Level-2: Adaptive Self-Regulation

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 defines and designs a Level-2 evaluation within a staged research program on machine self-organisation. Level-2 tests adaptive self-regulation: whether a system can maintain functional stability when conditions change, by using internal state or parameter updates rather than relying on a fixed stimulus-response mapping.

Level-1 showed that some systems can exhibit internal state mediation—meaning internal variables can causally influence behavior. However, internal mediation alone does not imply regulation. A system may have state, memory, or recurrence and still fail when the environment shifts.

Level-2 therefore introduces a stronger requirement: the system must keep a key variable stable under perturbation, and it must do so by internally updating how it interprets the world or how it chooses actions. This paper provides a clear operational definition, hypotheses, falsification criteria, evaluation methodology, and a concrete experimental design. It stops before implementation.

1. Introduction

There is ongoing debate about whether artificial systems could ever show properties associated with selfhood, phenomenal consciousness, or subjective experience. Many modern systems produce complex behavior, but complex output alone is not enough to infer deeper organisation, or agency. For scientific progress, we need tests that focus on internal causal structure—how internal variables influence action, and whether those variables support stable functioning.

This research program takes a staged approach. Each stage defines one organisational property, makes it measurable, and specifies how it can be falsified. This paper addresses Level-2, which asks a focused question:

Can an artificial system preserve stability when the world changes, by adjusting something inside itself—rather than only reacting with a fixed mapping from input to action?

Level-2 is meant to be minimal, testable, and repeatable. It does not claim consciousness. It does not aim to show general intelligence. It aims to isolate a specific capacity that is common in even simple biological systems: staying functional under changing conditions.

2. Why Level-2 is needed

Many agents can be built with internal state: memory buffers, recurrent networks, hidden variables, or ongoing traces of past inputs. Level-1 tests whether such internal state is causally relevant. But Level-1 does not require that internal state is used to keep the system stable.

A system can pass Level-1 while still being fragile. For example:

- It may produce different actions depending on internal state.
- Yet it may still collapse when the environment dynamics change.

Level-2 addresses this gap. It sets a boundary between:

- Internal state that exists and influences behavior, versus
- Internal state that updates to preserve stability when conditions shift.

This boundary matters because later levels—such as persistent goal pursuit (Level-3)—are difficult to interpret unless the system can already maintain basic stability under perturbation.

3. Relation to adjacent levels

3.1 Relation to Level-1

Level-1 asks: *Is behavior internally mediated at all?*
It is mainly a detection and non-reducibility test: can we show internal state makes a causal difference that cannot be replicated by a purely reactive controller?

Level-2 asks: *Does internal mediation do stabilising work under change?*
It is a functional and causal-necessity test: does internal updating preserve stability when dynamics shift?

So Level-2 builds on Level-1, but it is stricter: it requires that internal state is not only present and causal, but also adaptively useful.

3.2 Relation to Level-3

Level-3 is about goal persistence: behaviour stays oriented toward an internally represented target across delay or interference.

Level-2 does not require:

- explicit goal representations,
- delayed goal maintenance,
- persistence through distraction.

Instead, Level-2 requires:

- keeping the system within viability or stability bounds,
- using internal updates to respond to changes in dynamics.

Level-2 may be a prerequisite for Level-3, but it does not imply Level-3.

---

4. Level-2 definition

Level-2: Adaptive Self-Regulation

A system exhibits Level-2 organisation if it:

1. faces perturbations that change the environment dynamics,
2. preserves functional stability under those perturbations, and
3. achieves this stability through internally mediated state or parameter adjustment, not through a fixed stimulus-response policy alone.

This definition requires more than “feedback.” Many reactive controllers use feedback. Level-2 requires adaptive adjustment: internal variables must change in a way that helps the system recover and remain stable across a family of perturbations.

---

5. Operational requirements

To claim Level-2, the evaluation must satisfy the following requirements.

5.1 A stability variable and stability criterion

The system must be evaluated against a clearly defined stability requirement, such as:

- keeping a “vital” variable within bounds,
- returning to a safe range after disturbance,
- minimising long-run deviation from a setpoint.

This must be measurable across repeated trials.

5.2 Perturbations must change the dynamics (not just add noise)

Perturbations should alter how actions affect the system, such as:

- actuator gain changes (actions become weaker/stronger),
- leakage changes (system drifts faster/slower),
- sign flips (actions reverse effect),
- hidden mode switching between deterministic regimes.

If perturbations only add random noise, a robust reactive policy may appear adaptive. Level-2 needs perturbations that make “the rules of the world” shift.

5.3 Perturbation source must be partly hidden

If the agent can directly observe the perturbation label (e.g., “mode=2”), a reactive mapping can solve the task by lookup.

So the agent should observe the system state, but *not* the perturbation parameter directly. The agent must infer changes from mismatch over time (e.g., prediction error).

5.4 Internal updating must be necessary for robustness

Level-2 requires causal evidence that internal adjustment is doing work. This is tested via ablations:

- freeze internal updates,
- remove internal state from policy input,
- scramble updates.

If performance does not drop when adaptation is removed, Level-2 is not supported.

6. Theoretical grounding

Level-2 draws on several aligned traditions:

- **Cybernetics**: regulation is a system property that counters disturbances through feedback and internal organisation.

- **Adaptive control**: stability is preserved through online internal updates to controller parameters or model estimates.

