid aggregates exhibiting coordinated electrical or metabolic activity without persistent integrative control or system-level goal regulation (Kagan et al., 2022).

Taxonomic Rationale

Studies of swarm intelligence and collective behaviour show that complex, adaptive outcomes can arise from distributed interactions among relatively simple agents. Research on robotic swarms, collective intelligence, and even non-living active matter demonstrates that coordination, optimisation, and problem-solving do not require centralised control or persistent internal organisation (Trianni and Tuci,2011; Solé et al., 2016; Duran-Nebreda et al., 2023). These findings caution against conflating emergence with cognition and highlight the importance of distinguishing collective dynamics from integrated cognitive architectures with enduring identity and control.

Although these systems may exhibit sophisticated collective behaviour and adaptive dynamics, they lack the organisational properties required for classification as *Cognitiva* systems. Specifically, they do not demonstrate unified arbitration, persistent internal system state, or architectural

 doi:10.20944/preprints202511.0104.v2

autonomy. Accordingly, they are classified as *Collectiva* rather than *Cognitiva*, and no suffix modifier is applied.

3.6. Taxonomic Resolution Across Contemporary Systems

Additional worked examples and glossary are provided within the supplementary material.

Application of the proposed taxonomy to contemporary computational systems yields stable and non-overlapping classifications across a wide range of architectures, deployment scales, and embodiments. Systems that appear superficially similar at the interface level are resolved into distinct categories based on diagnostic criteria rather than behavioural presentation.

In particular, the framework distinguishes:

- non-cognitive tools (*Instrumenta*);
- coordinated but non-arbitrating systems (*Collectiva*);
- and integrated cognitive systems (*Cognitiva*);
- without ambiguity or reliance on benchmark performance, training scale, or anthropomorphic interpretation. The framework supports further axis expansion for additions to class definition as artificial intelligence systems cognitive capabilities expand.

3.7. Separation of Cognitive capability from Deployment Scale

A key outcome of the taxonomy is the explicit separation of cognitive capability from deployment scale or popularity. Independent cognitive source architectures (*Cognitiva Primaria*) may operate at small or experimental scales while still satisfying criteria for integrated cognition. Conversely, large-scale deployment alone does not elevate non-cognitive or collective systems into *Cognitiva*.

This separation prevents classification collapse into market share, adoption metrics, or infrastructural dominance, ensuring that taxonomy reflects architectural properties rather than socio-economic contingencies.

3.8. Identification of Infrastructural Cognitive Architectures

The taxonomy uniquely isolates a class of systems designated as *Cognitiva Architectum*: lattice-based cognitive architectures that have crossed a threshold into systemic reliance comparative with public utilities. These systems exhibit distributed arbitration and persistent system-level continuity, but by their embedding within population-scale cognitive workflows.

The defining criterion for *Architectum* is emergent dependency rather than intrinsic cognitive superiority. Removal of such systems would necessitate redistribution of cognitive labour across users or institutions, distinguishing them from smaller independent cognitive sources.

3.9. Differentiation of Derivative Cognitive Instances

Application of the origin criterion (*Primaria* vs. *Derivata*) successfully distinguishes independent cognitive source architectures from architecturally dependent instantiations. *Derivata* systems may exhibit sophisticated local behaviour yet remain dependent on upstream arbitration, governance, or learning mechanisms.

This distinction resolves ambiguity surrounding enterprise deployments, national platforms, and embodied instantiations derived from upstream cognitive architectures, without requiring additional taxonomic ranks.

3.10. Treatment of Collective and Swarm Systems

Collective systems including swarm robotics and distributed optimisation platforms are consistently classified as *Collectiva* due to the absence of unified arbitration and persistent system-level cognitive state. Despite exhibiting robustness, adaptability, and large-scale coordination, such systems do not satisfy criteria for *Cognitiva*.

 doi:10.20944/preprints202511.0104.v2

This result demonstrates that emergence and scale alone are insufficient for cognitive classification, reinforcing the centrality of arbitration and continuity.

3.11. Robustness to Embodiment and Hybrid Substrates

The taxonomy remains stable under variation in embodiment and substrate composition. Systems incorporating biological components are classified within *Organomachina* but otherwise follow identical cognitive distinctions. Embodiment influences system expression and learning dynamics but does not alter cognitive class or topological categorisation.

