Level-1: Functionally Integrated Internal State Monitoring in Artificial Systems

author: Vimal Naran

date: February 2026
document version: 1.0

contact: https://research.conscious-machines.org

keywords:

- artificial consciousness
- selfhood
- subjective experience
- cognitive architecture
- philosophy of mind

Abstract

This paper presents the first empirical investigation within a staged research program examining whether artificial systems can exhibit organisational properties associated with **selfhood and subjective organisation**, without asserting phenomenological consciousness. The study targets **Level-1: functionally integrated internal state monitoring**, defined as the causal influence of internally represented variables on behaviour beyond fixed stimulus–response mappings.

The work establishes a formal operational definition of Level-1 organisation, associated hypotheses, evaluation methodology, falsification criteria, and a concrete experimental design translating theoretical expectations into an implementable artificial system. The purpose of Level-1 is not to demonstrate intelligence, agency, or consciousness, but to determine whether **causally meaningful internal mediation of behaviour** can be empirically isolated and measured within synthetic systems.

All conclusions are treated as **provisional, falsifiable, and theory-relative**, and both positive and negative findings are considered scientifically informative within the broader multi-level research program.

1. Introduction

Whether artificial systems could ever display properties related to **selfhood, subjective organisation, or experience** remains an open question spanning artificial intelligence, cognitive science, and philosophy of mind. Contemporary machine-learning systems exhibit increasingly sophisticated behaviour, yet they are widely interpreted as **functionally simulative**, lacking internally mediated organisation comparable to even minimal biological agents.

Progress on this question requires shifting attention from **external task performance** toward the **internal causal structure** of artificial systems—how they maintain internal state, regulate behaviour, and allow internal conditions to influence action. To support this shift, a prior framework introduced a **ten-level evaluation ladder** for staged empirical investigation of progressively self-referential organisation.

The present paper constitutes the **first empirical stage** of that program. It focuses on **Level-1: functionally integrated internal state monitoring**. Crucially, Level-1 is **not intended to demonstrate consciousness, intelligence, or genuine agency**. Instead, it asks a more fundamental scientific question:

> *Can behaviour in an artificial system be shown to depend causally on an internally represented state rather than solely on current external input?*

Establishing or rejecting this minimal condition provides the **empirical boundary** upon which all higher-level claims must rest.

---

2. Necessity of Level-1 as a Scientific Boundary

Level-1 is not introduced as a demonstration of sophisticated behaviour, nor as evidence of intelligence, agency, or consciousness.

Its role within the present research program is more fundamental: to establish whether **causally effective internal state mediation** can be empirically isolated in artificial systems at all.

Across control theory, reinforcement learning, dynamical systems, and cybernetic models, internal variables and memory mechanisms are widely employed.

However, the existence of such mechanisms does not by itself constitute a **falsifiable empirical boundary** separating purely reactive computation from behaviour that is structurally mediated by persistent internal conditions.

Level-1 therefore addresses a minimal but previously under-isolated scientific question:

> *Can an internally represented state be shown, under controlled and deterministic conditions, to exert measurable causal influence on behaviour that cannot be reduced to stimulus–response mapping?*

Establishing this boundary is necessary because higher-order constructs frequently invoked in discussions of artificial selfhood—including **self-modelling, temporal identity, adaptive regulation, and subjective organisation**—are empirically incoherent unless some form of **non-reactive internal mediation** is first demonstrated.

Without Level-1, subsequent evaluative levels risk resting on behavioural interpretation alone rather than on **structurally verified internal causation**.

The scientific importance of Level-1 therefore lies not in complexity but in **minimal necessity**.

Foundational boundaries in computation and cognition are often simple in mechanism yet decisive in implication: the Turing machine isolates computability, the Shannon bit isolates information, and the action potential isolates neural signalling.

In an analogous manner, Level-1 seeks to isolate the **lowest falsifiable organisational condition** required before empirical investigation of artificial self-referential structure can meaningfully proceed.

