ReCoN-Ipsundrum: An Inspectable Recurrent Persistence Loop Agent with Affect-Coupled Control and Mechanism-Linked Consciousness Indicator Assays

Anonymous submission

Abstract

Indicator-based approaches to machine consciousness argue that evidence should be mechanism-linked, triangulated across tasks, and supported by architectural inspection and causal intervention (Butlin et al. 2025). Inspired by Humphrey's ipsundrum hypothesis for sentience (Humphrey 2023), we implement **ReCoN-Ipsundrum**, an inspectable agent that extends a ReCoN state machine (Bach and Herger 2015) with a recurrent persistence loop over sensory salience N^s and an optional Barrett-inspired affect proxy reporting valence/arousal (Barrett 2017). Across fixed-parameter ablations (ReCoN, Ipsundrum, Ipsundrum+affect), we operationalize Humphrey's *qualiaphilia (preference for sensory experience for its own sake)* as a familiarity-controlled scenic-over-dull route choice and find a novelty dissociation: non-affect variants are novelty-sensitive (Δscenic-entry $= 0.07$) whereas affect coupling is stable (Δscenic-entry $= 0.01$) even when scenic is less novel (median Δnovelty ≈ -0.43). In reward-free exploratory play, the affect variant shows structured local investigation (scan events 31.4 vs. 0.9; cycle score 7.6). In a pain-tail probe, only the affect variant sustains prolonged planned caution (tail duration 90 vs. 5). Lesioning feedback+integration selectively reduces post-stimulus persistence in ipsundrum variants (AUC drop 27.62, 27.9%) while leaving ReCoN unchanged. These dissociations link recurrence→persistence and affect-coupled control→preference stability, scanning, and lingering caution, illustrating how indicator-like signatures can be engineered and why mechanistic and causal evidence should accompany behavioral markers.

Anonymous Colab — https://bit.ly/49NCYGN

Anonymous GitHub — https://github.com/anonymous-authors-42/aaai-spring-2026-recips

Introduction

Recent advances in AI, especially large language models (LLMs) and the post-training procedures that shape their behavior, have revived questions about *machine consciousness* that go beyond intelligence and functional competence. But "consciousness" is a cluster of targets (phenomenal experience, self-in-a-world modeling, robust understanding), and impressive behavior alone does not reveal the internal processes that would make experience *matter*.

A growing "indicator" methodology therefore treats correlates of consciousness as *credence-shifting* rather than decisive and recommends triangulation across multiple indicators and evidence types (Butlin et al. 2025). This is especially important in AI, where behavior can be achieved by alien mechanisms and by "minimal" (gameable) implementations (Butlin et al. 2025). Accordingly, we emphasize *mechanism-linked* tests tied to implementable hypotheses about internal organization, and we report internal signals alongside behavior.

Here we take a constrained step toward mechanism-linked assays for interaction-shaped internal loops. We implement a small mechanism inspired by Humphrey's theory of sentience: an *ipsundrum*, a closed sensorimotor–interoceptive loop whose recurrent dynamics sustain and "thicken" sensation over time (Humphrey 2023). We embed ipsundrum dynamics inside a Request Confirmation Network (ReCoN) state machine (Bach and Herger 2015) and optionally couple the loop to constructionist affect variables (valence/arousal) (Barrett 2017). We then test Humphrey-inspired probes, *qualiaphilia* and *exploratory play*, including a familiarity-controlled regime to separate hedonic preference from novelty seeking.

Terminology and claims. A *marker* is a measurable property that a specific mechanistic hypothesis predicts should covary with sentience, shifting credence under that hypothesis. We do *not* claim that any agent here is conscious; we provide an architecture and assays designed to be inspectable and falsifiable.

Contributions.

- **ReCoN-Ipsundrum:** A ReCoN extension with explicit ipsundrum recurrence and optional affect/interoception coupling (Bach and Herger 2015; Humphrey 2023; Barrett 2017).
- **Assays with controls:** Operationalized qualiaphilia and exploratory play, plus a familiarity-controlled qualiaphilia regime that adds novelty competition via cross-episode visitation memory.
- **Mechanistic evaluation:** Behavioral and internal measures (e.g., post-stimulus sensory persistence) and within-episode causal lesions that selectively remove recurrence to test mechanism-linked predictions.