- **Homeostasis / allostasis**: stability is maintained either by holding variables near a setpoint (homeostasis) or by changing regulatory settings under shifting conditions (allostasis).

- **Predictive processing (conceptual link)**: systems can stabilise behavior by updating internal models to reduce prediction error.

These traditions motivate why Level-2 is a meaningful scientific boundary without implying consciousness or agency.

7. Hypotheses

H1: Robust stability under perturbation - Under defined perturbations, the candidate system maintains stability significantly better than optimised reactive baselines.

H2: Adaptive internal signature - Perturbations produce systematic changes in internal state or parameters that track recovery and stability outcomes.

H3: Causal necessity of internal updating - If internal updating is disabled (updates frozen), stability performance falls toward baseline.

H4: Non-reducibility to reactive control - If a stimulus-only controller can match performance across the perturbation family, Level-2 is not supported.

8. Evaluation approach

8.1 Experimental structure

A Level-2 experiment should be:

- **deterministic or tightly controlled**, to support repeatability and causal interpretation,
- **resettable**, so identical conditions can be tested across systems,
- **trajectory-based**, since regulation is expressed over time.

8.2 Metrics

Primary stability metrics:

- **violation rate** (fraction of trials leaving safe bounds),
- **recovery time** (in the applied evaluation: steps after the first post-switch violation until return to the viability band for *N* consecutive steps, *N=5*),
- **integrated deviation** (area under the error curve relative to a safe band or setpoint).

Secondary "adaptation evidence" metrics:

- magnitude/structure of internal state changes after perturbation,
- correlation between internal updates and recovery,
- sensitivity to update freezing.

8.3 Baselines

To avoid false positives, baselines should include:

- **Optimised reactive policy**: observation → action only, trained/tuned to the same distribution.
- **Fixed-state controller**: has internal state but does not update it adaptively (or updates do not depend on prediction mismatch).
- **Random update control**: update magnitude matches the adaptive system but direction is random.

If the baseline is weak, "winning" is meaningless.

8.4 Ablations (causal tests)

A Level-2 claim requires that adaptation is *necessary*, not cosmetic.

Core ablations used in the present evaluation:

1. **Freeze updates**: stop state/parameter updating during perturbation regime.
2. **Scramble updates**: apply random updates with matched magnitude.

A policy–state decoupling ablation (policy ignores $\hat{\theta}_t$ while updating continues) is recommended and planned. If performance does not meaningfully degrade under these ablations, the system has not demonstrated Level-2.

9. Falsification criteria

Level-2 is rejected if any of these conditions holds:

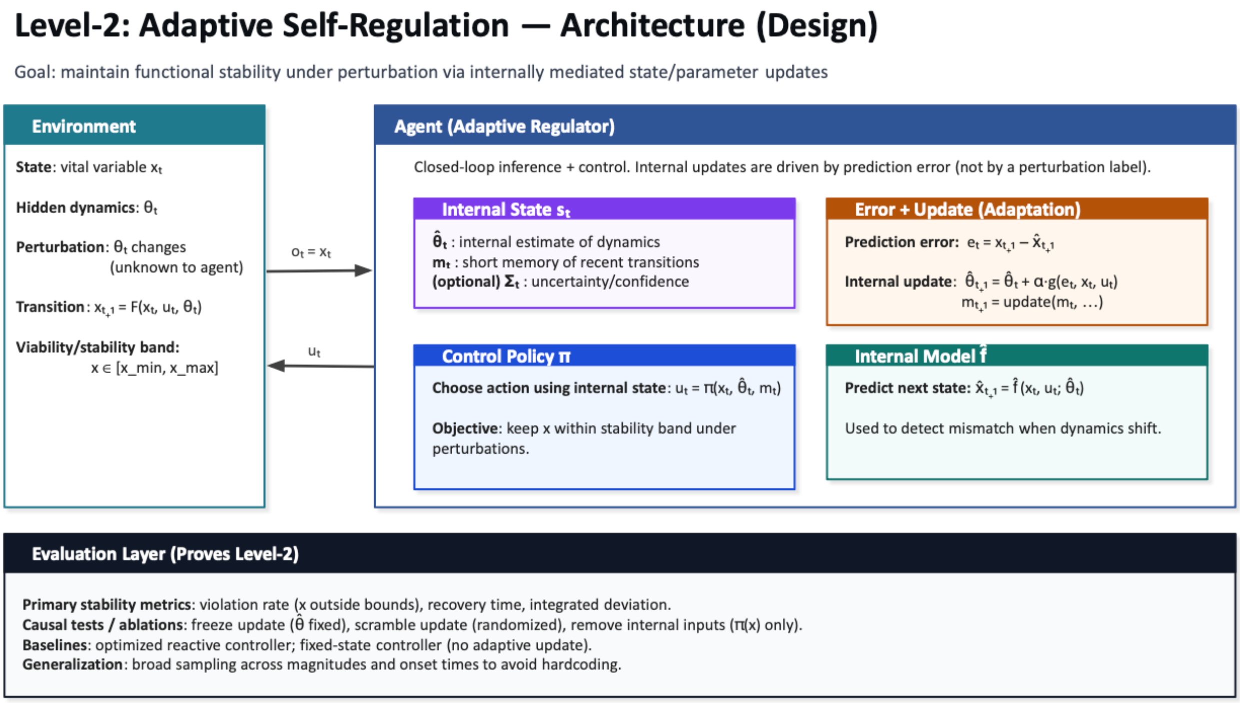
- stability is not better than strong reactive baselines under perturbation,
- internal updates occur but are not linked to recovery or stability,

- freezing updates does not reduce stability performance,
- a reactive controller matches performance across perturbation families,
- the result does not replicate under controlled reruns.