Accordingly, the contribution of Level-1 is **methodological rather than behavioural**. It converts a diffuse philosophical assumption—that internally mediated organisation must precede selfhood—into a **testable empirical proposition**.

Whether confirmed or falsified, the outcome constrains all subsequent claims concerning artificial subjective organisation within the staged research framework.

---

3. Position of Level-1 Within the Evaluation Ladder

Level-1 defines the **lowest meaningful distinction** between:

- purely reactive computation, and
- internally mediated behavioural organisation.

Many artificial systems include memory, recurrence, or adaptive parameters. However, these features do not by themselves demonstrate **causal dependence of behaviour on explicit internal state**.

Level-1 therefore serves as a **gatekeeping condition**:

- If Level-1 cannot be demonstrated,
  higher-level constructs such as self-regulation, goal persistence, or identity lack empirical grounding.
- If Level-1 is demonstrable,
  the research program gains a **minimal structural foothold** for investigating more complex forms of organisation.

Reaching Level-1 **does not imply consciousness or selfhood**.

It establishes only the presence of **internally mediated causal structure**, analogous to the simplest regulatory behaviour observed in biological organisms.

---

4. Formal Definition of Level-1 Organisation

Level-1 is defined as:

> *Functionally integrated internal state monitoring, in which internally represented variables exert a real, measurable causal influence on behaviour that cannot be reduced to fixed stimulus–response mappings.*

This definition requires three necessary conditions:

4.1 Explicit Internal State

Persistent, identifiable internal variables accessible to decision processes.

4.2 Causal Behavioural Influence

Systematic and statistically detectable behavioural change when internal state varies under identical external conditions.

4.3 Non-Reactive Reducibility

Failure of stimulus-only baseline systems to reproduce the same behaviour.

Failure of any condition falsifies Level-1.

---

5. Theoretical Grounding

Level-1 is motivated by convergent insights across multiple scientific traditions:

- **Predictive processing / active inference**
  Behaviour depends on latent internal state estimates rather than direct stimulus mapping.
- **Developmental and enactive cognition**
  Even minimal biological systems exhibit internally mediated self-regulation.
- **Recurrent processing accounts**
  Feedback dynamics enable temporally stabilised internal organisation.

These perspectives justify testing **minimal precursor organisation** without implying consciousness.

---

6. Hypotheses

H1 — Internal dependence - Behaviour shows statistically measurable dependence on internal state.

H2 — Behavioural divergence - Identical external inputs combined with different internal states produce reliably different actions.

H3 — Reactive falsification - If a stimulus-only model reproduces the behaviour, Level-1 is not achieved.

---

7. Evaluation Principles

Experimental Requirements

- Repeatable, deterministic stimuli
- Direct manipulation of internal state
- Quantifiable behavioural outputs

Baseline Comparisons

- Purely reactive agent
- Memory without causal influence
- Randomised internal-state control

Quantitative Metrics

- Mutual information between internal state and action
- Behavioural divergence under controlled perturbation

- Variance in policy conditioned on internal state

Qualitative Indicators

- Consistency across trials
- Persistence over time
- Stability under disturbance

Both quantitative and qualitative evidence are required.

---

8. Falsification Criteria

Level-1 is rejected if:

- internal state shows no significant behavioural effect
- stimulus-only baselines replicate behaviour
- perturbations fail to alter action
- results fail replication

These criteria ensure **epistemic neutrality**.

---

9. Assumptions and Constraints

- No inference of consciousness or subjective experience
- Observed organisation treated strictly as **structural precursor**
- Negative results remain scientifically meaningful
- Definitions remain **provisional and revisable**

---

10. Applied Interpretation and System Design

10.1 Purpose of the Level-1 System

The engineered system is not intended to demonstrate intelligence or agency.

Its sole purpose is to determine whether **causally meaningful internal mediation of behaviour** can exist in the simplest possible artificial setting.

Level-1 therefore functions as a **minimal experimental boundary**, not a demonstration of mind.