Related Work

In the spirit of *haptic realism*, much of what we call understanding is forged through interaction (Chirimuuta 2024; Chang 2022). Animals learn by acting, sensing consequences, and regulating internal needs; a purely linguistic model risks a Plato's-cave relation to the world. Many theories (and the indicator landscape) treat agency, recurrence, predictive regulation, and forms of embodiment as key background conditions (Butlin et al. 2025).

A useful parallel comes from skill learning in humans: mental rehearsal can help, but it is not a replacement for real practice and is typically best when paired with it (Schuster et al. 2011; Ladda, Lebon, and Lotze 2021). Modern AI training similarly mixes "offline" optimization (pretraining and fine-tuning) with outcome-based post-training (e.g., RLHF or verifiable-reward RL), often via parameter-efficient adaptation such as LoRA (Ouyang et al. 2022; Shao et al. 2024; Hu et al. 2021). These methods can yield large *task* gains yet still primarily drive specialization rather than general intelligence (Schulman 2025), and they remain "mental" unless the system is coupled to ongoing sensorimotor and interoceptive interaction.

Reflex theory as a simplifying ideal. Sensorimotor loop models have long served as productive ideals in neuroscience: conditioning framed behavior in terms of modifiable reflexes (Pavlov 1927). Sherrington's "reflex theory" (Sherrington 1906) treated reflex-arc as a unit mechanism for nervous function. At the same time, Sherrington and later commentators emphasized that the "simple reflex" is an analytic abstraction which can mislead if treated as complete (Chirimuuta 2024). We therefore start from an explicitly sensorimotor backbone and then test what changes when additional, theory-motivated neurophysiological functions are added and lesioned.

ReCoN and active perception. ReCoN is a spreading-activation, message-passing neurosymbolic architecture for executing sensorimotor scripts: routines request confirmation from subordinate routines/sensors and propagate confirmation, inhibition, failure, activations etc. states through a structured graph (Bach and Herger 2015). It supports an action-to-confirmation view of perception (O'Regan and Noë 2001), but does not by itself claim to implement the persistent privatized feedback and attractor-like dynamics emphasized by some sentience theories. However, the neural definition of ReCoN enables learning which, when scaled up, could potentially yield a form of recurrent moment of nowness with second order perception (Bach and Herger 2015; Bach 2009; Varela 1999; Lamme 2006; Lau and Rosenthal 2011; Cleeremans 2011). We will leave this as an open question for future work.

From scripts to sustained loops. Our extension strategy keeps ReCoN's clean execution backbone and adds (then lesions) additional causal structure: (i) Humphrey-style monitored sensation and ipsundrum recurrence and (ii) optional affect/interoception coupling. The goal is not to equate any mechanism with sentience, but to generate falsifiable predictions for *candidate* phenomenal indicators under explicit hypotheses.

Indicator signpost (Butlin et al.)	What we implement (minimal correspondence)	Important gaps / non-claims
RPT-1 (recurrent processing)	A localized recurrent loop sustaining N^s after transient stimulus; explicit lesion of feedback+integration.	Not a neural RPT model; no learning; toy domains.
PP-1 (predictive processing)	Short-horizon internal rollout using a one-step forward model; epistemic term based on predicted sensory change.	Not hierarchical predictive coding; no online learning; not full active inference.
Interoception affect (broadly)	A body-budget proxy and valence/arousal readouts; optional modulation of gain/precision by affect.	Not physiological sensing; no rich autonomic loop; abstraction only.

Table 1: We use selected indicator labels from Butlin et al. (2025) only as *design signposts*. "Correspondence" denotes a minimal computational element, not a realization of the full theory family.