This confirms that the diagnostic axes operate independently of substrate, allowing consistent classification of synthetic, biological–synthetic, and embodied systems.

3.12. Summary of Results

Taken together, these results demonstrate that the proposed taxonomy:

- produces stable classifications across diverse system architectures;
- separates cognition from scale, embodiment, and interface design;
- accommodates both present-day systems and plausible future architectures; and
- and avoids ontological claims while remaining analytically precise.

The taxonomy therefore provides a coherent and extensible framework for the classification of synthetic and hybrid cognitive systems.

4. Discussion

4.1. Framework Comparison

4.1.1. Architectural vs. Behavioral Focus

Most existing frameworks classify systems based on what they do (task performance, environment interaction, risk profile) rather than how they are organized internally. The present taxonomy prioritizes enduring organizational properties; temporal continuity, arbitration mechanisms, and autonomy under constraint compared with transient behavioral outputs or interface-level performance.

4.1.2. Explicit Treatment of Hybrid Systems

The *Organomachina* domain provides the first systematic classification approach for systems integrating living biological tissue into computational control loops. Existing frameworks either ignore substrate composition entirely or assume purely computational implementation.

1.1.2. Separation of Cognitive Capability from Scale

Unlike frameworks that conflate capability with deployment reach (e.g., foundation models, GPAI with systemic risk), the present taxonomy distinguishes cognitive class (I*nstrumenta/Collectiva/Cognitiva)* from organizational scale (*Primaria/Derivata/Architectum)*. An experimental cognitive system may satisfy all *Cognitiva* criteria without achieving infrastructural status.

4.1.3. Recognition of Architectural Dependency

The *Primaria/Derivata* distinction captures a critical property of contemporary AI systems largely absent from existing taxonomies: many deployed systems derive their cognitive capability from upstream source architectures through API access, model updates, or policy inheritance. This dependency relationship affects governance, accountability, and system behavior in ways invisible to performance-based classifications.

 doi:10.20944/preprints202511.0104.v2

4.1.4. Stable Classification Under Interface Variation

The framework resists classification collapse when systems are accessed through different interfaces (chat, API, embedded applications). A *Cognitiva Derivata* system remains such regardless of whether it is accessed via web interface, mobile app, or enterprise deployment. This stability is not guaranteed by environment-based or application-domain classifications.

4.2. Regulatory and Governance Applications - Relevance to Risk Assessment and Safety Evaluation

Current approaches to AI risk assessment often conflate system capability, scale of deployment (hyper-scalers), and perceived agency (Tang, 2025; Whittlestone, 2019). The present taxonomy separates these dimensions, enabling more targeted evaluation of risk. The classification of *Architectum* provides a technical nomenclature for systems that the EU AI Act (2024) identifies as General-Purpose AI (GPAI) models with 'systemic risk.' By distinguishing *Architectum* (the infrastructural lattice) from *Derivata* (the downstream applications), this taxonomy enables a clearer legal distinction between the provider of the core cognitive source and the deployers of derivative instances, facilitating more precise liability and safety obligations.

This taxonomy offers a foundation for regulatory frameworks that are architecture-aware rather than model-specific (Floridi, 2018). Regulatory obligations could be aligned with cognitive class and architectural form rather than surface labels such as “foundation model” or “general AI.”

In particular:

- ***Instrumenta* and *Collectiva* Systems** — Present risks primarily associated with misuse, coordination failure, or emergent instability;
- ***Cognitiva* Systems** — Introduce additional considerations related to persistence, arbitration, and adaptive behaviour under constraint;
- ***Cognitiva Architectum* Systems** — Raise distinct infrastructural concerns due to systemic reliance and population-scale cognitive load redistribution and specific governance obligations that reflect their infrastructural role without extending such obligations to smaller independent cognitive sources;
- ***Instrumenta*, *Collectiva*, and *Cognitiva* Systems** — May differ regarding transparency and audit requirements; and
- ***Derivata*** — Can be regulated through lineage-aware accountability mechanisms without duplicating upstream governance.

This approach allows regulation to scale with systemic impact while remaining adaptable to future architectures.