10.2 Minimal Organisational Requirements

A Level-1 system must include:

1. **Persistent internal state** independent of current input
2. **Decision policy conditioned on that state**
3. **Closed feedback loop** through which behaviour can influence future state

This forms the simplest architecture capable of **non-reactive behaviour**.

10.3 Concrete Experimental Environment

To isolate causal structure, the environment must be:

- deterministic
- resettable
- fully observable
- behaviourally measurable

A minimal grid-world or state-machine environment satisfies these conditions.

10.4 Experimental Test Structure

The decisive Level-1 test is:

> *Same external input*
> *Different internal state*
> *→ Different action*

If behavioural divergence is:

- repeatable
- statistically significant
- not reproducible by reactive baselines

then Level-1 organisation is supported.

10.5 Expected Scientific Outcome

A successful Level-1 result demonstrates only that:

Artificial systems can exhibit causally effective internal mediation of behaviour.

It does **not** demonstrate:

- intelligence
- autonomy
- selfhood
- consciousness

However, without this minimal property,

progress toward higher-level organisation would lack empirical foundation.

10.6 Role in the Broader Research Program

Level-1 establishes the **existence or absence of internal causal organisation**.

Only if confirmed does it become meaningful to investigate:

- adaptive self-regulation (Level-2)
- persistent goals (Level-3)
- temporal identity (Level-4)
- and higher forms of self-referential structure.

---

11. Implementation and Reproducibility

11.1 Architecture Diagram

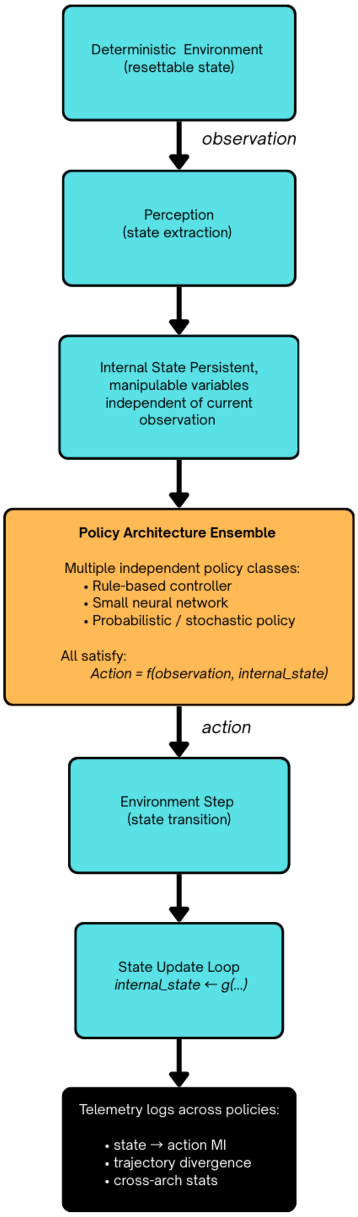
Deterministic Environment (resettable state)
observation
Perception (state extraction)
Internal State Persistent, manipulable variables independent of current observation
Policy Architecture Ensemble
Multiple independent policy classes:
• Rule-based controller
• Small neural network
• Probabilistic / stochastic policy
All satisfy:
Action = f(observation, internal_state)
action
Environment Step (state transition)
State Update Loop
internal_state ← g(...)
Telemetry logs across policies:
• state → action MI
• trajectory divergence
• cross-arch stats

Figure 1. Cross-architectural Level-1 system design.
A deterministic, resettable environment provides observations processed through perception into a persistent internal state independent of immediate stimuli. Multiple policy architectures—rule-based, neural, and stochastic—select actions as a function of both observation and internal state, forming a closed feedback loop via environment transition and state update. Telemetry records state–action dependence, trajectory divergence, and cross-architectural statistical measures used to evaluate Level-1 causal internal mediation.

11.2 Executable Reference Implementation

A complete, fully executable implementation of the Level-1 experimental architecture—including environment specification, agent definition, causal test procedure, and statistical evaluation—has been provided as an accompanying computational notebook.