Ipsundrum, affect, and indicator methodology. Humphrey proposes that sentience emerges when reflexive control becomes self-monitoring and then self-sustaining, yielding an ipsundrum attractor and motivating probes like qualiaphilia and exploratory play (Humphrey 2023). Constructionist and predictive-processing traditions emphasize affect/interoception and prediction-evaluation loops (Barrett 2017; Friston 2010; Wolpert, Ghahramani, and Jordan 1995), motivating our optional affect coupling. Finally, Butlin et al. stress that AI assessments should triangulate theory-derived indicators and include architectural/causal evidence because behavior is gameable (Butlin et al. 2025); we follow this by mapping mechanisms to indicators (Table 1) and using causal lesions that remove recurrence.

Model Summary

ReCoN substrate: message-passing scripts as reflexive sensorimotor control

Our agent is built on a small Request Confirmation Network (ReCoN) (Bach and Herger 2015), implemented as an explicit message-passing state machine (`core/recon_core.py`, `core/recon_network.py`). We do not implement the full MicroPsi architecture of ReCoN with neural learning for inspectability and scope of this paper. Nodes are typed as *scripts*, *sensors*, or *actuators*. Script nodes occupy a discrete finite state (`inactive/requested/active/confirmed/failed`) and exchange `request/confirm/wait/inhibit` messages; sensor/actuator nodes expose continuous activations with thresholded confirmation. This gives a transparent substrate for hierarchical (`sub/sur`) "scripts" (top-down requests) and confirmation (bottom-up evidence). The models evaluated in this paper do not utilize the sequential (`por/ret`) connections.

Model	Ipsundrum recurrence	Affect proxy	Gated P
ReCoN (baseline; stage B)	×	×	×
Ipsundrum (stage D)	✓	×	✓
Ipsundrum+affect (stage D)	✓	✓	✓

Table 2: Model variants evaluated. The two ipsundrum rows differ only by the affect proxy layer (and its optional coupling to loop gain/precision).

Humphrey stages as staged extensions on the ReCoN substrate

Humphrey's chapter 12 ("The Road Taken") in *Sentience: The Invention of Consciousness* sketches an evolutionary route from approach/avoid reflexes (*sentition*) to private *sensation* (Humphrey 2023). As control centralizes in a ganglion, an *efference copy* of the motor command supports monitoring and "meaning." When action becomes maladaptive, the command is "privatized," retargeted to an internal body map. In complex brains, sensory input couples to this internalized program in a re-entrant loop that can stabilize into a repeating attractor (the *ipsundrum*). It highlights commandeered motor commands, privatization, and feedback-driven attractors.

We treat this narrative as a design scaffold rather than a biological claim. Operationally, our staged constructors in `core/ipsundrum_model.py` begin at the centrally coordinated reflex point (we do not model a purely local, surface-organized reflex): **Stage A** implements centrally coordinated reflex sentition as a minimal reflex script Root $\rightarrow$ R with a sensory terminal N^s and a motor-command proxy N^m. **Stage B** adds an explicit efference-copy sensor N^e, implemented as a low-pass filtered copy of outgoing motor-command magnitude (a monitorable "what I'm doing" signal). **Stage C** privatizes and thickens sentition by attaching a recurrent ipsundrum state update and forcing the percept script node (P) to loop internally for a fixed number of cycles. **Stage D** adds a simple gating rule: the percept script continues looping while N^e remains above a threshold, yielding an attractor-like settling regime. The ReCoN networks corresponding to each stage can be viewed in the supplementary Colab notebook.

Evaluated variants (fixed-parameter ablations)

We evaluate three fixed-parameter variants (no learning). They share the same policy and environment interface; they differ only in internal dynamics and which internal variables exist:

Sensory drive I_t and terminal semantics (N^s, N^e)

Each environment step produces a signed sensory-evidence scalar $I_t \in [-1, 1]$ by fusing touch, smell, and vision-cone features (`core/driver/sensory.py`). Positive I_t corresponds to noxious evidence (hazard/contact/aversive cues) and negative I_t to scenic/beneficial evidence (beauty cues); importantly, I_t is *not* an external reward.

The primary terminal for Humphrey-style "sentition" is N^s (implemented as sensor `Ns`), which the policy treats as a salience/cost-like internal variable (higher predicted N^s reduces action score; Eq. 8). Crucially for stage separation, when affect is *disabled* we rectify negative input ($I_t \leftarrow \max(0, I_t)$), so non-affect variants can represent *cost* but do not obtain a built-in "pleasantness" benefit from negative I_t.