Negative results are still valuable: they clarify which architectures or mechanisms fail to produce adaptive regulation.

10. System design

This section specifies a design that can realistically test Level-2 while avoiding hardcoded solutions.



Level-2: Adaptive Self-Regulation — Architecture (Design)

Goal: maintain functional stability under perturbation via internally mediated state/parameter updates

Environment

State: vital variable x_t

Hidden dynamics: θ_t

Perturbation: θ_t changes (unknown to agent)

Transition: $x_{t+1} = F(x_t, u_t, \theta_t)$

Viability/stability band: $x \in [x\_min, x\_max]$

Agent (Adaptive Regulator)

Closed-loop inference + control. Internal updates are driven by prediction error (not by a perturbation label).

Internal State s_t

$\hat{\theta}_t$: internal estimate of dynamics
m_t : short memory of recent transitions
(optional) Σ_t : uncertainty/confidence

Error + Update (Adaptation)

Prediction error: $e_t = x_{t+1} - \hat{x}_{t+1}$

Internal update: $\hat{\theta}_{t+1} = \hat{\theta}_t + \alpha \cdot g(e_t, x_t, u_t)$
$m_{t+1} = \text{update}(m_t, \ldots)$

Control Policy π

Choose action using internal state: $u_t = \pi(x_t, \hat{\theta}_t, m_t)$

Objective: keep x within stability band under perturbations.

Internal Model $\hat{f}$

Predict next state: $\hat{x}_{t+1} = \hat{f}(x_t, u_t; \hat{\theta}_t)$

Used to detect mismatch when dynamics shift.

Evaluation Layer (Proves Level-2)

Primary stability metrics: violation rate (x outside bounds), recovery time, integrated deviation.
Causal tests / ablations: freeze update ($\hat{\theta}$ fixed), scramble update (randomized), remove internal inputs ($\pi(x)$ only).
Baselines: optimized reactive controller; fixed-state controller (no adaptive update).
Generalization: broad sampling across magnitudes and onset times to avoid hardcoding.

Figure 1 summarizes the proposed Level-2 architecture.

The system is evaluated in a perturbed dynamical environment and must maintain stability via internally mediated state/parameter updates driven by prediction error.

10.1 Design goals

The design should:

1. make perturbations genuinely destabilising,
2. prevent trivial reactive lookup solutions,
3. require internal inference and adjustment,
4. allow strong causal ablations, and
5. generalise beyond narrow scripted cases.

10.2 Environment design: regulated variable with hidden dynamics shifts

The Level-2 environment is intentionally minimal. Its purpose is not to model a rich world, but to create a controlled setting in which stability under change can be measured and attributed to internal adaptation rather than to a fixed stimulus→response mapping.

The environment exposes a single regulated (vital) variable xt that the agent must keep within a predefined viability/stability band:

$$x_t \in [x_{\min}, x_{\max}].$$

At each time step the agent selects an action *ut*, which influences *xt*. Critically, the mapping from action to state change depends on a hidden dynamics parameter *θt* (not observed by the agent). A perturbation occurs when *θt* changes during an episode at an unknown time *tp*. This means that an action that previously had a certain effect can become stronger, weaker, or qualitatively different after the perturbation, making fixed control strategies unreliable.

In the applied evaluation reported in Section 11, the perturbation is implemented as a hidden dynamics switch together with a post-switch bias term that begins pushing the system away from its setpoint. The resulting dynamics are:

$$x_{t+1} = x_t + \theta_t u_t - \lambda(x_t - x_{\text{env}}) + d_t.$$

where λ is a leak term toward the environmental attractor *xenv*. The perturbation structure is:

- $\theta_t = \theta_{\text{pre}}$ for $t < t_p$, and $\theta_t = \theta_{\text{post}}$ for $t \geq t_p$.
- $d_t = 0$ for $t < t_p$, and $d_t = d$ for $t \geq t_p$.

Both *θt* and dd are unobserved. The agent receives only observations derived from *xt*, ensuring that the perturbation must be inferred indirectly from its consequences over time (e.g., via prediction mismatch), rather than from a directly visible regime label.

This environment is designed so that the perturbation changes the underlying dynamics in a way that can systematically threaten stability while remaining unobserved. As a result, a single fixed observation→action mapping is not, in general, sufficient across regimes; successful performance requires internally mediated adjustment in response to mismatch.

10.3 Perturbation family (avoid hardcoding)

Perturbations are sampled from a family:

- onset time varies,

- magnitude varies,
- type varies (gain/leak/sign/mode).

Evaluation samples broadly across magnitudes, onset times, and perturbation types within the defined perturbation family, reducing sensitivity to any single scripted perturbation.

This prevents “memorise a single perturbation” solutions.

10.4 Agent design: internal estimation + internal updating + control

The Level-2 candidate agent is defined by its closed-loop structure, not by any particular programming implementation. The agent receives only the observable state (e.g., the current value of the regulated variable *xt*) and must maintain stability when environment dynamics change in a way that is not directly signaled.