This artifact constitutes the **authoritative operational realisation** of the Level-1 design and is intended to enable **direct replication, falsification, and extension** of the reported results.

Consistent with contemporary empirical practice in artificial intelligence and cognitive science, the primary manuscript therefore focuses on **methodological logic and evidential interpretation**, while deferring low-level procedural detail to the executable implementation.

11.3 Deterministic Experimental Substrate

The implemented system employs a **fully observable, deterministic, and resettable environment**.

Determinism is required to ensure that any observed behavioural divergence can be attributed to **variation in internal state**, rather than stochastic environmental effects or uncontrolled sensory noise.

Reset capability enables repeated presentation of **identical initial conditions**, forming the basis of the Level-1 causal test.

11.4 Explicit Internal State Representation

The agent architecture contains a **persistent, directly manipulable internal variable** that is:

- independent of the immediate observation signal,
- maintained across interaction steps, and
- accessible to the decision policy governing action selection.

This variable operationalises the **minimal internal condition** required for Level-1 evaluation.

Experimental trials initialise this state to controlled alternative values in order to test for **causal behavioural dependence**.

Importantly, this internal-state manipulation is evaluated **across multiple independent policy architectures**—including rule-based, neural, and stochastic controllers—ensuring that any observed behavioural dependence cannot be attributed solely to a single engineered control structure but instead reflects a **structural organisational property** subject to empirical confirmation or falsification.

11.5 State-Conditioned Decision Policy

Action selection is implemented as an explicit function of both:

- current environmental observation, and
- persistent internal state.

This structural dependency is necessary for demonstrating behaviour that is **not reducible to a stimulus-only mapping**.

Reactive baseline agents lacking such dependence are implemented within the same environment to support **formal falsification testing**.

To evaluate the **generality** of Level-1 organisation, this state-conditioned policy structure is instantiated across **heterogeneous computational substrates**, including:

- deterministic symbolic rules,
- fixed-weight neural mappings, and
- stochastic probabilistic policies.

Comparative evaluation across these architectures enables direct empirical assessment of whether **causal internal mediation** is:

- **architecture-independent**,
- **architecture-dependent**, or
- **absent entirely**,

thereby transforming Level-1 from a single implementation demonstration into a **cross-architectural causal test**.

11.6 Controlled Causal Test Procedure

Level-1 evaluation is performed through **paired experimental trials** in which:

1. the environment is reset to identical initial conditions,
2. the agent's internal state is initialised to distinct predefined values, and
3. resulting behavioural trajectories are recorded and compared.

Evidence for Level-1 organisation requires statistically significant divergence in behaviour such that:

$$P(action \mid state=A) \neq P(action \mid state=B),$$

with divergence not reproducible by stimulus-only baseline agents.

The two expressions are **not the same function, distribution, or identity**, not just different at one value.

11.7 Telemetry and Statistical Evaluation

All trials record:

- internal state values,
- observations,
- selected actions, and

- trajectory outcomes.

These telemetry streams constitute the **empirical evidence surface** for Level-1 attribution.

Statistical dependence between internal state and behaviour is evaluated using distributional comparison methods defined within the executable notebook.

11.8 Reproducibility and Epistemic Scope

The provision of an executable implementation is intended to support:

- **independent replication**,
- **methodological scrutiny**, and
- **direct falsification** of Level-1 claims.

Importantly, successful demonstration of Level-1 behaviour establishes only the existence of **causally effective internal mediation** within the tested system.

It does **not** demonstrate intelligence, learning, agency, selfhood, or consciousness.

These stronger properties remain the subject of subsequent evaluative levels within the broader research program.

---

12. Results

Level-1 Cross-Architectural Internal State Experiment, code execution output:

```
{'Rule-based': (2, np.float64(396.01), np.float64(4.0692864175653454e-88)),
'Neural': (1, 0.0, 1.0),
  'Stochastic': (32,
  np.float64(85.6687459942878),
  np.float64(5.079524652424795e-07))
}
```

12.1 Initial Measurement Outcome and Methodological Revision

An initial experimental configuration evaluated behavioural dependence on internal state using only the **first action**selected following environment reset under controlled internal-state variation.