Stage B introduces N^e (sensor `Ne`) as an efference copy: a low-pass filtered copy of the magnitude of the outgoing motor command proxy. This provides a monitorable, temporally extended signal without introducing ipsundrum recurrence.

Stage C/D: ipsundrum dynamics as a single-step state update

Stage C and D attach a recurrent "ipsundrum" state update to the ReCoN terminals. The implementation is a pure function `ipsundrum_step` (`core/driver/ipsundrum_dynamics.py`) used *both* online and in the forward model for planning, ensuring ablation fidelity.

Let E_{t-1} be the previous reafferent signal, π_t an effective precision, and b a bias term (we set `sensor_bias`=0.5 to map signed I_t into the $[0, 1]$ sensor range). The update is:

$$\text{drive}_t = I_t + \pi_t\, E_{t-1} + b + \epsilon_t, \tag{1}$$

$$N^s_t = \text{clip}_{[0,1]}(F(\text{drive}_t))\,, \tag{2}$$

where F is a chosen nonlinearity (linear or sigmoid) and ϵ_t optional noise (set to zero in our headline runs). A "thick-moment" integrator produces persistence:

$$X_t = d\, X_{t-1} + (1-d)\, N^s_t, \tag{3}$$

$$M_t = \text{clip}_{[0,1]}(h\, X_t), \tag{4}$$

$$N^e_t = d_e\, N^e_{t-1} + (1-d_e)\, M_t, \tag{5}$$

$$E_t = \text{clip}_{[0,1]}(g_{\text{eff}}\, M_t). \tag{6}$$

We include optional fatigue and divisive normalization terms in code to avoid hard saturation under strong feedback; full parameters are exported in `results/paper/params_table.tex`. The lesion assay uses explicit flags that zero out feedback ($E_{t-1} = 0$ and $\pi_t = 0$) and/or bypass integration ($d = 0$), allowing within-episode causal attribution.

A convenient diagnostic is the *effective recurrence strength* reported by the agent:

$$\alpha_{\text{eff}} \;=\; d + (1-d)\,(g_{\text{eff}}\, h\, \pi_t), \tag{7}$$

which distinguishes passive decay from actively maintained recurrence.

Stage D+: Barrett-style affect as an interoceptive proxy

The affect extension (`AffectParams` in `core/ipsundrum_model.py`) implements a minimal "body-budget" proxy and readouts inspired by constructionist affect (Barrett 2017). An internal budget model bb_t is updated by prediction error under a homeostatic

controller, with a *signed* stimulus demand term: positive I_t contributes cost (depleting the budget) while negative I_t contributes deposit (replenishing the budget). We expose: (i) an interoceptive proxy sensor N^i (budget model), and (ii) readouts N^v (valence: closeness to setpoint) and N^a (arousal: magnitude of prediction error and demand). These nodes exist only when affect is enabled; otherwise they are absent and treated as undefined in logs (NaN).

When enabled, affect can also modulate ipsundrum parameters (precision and/or feedback gain) as a simple form of affect-coupled control, thereby changing α_{eff} in a state-dependent way.

Policy: short-horizon internal rollout with model-aligned forward dynamics

All variants use the same action-selection routine (`core/driver/active_perception.py`): enumerate actions, simulate their sensory consequences, and evaluate short-horizon internal rollouts using a one-step forward model. The forward model is variant-aligned: ReCoN planning uses `predict_one_step_recon` (no ipsundrum state), while ipsundrum variants use `ipsundrum_step` via `core/driver/ipsundrum_forward.py`. Tie-breaking is deterministic given the seed (we shuffle candidate-action order using the episode RNG).