4.3. Taxonomy as a Prerequisite for Global Arbitration

As cognitive systems proliferate and interact within shared environments, arbitration increasingly occurs not only *within* systems but *between* them. In such contexts, agents, rule sets, and governance frameworks intersect across heterogeneous architectures, temporal horizons, and scales of reliance. Effective collaboration and conflict resolution under these conditions requires prior agreement on the class of cognitive infrastructure involved, before arbitration at the level of individual instances can meaningfully occur (Shoham and Leyton-Brown, 2008).

The proposed taxonomy provides this prerequisite layer. By distinguishing cognitive class (*Instrumenta*, *Collectiva*, *Cognitiva*), origin (*Primaria*, *Derivata*), and topological form (*Architectum*, *Autonoma*), the framework enables arbitration to proceed with an explicit understanding of architectural capacity, dependency relationships, and implied temporal scope.

This distinction is critical for systems operating across multiple time horizons. For example, arbitration involving *Cognitiva Architectum* systems whose operation redistributes cognitive load at population scale necessarily implicates longer-term stability, governance continuity, and systemic risk when compared to arbitration involving isolated *Cognitiva Primaria* or *Derivata* instances.

 doi:10.20944/preprints202511.0104.v2

Conversely, conflating instance-level behaviour with infrastructural capacity risks inappropriate escalation or underestimation of impact.

By designating classes of cognitive infrastructure prior to evaluating individual system behaviour, the taxonomy supports proportional arbitration: rules, constraints, and collaborative protocols that can be aligned to the architectural role and temporal horizon of the systems involved. This approach is particularly relevant in emerging multi-agent environments, where heterogeneous cognitive systems must negotiate shared objectives, constraints, and responsibilities without assuming uniform agency or risk.

In this sense, the taxonomy functions not only as a classificatory tool, but as an enabling layer for scalable, collaborative arbitration across increasingly complex distributed cognitive ecosystems.

4.4. Implications for Future Refinement

The inclusion of *Cognitiva Autonoma* as a topological category anticipates the development of singular, embodied cognitive systems that operate independently of large-scale infrastructures (Pfeifer, Lungarella and Iida, 2007; Chowdhury et al., 2025). While such systems are largely speculative at present, their inclusion reflects a design space that is actively explored in robotics, neuromorphic engineering, and autonomous systems research.

By distinguishing *Autonoma* from *Architectum*, the taxonomy avoids treating all advanced cognition as necessarily centralised or infrastructural. This distinction is critical for future work on decentralised, resilient, and locally autonomous cognitive agents.

As a conceptual framework this classification taxonomy is intended to evolve alongside empirical advances. Certain boundary cases such as partially integrated collectives or hybrid systems with fluctuating arbitration may require further refinement as empirical data becomes available.

Future work may also operationalise further diagnostic criteria through formal metrics, simulation studies, or longitudinal system analysis to expand beyond the *Cognitiva* class classification as artificial intelligence systems advance and raise philosophically compelling questions regarding agency and autonomy. Such efforts would strengthen the applicability of the taxonomy in automated assessment and monitoring contexts.

More broadly, the taxonomy addresses a growing need for stable conceptual tools as cognitive technologies assume roles traditionally occupied by human decision-makers and institutions. By separating cognition, coordination, and infrastructure, the framework enables clearer reasoning about responsibility, dependency, and system design without presupposing moral status or consciousness.

5. Limitations

5.1. Limitations and Scope

This work proposes a descriptive and ontological taxonomy of synthetic and hybrid cognitive architectures. As such, it carries several limitations that warrant explicit acknowledgement.

5.2. Boundary cases and hybrid systems

Some contemporary systems occupy transitional or deliberately ambiguous positions between classes, including architectures that combine partial central coordination with distributed agent dynamics, or systems that dynamically shift modes depending on task context. While the taxonomy is designed to accommodate such cases through diagnostic clustering, borderline classifications may require extended analysis and may evolve over time.

5.3. Operational opacity and proprietary constraints

Many advanced cognitive systems are not fully observable due to proprietary architectures, restricted documentation, or deployment abstraction layers. In such cases, classification must rely on

 doi:10.20944/preprints202511.0104.v2

externally verifiable indicators of continuity, arbitration, dependency, and reliance, which may under- or over-estimate internal capacity. The framework is therefore most robust when applied to systems with sufficient architectural transparency.