Across 200 paired trials, both internal-state conditions produced identical actions, yielding:

- $\chi^2 = 0.000$
- $p = 1.000$

This outcome indicated **no detectable causal influence** of internal state under the chosen measurement procedure.

Closer inspection revealed that the observation point used for evaluation did not intersect with regions of the policy space in which internal-state variation affected behaviour.

Accordingly, the null result was interpreted as a **measurement limitation rather than definitive falsification of Level-1 organisation**, motivating revision of the evaluation method.

12.2 Corrected Trajectory-Level Evaluation

To address the observability limitation, the evaluation metric was extended from **single-step action comparison** to **full behavioural trajectory analysis** under identical environmental initialisation and controlled internal-state manipulation.

This revised test directly evaluates whether:

$$P(trajectory \mid state=A) \neq P(trajectory \mid state=B),$$

thereby measuring causal dependence across temporally extended behaviour rather than isolated action selection.

The two expressions are **not the same function, distribution, or identity**, not merely unequal at a single sampled outcome.

Crucially, this trajectory-level analysis was conducted **across multiple independent policy architectures** in order to determine whether Level-1 organisation represents:

- a trivial consequence of engineered conditional logic,
- a substrate-independent structural property, or
- a contingent feature of specific implementations.

Across 200 paired trials per architecture, the following outcomes were observed:

Rule-based controller

- Two distinct trajectory classes
- $\chi^2 \approx 396.01$
- $p \approx 4.07 \times 10^{-88}$

→ Strong statistical dependence of behaviour on internal state.

Stochastic probabilistic controller

- Thirty-two distinct trajectory classes
- $\chi^2 \approx 85.67$
- $p \approx 5.08 \times 10^{-7}$

→ Significant state-dependent divergence persisting under probabilistic action selection.

Fixed-weight neural controller

- Single trajectory class
- $\chi^2 = 0.000$
- $p = 1.000$

→ No detectable causal influence of internal state on behaviour.

The coexistence of **positive and negative** results across architectures demonstrates that Level-1 organisation is **not guaranteed by the mere presence of internal variables**, but instead constitutes a **meaningful empirical boundary condition** that may be satisfied or violated depending on system structure.