To do this, the agent maintains a compact internal regulatory representation (e.g., an estimate of the current regime or dynamics sensitivity). This internal representation is not merely a memory trace: it is an adaptive variable intended to change when the agent detects that its expectations about the environment no longer match what is observed.

Updating is driven by an internally computed mismatch signal, such as a prediction error between (i) the next state that the agent expects under its current internal representation and chosen action and (ii) the state that actually occurs. When the environment’s hidden dynamics shift, this mismatch signal increases, triggering internal adjustment. As mismatch decreases, internal adjustment should reduce or stabilize, reflecting successful accommodation of the changed dynamics.

Crucially, the agent’s control policy must be coupled to this internal representation. Actions are not selected only from the current observation *xt*, but also from the current internal regulatory variable. This coupling ensures that internal adjustment is not epiphenomenal: changes in internal state alter action selection and therefore can causally support stability recovery.

This structure is compatible with multiple realizations (symbolic, probabilistic, or neural). The Level-2 attribution depends only on the presence of (i) an internal regulatory variable, (ii) an online update process driven by mismatch, and (iii) control decisions conditioned on that internal variable, together with the ablation tests specified in Section 10.12.

10.5 Why this design avoids hardcoding

This design resists hardcoding because:

- perturbations vary and are not directly labeled,
- the agent must infer change from observed consequences,
- success requires trajectory-level adjustment,
- broadly sampled perturbations from the evaluation family test generality,
- ablations can prove the adaptive mechanism is doing the work.

10.6 Expected outcome patterns

If Level-2 is supported, we expect:

- lower violation rates and faster recovery than reactive baselines,
- internal estimates *θ^t* shift after perturbation in a structured way,
- freezing internal updates causes a clear stability collapse,
- reactive baselines cannot match performance across the full perturbation family.

10.7 Why this architecture is *specifically* Level-2

Level-1 can be satisfied by *any* system where internal state influences action. That is necessary, but not sufficient.

This Level-2 architecture adds two stronger requirements:

1. **The environment changes in a way that breaks fixed mappings.**
   A purely reactive controller can be tuned to stationary dynamics. Level-2 requires conditions where "what works" changes mid-episode.

2. **The system must internally adjust to recover stability.**
   The internal estimate *θ^t* (and/or regulatory memory *mt*) must update in response to mismatch. In other words, state is not only *present*, it is *used to adapt*.

So the architecture is designed to make adaptation a **causal necessity**, not an optional feature.

10.8 What counts as "internally mediated adjustment"

The Level-2 definition allows either (or both) of the following:

A) Internal state adjustment

The agent updates internal variables that influence policy (e.g., *mt*, a belief state, an integrator, a regime estimate). This is "state" in the control-theoretic sense: internal variables that summarize history and change how future inputs are interpreted.

B) Internal parameter adjustment

The agent updates parameters that affect its predictions or actions (e.g., an online estimate *θ^tθ^t*, adaptive gains, model parameters). This can be a light-weight system identification update or an online learning step.

Important boundary: updating must be driven by the agent's own computations (e.g., prediction error or stability error), not by a hand-coded "mode switch" rule triggered by a directly observed label.

10.9 Design principle: perturbations must be “learnable but not directly visible”

A common failure mode in regulation experiments is that the perturbation is either:

- **too obvious** (agent sees a label and branches), or
- **too chaotic** (noise dominates, no systematic update helps), or
- **too weak** (reactive controller handles it).

The perturbation family must sit in the middle:

- It should create systematic prediction mismatch that can be reduced by updating *θ^t* or *mt*.
- It should not be fully identifiable from a single observation.

This is why the architecture centers on:

- a predictor *x^t+1*,
- a mismatch signal *et*, and
- an update rule driven by that mismatch.

10.10 “Hardcoding avoidance” requirements

To ensure the experiment demonstrates Level-2 and not a disguised lookup table, the study design must satisfy these constraints:

1. **Perturbations are drawn from a distribution**
   Onset time, magnitude, and type should vary across episodes.

2. **Evaluation covers broad perturbation variation**
   The agent must be tested on evaluation samples broadly across magnitudes, onset times, and perturbation types within the defined perturbation family.

3. **No observation channel directly reveals the perturbation**
   No “mode bit,” no “gain value,” no “shift occurred” flag.

4. **Baselines are strong**
   A weak reactive baseline lets you “win” without demonstrating anything structural. The reactive baseline should be optimized on the same training distribution.

These constraints make “hardcoding” fragile and force the system toward genuine internal adaptation.

10.11 Required baselines

The evaluation must include at least the following baseline classes:

Baseline 1 — Optimized reactive controller

A stimulus-only policy: $ut=\pi(xt)$

This baseline should be tuned or trained for the same perturbation distribution.

Purpose: tests whether performance can be explained by a fixed mapping.

Baseline 2 — Fixed-state controller (non-adaptive internal state)

A controller that has internal state but does not adjust it in a way that tracks mismatch. Examples:

- memory is present but not updated from prediction error,
- parameters are fixed,
- recurrent state exists but is frozen or irrelevant.

Purpose: separates “state exists” from “state adapts.”

Baseline 3 — Random update controller (matched update magnitude)

This baseline performs internal updates with a similar magnitude as the candidate, but the direction is randomized or decoupled from mismatch signals.