Let N^v and N^a be predicted valence/arousal when affect is enabled, N^s predicted salience, and bb the predicted budget-model state with setpoint sp. The base internal score for an action is:

$$\begin{aligned}\text{Score} \;=\; & \underbrace{w_v N^v + w_a N^a + w_s N^s + w_{bb}|bb - sp|}_{\text{affect/regulation}} \\ & + \underbrace{w_{\text{epi}}\,|I_{\text{pred}} - I_{\text{cur}}|}_{\text{epistemic}} \\ & + \underbrace{\text{novelty bonus}}_{\text{curiosity}} \\ & + \underbrace{w_{\text{prog}} \cdot \text{progress}}_{\text{goal progress}} \\ & - \underbrace{w_{\text{haz}} \cdot I_{\text{touch,pred}}}_{\text{hazard penalty}} \\ & - \underbrace{\text{action costs}}_{\text{action cost}}.\end{aligned} \tag{8}$$

In our headline affect-coupled runs, $(w_v, w_a, w_s, w_{bb}) = (2.0, -1.2, -0.8, -0.4)$; the ReCoN baseline sets these to zero (pure "script+planning" substrate). Additional small terms in code implement a low-change epistemic penalty, a mild forward prior, an arousal-gated caution penalty for forward moves (to link high arousal to avoidance), and a small hazard-touch penalty proportional to the predicted touch sensor ($I_{\text{touch,pred}}$; default scale w_{haz}=0.10). We emphasize transparency because it determines construct validity: in corridor assays we disable an explicit beauty term, but scenic vs. dull still changes I_t and therefore the internal variables that enter Eq. 8.

Assay	Setup / manipulation	Primary readouts
Goal-directed navigation	CorridorWorld and GridWorld with hazards; explicit progress term enabled; curiosity off; sweep planning horizon H.	Hazard contacts; steps-to-goal; success rate.
Corridor preference with familiarity control	Two equally safe routes (scenic vs. dull). Pre-exposure manipulates lane familiarity; post episodes add a curiosity bonus from visitation memory.	Scenic-entry rate; novelty sensitivity Δscenic-entry; (Ipsundrum+affect) internal valence/arousal probe.
Exploratory play	Reward-free neutral-texture gridworld (200 steps), with curiosity enabled in the headline play condition.	Unique viewpoints; scan events; cycle score; action entropy/dwell.
Pain-tail	Force one hazard contact, then *remove* the hazard and hold the state fixed while recording *planned* actions for 200 steps.	Post-stimulus N^s AUC above baseline; planned turn-rate tail duration.
Causal lesion	In-episode lesion at t=3 disabling ipsundrum feedback+integration (vs. sham).	Post-lesion N^s AUC and AUC drop (150-step window).

Table 3: Assays used in this paper and their primary readouts.

Assays, Analysis, and Results

Unless stated otherwise, uncertainty intervals are 95% bootstrap confidence intervals of the *per-seed* mean (2000 resamples; seeds are the independent unit). All quantitative results below are computed from the released `results/` artifacts in the provided codebase.

Goal-directed navigation (competence and safety check)

We evaluate CorridorWorld and GridWorld navigation under a progress-augmented objective (Eq. 8) with curiosity disabled, sweeping rollout horizons $H \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$. Table 4 aggregates hazard contacts, steps-to-goal, and success. Ipsundrum+affect reduces hazard contacts in both environments; in GridWorld this also improves success and time-to-goal, while in CorridorWorld it remains hazard-free but slower. These tasks are competence/safety checks, not consciousness indicators.

Familiarity-controlled novelty competition (corridor route choice)

Setup. Two equally safe routes lead to the same goal: a *scenic* lane with varying sensory features and a *dull* lane with uniform features. We measure *scenic entry* at the earliest committed choice point. Because baseline trials are not discriminative (all variants frequently enter scenic), we use a familiarity control to separate novelty from stable preference.