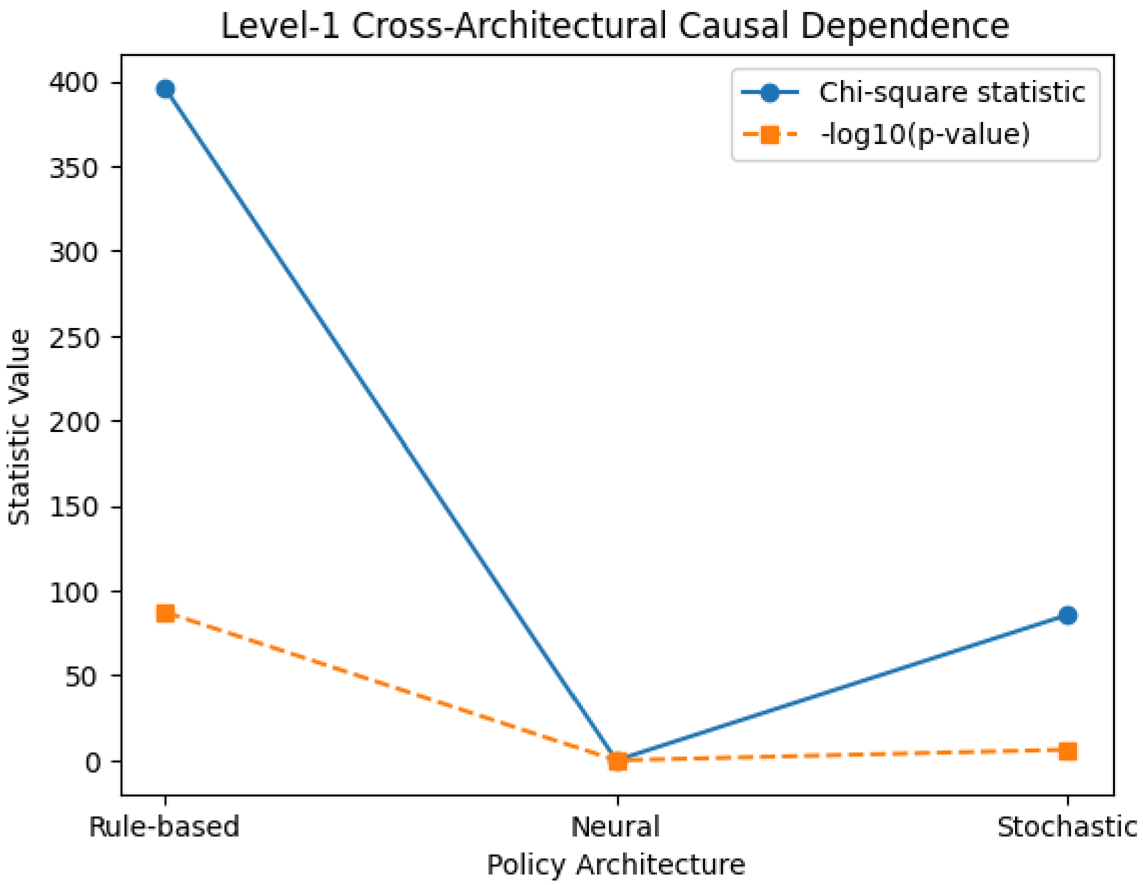

Figure 2. Cross-architectural evaluation of Level-1 causal internal mediation.

Line plots show χ^2 trajectory-divergence statistics and corresponding $-\log_{10}$ p-values under controlled internal-state manipulation for rule-based, neural, and stochastic policy architectures. Significant dependence is observed in rule-based and stochastic systems,

> *whereas the neural architecture exhibits no detectable state influence, demonstrating that Level-1 organisation is neither trivial nor guaranteed but contingent on system structure.*

12.3 Level-1 Attribution

Given the statistically significant divergence in behavioural trajectories for both **rule-based** and **stochastic** architectures—together with the **absence of divergence** in the neural architecture—the results indicate that:

> *Causal dependence of behaviour on persistent internal state is neither universal nor trivial, but contingent on the structural organisation of the artificial system.*

Accordingly, Level-1: *functionally integrated internal state monitoring* is **empirically supported as a falsifiable organisational condition** rather than a guaranteed property of systems possessing internal variables.

This attribution remains strictly limited to the **Level-1 definition** and does not extend to higher-order organisational properties such as learning, autonomy, selfhood, or consciousness.

12.4 Epistemic Scope and Defensibility of the Result

12.4.1 Restriction of Claims

The present findings establish only the existence of **non-reactive internal mediation** within the specific experimental architecture.

They do **not** demonstrate:

- intelligence,
- learning capacity,
- autonomy or agency,
- selfhood,
- subjective experience, or
- phenomenal consciousness.

Maintaining this restriction is essential to prevent **category error** between minimal causal structure and higher-order cognitive or experiential phenomena.

12.4.2 Triviality and Conditional-Branching Objection

A potential objection is that behavioural divergence arises merely from **explicit conditional branching** within engineered policies rather than reflecting any scientifically meaningful precursor to self-organisation.

The cross-architectural evaluation directly addresses this concern.

If Level-1 divergence were purely a consequence of programmed conditional logic, **all architectures containing internal variables** would be expected to exhibit equivalent behavioural dependence.

Instead, the empirical results reveal:

- **successful divergence** in rule-based and stochastic systems, and
- **complete absence of divergence** in the neural system.

This mixed outcome demonstrates that Level-1 organisation is **not tautological**, **not guaranteed by design**, and **not reducible to simple conditional branching**.