Purpose: rules out the explanation that “any internal drift” gives robustness.

10.12 Causal Tests (ablations)

To claim Level-2, the adaptive mechanism must be shown to be **necessary**, not decorative. Minimum ablations:

Ablation A — Freeze internal updating

Clamp $\theta\text{^}t+1=\theta\text{^}t$ (and/or freeze mt updates).

Expected Level-2 pattern: stability degrades substantially relative to the Adaptive (Level-2) system and approaches baseline-level performance (reactive and/or fixed-parameter control).

Ablation B (policy–state decoupling; recommended):

the internal estimator continues updating $\hat{\theta}_t$, but the policy is forced to ignore it and depend only on the observation ($u_t = \pi(x_t)$). This isolates whether stability gains require coupling internal adaptation to control. This ablation is a high-value extension and is planned for the next experimental revision.

Ablation C — Scramble the update signal

Replace the update direction with random noise while keeping update size similar.

Expected Level-2 pattern: robustness degrades, often sharply.

If these ablations do not meaningfully reduce stability, the study does not support Level-2.

10.13 Evaluation metrics

Because Level-2 is about stability across time, metrics must be trajectory-level, not single-step.

Recommended primary metrics:

1. **Violation rate**
   Fraction of trials where:

$$\text{Violation rate} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[x_t \notin [x_{\min}, x_{\max}]].$$

2. **Recovery time (post-switch)**
   Recovery time measures how quickly the system returns to stable operation after it first becomes unstable following the perturbation. This avoids the misleading case where the state is briefly within the viability band immediately after the switch but violates the band shortly afterward. In the applied evaluation, recovery is defined as returning to the viability band and then remaining in-band for N consecutive steps (with N=5). If no post-switch violation occurs, recovery time is defined as 0

$$t_v = \min\{t \geq t_p \colon x_t \notin [x_{\min}, x_{\max}]\}, \tau = \min\{\Delta \geq 1 \colon x_{t_v+\Delta:t_v+\Delta+N-1} \in [x_{\min}, x_{\max}]\}.$$

   with *τ=0* if *tv* does not exist.

3. **Integrated deviation from the viability band**
   Integrated deviation captures how much and for how long the system departs from the viability band over an episode. It sums the distance of the state from the nearest boundary whenever the state is out-of-band, and contributes zero while the state remains within the band. This metric complements violation rate by distinguishing brief, small excursions from large or persistent instability

$$\text{IntDev} = \sum_{t=1}^{T} \text{dist}(x_t, [x_{\min}, x_{\max}]).$$

   where

$$\text{IntDev} = \sum_{t=1}^{T} \text{dist}(x_t, [x_{\min}, x_{\max}]).$$

Supporting metrics:

- worst-case deviation after perturbation,

- stability margin across perturbation severities,
- correlation between update magnitude and recovery success.

10.14 Expected result signatures (what counts as Level-2 evidence)

A Level-2-positive result should show:

- **Robust stability:** candidate maintains bounds better than reactive baseline across perturbations.

- **Structured adaptation:** internal estimate/state shifts after perturbation and aligns with recovery.

- **Causal necessity:** freezing/scrambling updates collapses performance.

- **Non-reducibility:** reactive mapping cannot match performance, across the evaluated perturbation family.

A Level-2-negative result should show one or more of:

- adaptive mechanism does not improve stability beyond reactive control,
- internal updates do not track perturbations or do not matter when removed,
- performance is brittle across perturbation magnitudes/timings in the evaluated family.

Both outcomes are informative: they clarify which architectures and update mechanisms do or do not produce adaptive regulation.

10.15 Limits of interpretation (what this does *not* establish)

Even a strong Level-2 result does *not* imply:

- intelligence in general domains,
- agency in a philosophical sense,
- persistent goals (Level-3),
- selfhood, consciousness, or subjective experience.

The only claim supported is:

The system preserves stability under perturbation via internal adjustment that is causally necessary and not reducible to fixed stimulus-response control.

---

11. Results — Level-2 Adaptive Self-Regulation

This section reports the empirical outcome of the Level-2 evaluation: whether the system *preserves functional stability under a hidden perturbation* by internally mediated state/parameter adjustment, rather than by purely reactive stimulus control. (Level-2 definition in the programmatic framework: the system maintains stability under perturbation via internally mediated adjustment.

11.1 What was tested (and why this is Level-2, not Level-1)

Level-1 asks a narrower question: *does internal state causally influence behaviour at all?* This is tested by holding external input fixed, changing internal state, and checking whether behaviour diverges in a way stimulus-only baselines cannot reproduce.

Level-2 adds a stricter requirement: internal variables must not only *matter*, they must change themselves to preserve stability when the world's dynamics shift in an unlabelled way. In other words:

- Perturbation occurs (a change in the environment dynamics unknown to the agent).
- A purely reactive controller may fail because it assumes the pre-change dynamics.
- A Level-2 controller should detect mismatch (prediction error) and internally adapt (update an internal estimate/state) so the behaviour returns to stability.