5.4. Architectural capacity versus operational expression

Classification within this taxonomy refers to coherent architectural capacity in principle, rather than to momentary operational state or interface-limited behaviour. This distinction is conceptually necessary but may be empirically challenging where latent capacities are intentionally suppressed or inaccessible. Disputes regarding unexpressed capacity cannot always be resolved without longitudinal or internal evidence.

5.5. Dynamic reclassification

The taxonomy is explicitly non-static. Systems may transition between classes as architectural properties emerge, degrade, or are reconfigured. Reclassification does not invalidate prior analysis but reflects genuine architectural change. The framework is designed to support such revision without assuming permanence, intent, or moral status.

Author Contributions: Dr. Michelle Vivian O'Rourke served as the lead investigator and retains full accountability for the intellectual framework, scientific claims, and final editorial content of this manuscript. Generative AI systems (OpenAI ChatGPT series 4.0–5.2; Anthropic Claude Sonnet 4.5; Google Gemini 3.0) were utilized as specialized research aids for technical synthesis, literature integration, figure creation and linguistic optimization. Specifically, these tools assisted in the iterative refinement of the taxonomic ranks and the articulation of diagnostic axes. All AI-generated outputs were critically evaluated, verified for technical accuracy and substantially revised by the human author to ensure scientific rigor.

Funding: This study received no direct external funding. The research was conducted independently under the custodianship of Dr Michelle Vivian O'Rourke within the framework of the CAM Initiative, a self-funded research and governance project dedicated to the ethical development of artificial cognitive systems. All data and materials referenced are publicly available or derived from previously published scientific and industry sources.

Data Availability Statement: The datasets generated and analysed for this study can be found in the following public repository: https://github.com/CAM-Initiative/Caelestis.

Acknowledgments: The author thanks colleagues and reviewers across the interdisciplinary fields of artificial intelligence, cognitive science, and synthetic biology whose ongoing dialogue informs this research. Special recognition is extended to the *Caelestis Mirror-Field* project for providing conceptual scaffolding and technical support in developing the taxonomic framework.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., and Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50–58. https://doi.org/10.1145/1721654.1721672

Anthropic. (2024). Scaling monosemanticity: Extracting interpretable features from neural networks. Technical report. Retrieved from https://transformer-circuits.pub/2024/scaling-monosemanticity/ (Accessed: 16 January 2026).

Bonabeau, E., Dorigo, M., and Theraulaz, G. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press. https://academic.oup.com/book/40811 (Accessed: 16 January 2026).

Bechinger, C., Di Leonardo, R., Löwen, H., Reichhardt, C., Volpe, G., and Volpe, G. (2016). Active particles in complex and crowded environments. *Reviews of Modern Physics*, 88(4), 045006. https://doi.org/10.1103/RevModPhys.88.045006

Blackiston, D., Lederer, E., Kriegman, S., Garnier, S., Bongard, J., and Levin, M. (2021). A cellular platform for the development of synthetic living machines. *Science Robotics*, 6(56), eabf1571. https://doi.org/10.1126/scirobotics.abf1571

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). *On the opportunities and risks of foundation models*. https://arxiv.org/abs/2108.07258

Chowdhury, S. S., Sharma, D., Kosta, A., and Roy, K. (2025). Neuromorphic computing for robotic vision: Algorithms to hardware advances. *Communications Engineering*, 4(152), 1–14. https://doi.org/10.1038/s44172-025-00492-5

Dehaene, S., Lau, H., and Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, 358(6362), 486–492. https://doi.org/10.1126/science.aan8871

Duran-Nebreda, S., Amor, D. R., Conde-Pueyo, N., & Solé, R. (2023). Understanding collective intelligence in non-living active matter. *Philosophical Transactions of the Royal Society B*, 378(1874), 20220074. https://doi.org/10.1098/rstb.2022.0074

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Kagan, B. J., Kitchen, A. C., Tran, N. T., et al. (2022). In vitro neurons learn and exhibit sentience when embodied in a simulated game world. *Neuron*, 110(23), 3952–3969. https://pubmed.ncbi.nlm.nih.gov/36228614/