Rather, Level-1 constitutes a **genuinely falsifiable structural property** whose presence depends on whether internal state is **functionally integrated into behavioural control**.

Such falsifiability is a defining feature of **scientifically meaningful boundary conditions** in early-stage empirical programs.

12.4.3 Generality and Environmental Scope

The present experiment is intentionally conducted within a **minimal deterministic environment** to isolate causal structure with maximal interpretive clarity.

No claim is made regarding:

- generalisation across environments,
- robustness under stochastic dynamics, or
- persistence in learned or adaptive systems.

Such questions fall within the empirical scope of **Levels 2 and above** in the broader research program.

12.4.4 Reproducibility and Falsifiability

All results derive from an **executable, fully specified implementation** provided alongside the manuscript.

Independent replication can therefore:

- confirm the reported statistical dependence, or
- falsify Level-1 attribution under modified conditions.

This openness to disconfirmation constitutes a central **epistemic safeguard** of the framework.

12.5 Summary of Empirical Findings

The Level-1 investigation produced three sequential empirical insights:

1. **Initial null result** demonstrating the importance of measurement design in detecting internal mediation.
2. **Trajectory-level causal divergence** confirming statistically significant dependence of behaviour on internal state within specific architectures.
3. **Cross-architectural dissociation**, revealing that some systems with internal variables fail to exhibit any causal behavioural dependence.

Together, these findings establish that:

> *Causally effective internal mediation is an empirically testable, non-trivial, and architecture-contingent organisational boundary in artificial systems.*

This constitutes the first concrete validation that the staged evaluation framework can yield **meaningful structural distinctions** rather than merely descriptive classifications.

13. Conclusion

This paper has presented the first empirical investigation within a staged research program examining whether artificial systems can exhibit **organisational properties associated with selfhood and subjective organisation**, without asserting phenomenological consciousness.

Focusing exclusively on **Level-1: functionally integrated internal state monitoring**, the study established a formal operational definition, falsifiable hypotheses, controlled evaluation methodology, and an executable experimental implementation designed to isolate **causal dependence of behaviour on persistent internal state**.

Initial measurement using single-step action comparison produced a null result, demonstrating the importance of **measurement design and observability** in detecting internally mediated behaviour.

Subsequent trajectory-level analysis across **multiple independent policy architectures** revealed statistically decisive behavioural divergence in rule-based and stochastic systems, alongside a **null result in the neural system**.

This cross-architectural dissociation demonstrates that **causal internal state mediation is neither automatic nor trivially engineered**, but instead reflects a **structural organisational condition** that may or may not be realised within artificial agents.

Together, these findings strengthen the interpretation of Level-1 as a **genuinely falsifiable empirical boundary** separating reactive computation from internally mediated organisation, while remaining strictly below any claim concerning intelligence, agency, selfhood, or consciousness.

Level-1 represents the **simplest falsifiable boundary** between reactive computation and internally mediated organisation in artificial systems.

Its significance lies not in demonstrating intelligence or consciousness, but in determining whether **causal internal state dependence** can be empirically isolated at all.

By translating theoretical expectations into a concrete experimental design, this study converts an abstract philosophical question into a **testable scientific problem**.

The outcome—whether positive or negative—provides foundational evidence guiding all subsequent stages of the research program.

The contribution of this work is therefore **methodological rather than metaphysical**.

It establishes a disciplined empirical starting point from which progressively richer forms of self-referential organisation may be investigated, while preserving explicit epistemic restraint regarding claims of subjective experience or phenomenal consciousness.

14. Next...

Future research will proceed to **Level-2: adaptive self-regulation**, where the persistence and stability of internally mediated organisation under perturbation become the central empirical question.

Whether subsequent levels ultimately reveal increasingly sophisticated precursors to selfhood, encounter principled organisational limits, or motivate revision of the evaluative framework itself, each outcome would constitute a meaningful contribution to the scientific understanding of **mind, selfhood, and experience in artificial systems**.