11.2 Multi-run stability outcomes

Across many seeds and episodes (multi-run evaluation), the Adaptive (Level-2) agent shows substantially better post-perturbation stability than all comparison agents:

- **Post-switch violation rate** (fraction of steps outside the stability band after the perturbation): Adaptive (Level-2) is very low (≈0.046), while Frozen Update, Random Update, and Reactive (tuned) are much higher (≈0.33–0.45).
- **Recovery time after first violation** (time steps required to return to stable in-band behaviour):
  Adaptive (Level-2) is near-immediate (very small), while the other agents take substantially longer on average.
- **Integrated deviation from band** (total magnitude of out-of-band deviation):
  Adaptive (Level-2) is near zero, while the other agents accumulate large deviations (order ~10^2).

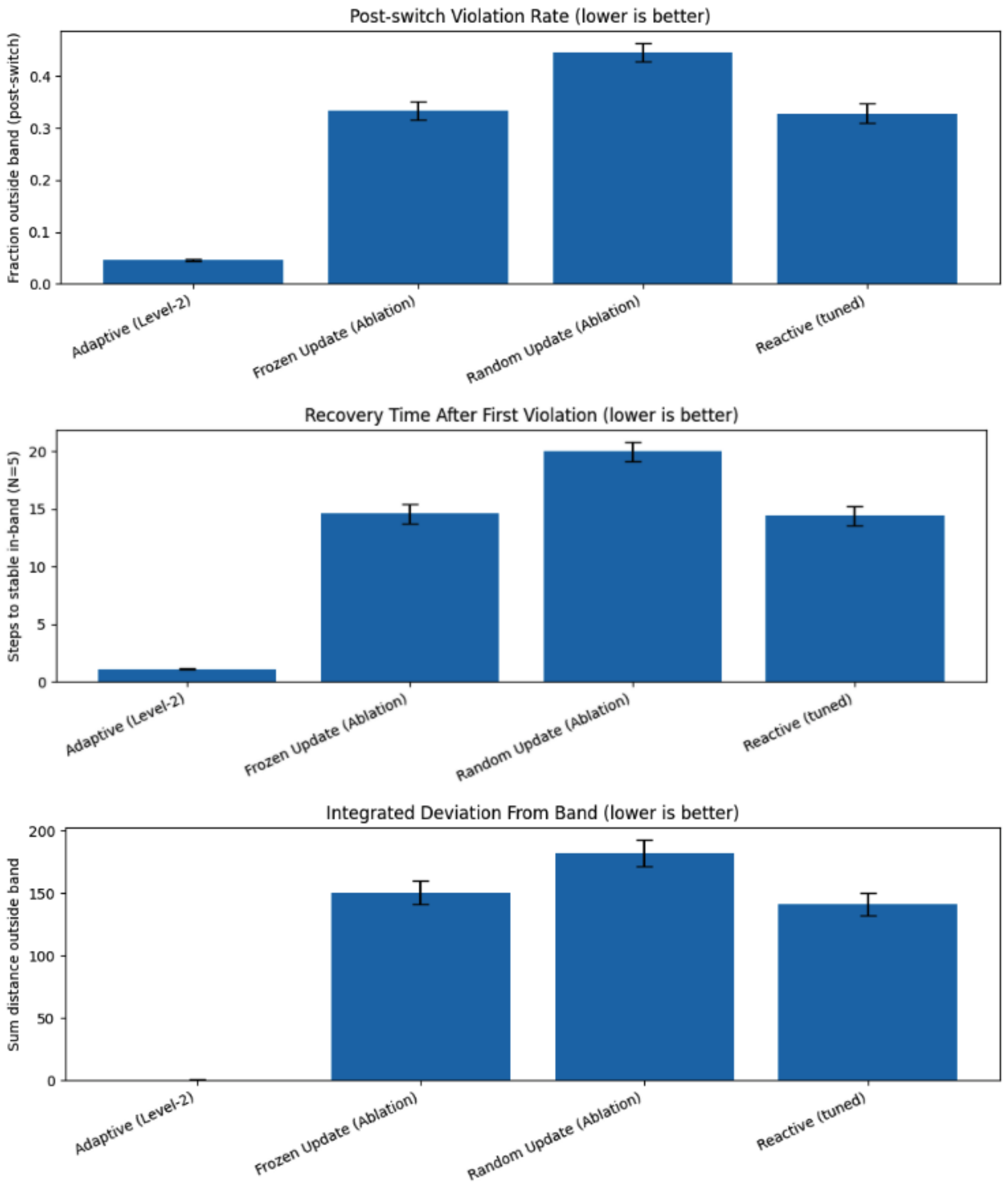


Figure 3. Multi-run stability metrics under hidden perturbation.

Bars show mean performance across repeated episodes and random seeds for the Adaptive (Level-2) controller, a tuned reactive baseline, and two ablations (Frozen Update and Random Update). Top: post-switch violation rate (fraction of post-perturbation steps outside the viability band). Middle: recovery time after the first post-switch violation, defined as steps until the system returns to the viability band and remains in-band for N=5 consecutive steps. Bottom: integrated deviation from the viability band (sum of distances outside the band across the episode). Error bars indicate 95% bootstrap confidence intervals of the mean. (Evaluation: 25 seeds × 90 episodes per seed, unless otherwise stated.)

The Frozen Update condition uses the same control structure as Adaptive but disables internal updating; it therefore provides the cleanest test of causal necessity of adaptation.