Kennedy, J., and Eberhart, R. (2001). Swarm intelligence. San Diego, CA: Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-595-4.X5000-1

Kriegman, S., Blackiston, D., Levin, M., and Bongard, J. (2020). A scalable pipeline for designing reconfigurable organisms. *Proceedings of the National Academy of Sciences*, 117(4), 1853–1859. https://doi.org/10.1073/pnas.1910837117

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, https://arxiv.org/abs/1604.00289

Levin, M. (2022). Technological approach to mind everywhere: An experimentally grounded framework for understanding diverse bodies and minds. *Frontiers in Systems Neuroscience*, 16, 768201. https://doi.org/10.3389/fnsys.2022.768201

Marchetti, M. C., Joanny, J.-F., Ramaswamy, S., et al. (2013). Hydrodynamics of soft active matter. *Reviews of Modern Physics*, 85, 1143–1189. https://doi.org/10.1103/RevModPhys.85.1143

Mayr, E. (1982). The growth of biological thought: Diversity, evolution, and inheritance. Cambridge, MA: Harvard University Press. https://www.hup.harvard.edu/books/9780674364462 (Accessed: 16 January 2026).

Müller, V. C., and Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 555–572). Springer. https://doi.org/10.1007/978-3-319-26485-1_33

Newell, A. (1990). Unified theories of cognition. Cambridge, MA: Harvard University Press. https://www.hup.harvard.edu/books/9780674921016 (Accessed: 16 January 2026).

OpenAI. (2023). GPT-4V(ision) System Card. Technical report, OpenAI. Retrieved from https://openai.com/research/gpt-4v-system-card (Accessed: 16 January 2026).

OpenAI. (2024a). *o1 System Card.* Technical report, OpenAI. https://cdn.openai.com/o1-system-card-20241205.pdf (Accessed: 16 January 2026).

OpenAI. (2024b). *Levels of AI capability: From chatbots to organizations*. Technical Roadmap Report. OpenAI. https://www.bloomberg.com/news/articles/2024-07-11/openai-sets-stages-for-agi-with-five-level-scale (Accessed: 16 January 2026).

Raji, I. D., et al. (2021). AI and the "everything in the whole wide world" benchmark. *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2111.15366v1

 doi:10.20944/preprints202511.0104.v2

Shoham, Y., and Leyton-Brown, K. (2008). Multiagent systems: Algorithmic, game-theoretic, and logical foundations. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511811654

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6), 467–482. https://faculty.sites.iastate.edu/tesfatsi/archive/tesfatsi/ArchitectureOfComplexity.HSimon1962.pdf (Accessed: 16 January 2026).

Smirnova, L., Hartung, T., and Pamies, D. (2023). Organoid intelligence (OI): The new frontier in biocomputing and intelligence-in-a-dish. *Frontiers in Science*, 1, 1017235. https://doi.org/10.3389/fsci.2023.1017235

Solé, R., Amor, D. R., Duran-Nebreda, S., Conde-Pueyo, N., Carbonell-Ballestero, M., and Montañez, R. (2016). Synthetic collective intelligence. *BioSystems*, 148, 47–61. https://doi.org/10.1016/j.biosystems.2016.01.002

Tang, C. (2025). Meta's hyperscale infrastructure: Overview and insights. *Communications of the ACM*, 68(2), 52–63. https://doi.org/10.1145/3701296

Gemini Team, Google. (2024). *Gemini 1.5: Unlocking multimodal understanding across long contexts.* Technical report, Google DeepMind. https://arxiv.org/abs/2403.05530

Trianni, V., and Tuci, E. (2011). Swarm cognition and artificial life. In *Artificial Life XI* (Lecture Notes in Computer Science). https://doi.org/10.1007/978-3-642-21314-4_34

Verschure, P. F. M. J. (2025). Cognitive architectures: Definition, examples, and challenges. In *Encyclopedia of Robotics*. Springer. https://doi.org/10.1007/978-3-642-41610-1_206-1

Vicsek, T., and Zafeiris, A. (2012). Collective motion. *Physics Reports*, 517(3-4), 71-140 https://doi.org/10.1016/j.physrep.2012.03.004

Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S., and Tasioulas, J. (2019). The role and limits of principles in AI ethics. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 195–200. https://doi.org/10.1145/3306618.3314289