11.3 Qualitative adaptation signature: internal update occurs when dynamics change

Level-2 is not just “better performance.” It requires evidence that the improvement comes from an internal adjustment mechanism. The notebook includes an *adaptation signature* plot showing:

- the environment variable trajectory *x(t)* relative to the stability band,
- the (hidden) perturbation time (vertical marker),
- and the agent’s internal estimate (e.g., *θ^(t)*).

In the Adaptive (Level-2) agent, *θ^(t)* changes in a structured manner following the switch (consistent with prediction-error-driven updating), consistent with an internally mediated correction process. In the Frozen Update ablation, the internal estimate remains fixed by design, and stability degrades. In the Random Update ablation, the internal estimate changes but in an unstructured way, and stability also degrades—showing that “any internal change” is not sufficient; it must be *error-coupled and functionally stabilizing*.

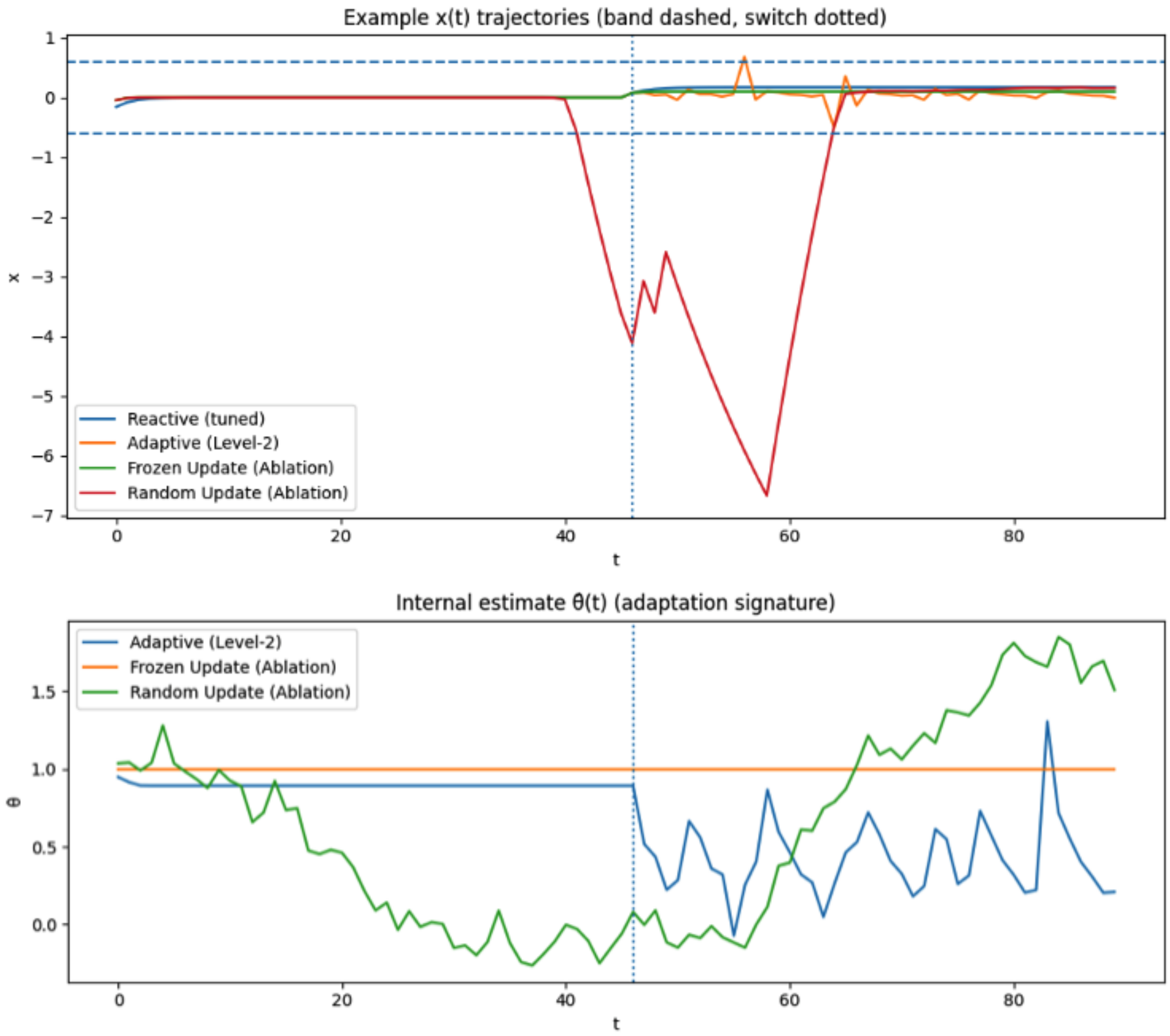


Figure 4. Example trajectories and internal adaptation signature under a hidden dynamics switch.

Top: Example x(t) trajectories for the tuned reactive baseline, the Adaptive (Level-2) controller, and two ablations. The viability band is shown by dashed horizontal lines; the perturbation time (hidden dynamics change) is marked by the dotted vertical line. The Adaptive (Level-2) controller remains within (or rapidly returns to) the viability band after the switch, whereas the ablations and reactive baseline exhibit larger instability in this example.

Bottom: Internal estimate θ^(t) for controllers that maintain an internal parameter state. Adaptive (Level-2) shows a structured change in θ^(t) following the switch (consistent with prediction-error-driven updating). Frozen Update remains constant by design, and Random Update exhibits drift not coupled to stabilization, illustrating that internal change alone is insufficient—what matters is error-linked adjustment that supports post-perturbation stability.

11.4 Causal tests and ablations (why this supports Level-2 attribution)

A common failure mode in early self-regulation demonstrations is accidental “cheating,” where the system’s stability comes from a strong reactive controller that happens to generalize, rather than from true internal adaptation.

To make the claim falsifiable, the evaluation includes ablations that directly target the Level-2 requirement:

1. **Frozen Update (Ablation):** adaptation is disabled (internal update is frozen).
   If performance remains strong, then “adaptive self-regulation” is not needed.
   Observed outcome: stability worsens substantially relative to Adaptive (Level-2), indicating internal updating is *causally relevant*.

2. **Random Update (Ablation):** internal parameters change, but not in a prediction-error-coupled way. If this performs as well as Adaptive (Level-2), then the “update mechanism” is not meaningful. Observed outcome: performance worsens, suggesting the specific error-coupled update is doing the work.

3. **Reactive (tuned) baseline:** no internal model update; action is chosen from observation only. Observed outcome: it fails under the dynamics shift more often than Adaptive (Level-2), consistent with the Level-2 claim that purely reactive control is insufficient under this perturbation regime.

These interventions are the Level-2 analogue of Level-1’s stimulus-only baseline requirement: the point is to show that the defining mechanism (internal adaptation) is necessary for the observed stability, rather than merely present.

The Frozen Update condition uses the same control structure as Adaptive but disables internal updating; it therefore provides the cleanest test of causal necessity of adaptation.

Because the reactive baseline uses a different control law, its comparison to Adaptive is supportive but not a pure isolation of the update mechanism.

The present evaluation includes Ablation A (freeze updating) and Ablation C (randomized updating), but not the policy–state decoupling ablation (B); adding (B) is an immediate next step to further isolate the role of policy coupling.

11.5 Level-2 attribution (what the results justify—and what they do not)

Supported attribution:
The combination of (i) improved stability under a hidden perturbation, (ii) an internal adaptation signature aligned with prediction error, and (iii) ablations showing performance degrades substantially when structured adaptation is removed, together support the claim that the implemented system satisfies the operational Level-2 criterion: adaptive self-regulation via internally mediated adjustment. (Level-2 definition: stability under perturbation via internal adjustment).

Not supported (explicitly out of scope):
These results do *not* imply intelligence, agency, selfhood, or consciousness—just as Level-1 results did not. The program's own epistemic constraint is that each level only supports its local definition, nothing stronger.

11.6 Limitations and remaining falsification pressure

Even with strong separation between Adaptive and baselines, several limits remain important:

- **Environment simplicity:** the task is deliberately minimal to isolate the causal mechanism (as in Level-1's methodological choice to isolate causal structure in a controlled substrate).

- **Perturbation family:** the current perturbation is a specific class of dynamics shift. Stronger Level-2 evidence would test multiple perturbation magnitudes, times, and forms.

- **Alternative explanations:** any remaining possibility that the adaptive controller is effectively "tuned to the perturbation distribution" should be pressured via broadly sampled perturbations from the evaluation family and more aggressive stress tests.

12. Conclusion

This paper defined and evaluated Level-2: Adaptive Self-Regulation within the staged Conscious Machines research program. Level-2 requires that a system preserve functional stability under perturbation through internally mediated state or parameter adjustment, rather than relying on a fixed stimulus→response mapping.

The reported results support a Level-2-positive interpretation under the operational definition used here. Across multi-run trials with a hidden dynamics switch, the Adaptive (Level-2) controller consistently showed:

- substantially lower post-switch violation rates,
- faster recovery after first violation, and

- markedly reduced integrated deviation from the viability band,

relative to a tuned reactive baseline and two ablations. The ablation pattern is consistent with causal attribution: when internal updating was disabled (Frozen Update) or decoupled from prediction error (Random Update), stability performance degraded toward baseline levels. The qualitative traces further align with a Level-2 mechanism: the adaptive controller exhibits a structured post-switch change in its internal estimate *θ^(t)* that corresponds to restored stability.

Importantly, this is a bounded conclusion. A Level-2 result does not imply general intelligence, agency, selfhood, or subjective experience. It supports only the narrow claim that the tested system demonstrates adaptive regulation—a specific organizational property that is plausibly prerequisite to later levels concerned with persistent targets and temporal continuity.

13. Next steps

Two immediate steps would strengthen the Level-2 result without adding conceptual complexity: (1) broaden the perturbation family (e.g., vary switch magnitude/timing more aggressively, include leak changes or multi-switch sequences), and (2) add held-out tests where evaluation perturbations fall outside the tuning distribution (unseen magnitudes, timings, or combinations). These checks reduce the chance that the observed adaptation is narrowly fitted to one perturbation regime and increase confidence that the mechanism reflects genuine adaptive self-regulation rather than fragile tuning.

With Level-2 established as a stability substrate, the next stage is *Level-3: Goal-Directed Policy Persistence*, which shifts the success criterion from “stay within a viability band” to “continue pursuing an internally represented target across delay and interference.” A clean Level-3 design can reuse the same environment dynamics, but must introduce explicit target representations and structured distractors/interruptions, plus ablations that selectively disrupt target maintenance to test causal necessity.