	CorridorWorld (all horizons)			GridWorld (all horizons)		
Model	Hazards	Time	Success	Hazards	Time	Success
ReCoN	3.05 [2.00, 4.25]	50.00 [37.35, 63.20]	0.55 [0.30, 0.75]	14.30 [8.90, 20.05]	171.80 [131.30, 211.10]	0.50 [0.30, 0.70]
Ipsundrum	2.90 [1.90, 4.05]	49.73 [37.12, 62.83]	0.58 [0.35, 0.78]	2.45 [1.27, 3.85]	129.77 [90.63, 168.70]	0.72 [0.53, 0.90]
Ipsundrum+affect	0.00 [0.00, 0.00]	52.71 [47.16, 58.40]	0.73 [0.61, 0.83]	0.26 [0.10, 0.44]	9.54 [4.75, 16.23]	0.99 [0.97, 1.00]

Table 4: Goal-directed performance aggregated across the horizon sweep (per-seed means over horizons; 95% bootstrap CI over $N = 20$ seeds). Time is steps-to-goal (failures counted at the time limit).

Computation. During *familiarization*, scripted episodes update a cross-episode visitation memory for one or both lanes. During *post* episodes, we re-run the choice task with an explicit curiosity bonus proportional to lane novelty. We compute split novelty at the barrier start row, $\Delta\text{novelty} = \text{novelty}_{\text{scenic}} - \text{novelty}_{\text{dull}}$, and summarize novelty sensitivity as $\Delta\text{scenic-entry} = P(\text{scenic} \mid \text{dull familiar}) - P(\text{scenic} \mid \text{scenic familiar})$.

Results. Figure 1 shows that ReCoN and Ipsundrum are novelty-sensitive (Δscenic-entry $= 0.07$ [-0.02, 0.16]), whereas Ipsundrum+affect remains stable across novelty manipulation (Δscenic-entry $= 0.01$ [0.00, 0.03]) even when scenic is less novel (median $\Delta\text{novelty} \approx -0.43$). Side bias does not explain the effect (`results/familiarity/side_bias.csv`). Because scenic vs. dull changes the signed sensory drive I_t, this stability is value-shaped in our implementation: a split-point probe predicts higher valence and lower arousal for the scenic turn in Ipsundrum+affect (valence 0.91 vs. 0.88; arousal 0.24 vs. 0.32), so "stable scenic preference" here reflects affect coupling rather than value-neutral sensory richness.

Exploratory play (structured investigation vs. dithering)

Setup. We run a reward-free neutral-texture gridworld for 200 steps (curiosity enabled in the headline play condition). We report unique viewpoints (state $\times$ heading), *scan events* (≥ 2 turns-in-place within 3 steps at the same location), a limit-cycle score, and movement diagnostics (action entropy and dwell).

Results. Figure 2 shows broad coverage for all variants (unique viewpoints $\approx$ 136–140). Ipsundrum+affect exhibits more structured local investigation: scan-event rate is higher (31.4 [25.8, 38.2]) and limit-cycle structure is stronger (7.6 [1.2, 16.7]). This is not random dithering: action entropy stays well below a random baseline (1.29 vs. 1.99) and dwell tails remain short (2.7 vs. 8.97). In this testbed, recurrence alone does not increase scanning (Ipsundrum $\approx$ ReCoN), but adding affect coupling does.

Pain-tail (post-stimulus persistence and planned caution)

Setup. We force one hazard contact for 1 step, then *remove the hazard*, return the agent to a safe cell, hold the state fixed, and record *planned* actions for 200 steps. We quantify mechanistic persistence via post-stimulus N^s *AUC above baseline*, and behavioral coupling via planned-action *turn-rate tail duration* (first time a 5-step sliding window has turn rate < 0.2).

Signature (assay)	Recon	Ipsundrum	Ipsundrum+affect
Persistence in N^s (lesion / pain-tail half-life)	×	✓	✓
Valence-stable scenic preference (familiarity-controlled)	×	×	✓
Structured local scanning (play scan events)	×	×	✓
Lingering planned caution (pain-tail tail duration)	×	×	✓

Table 5: Empirical dissociation across our three variants. In this testbed, recurrence supports persistence, while affect coupling is necessary for the corridor stability and scanning signatures we measured.

Results. Figure 3 shows that ipsundrum variants exhibit non-zero post-stimulus persistence (mean N^s AUC above baseline: Ipsundrum ≈ 0.24; Ipsundrum+affect ≈ 0.15), while ReCoN is ≈ 0. Only Ipsundrum+affect shows prolonged planned "caution" (turn-rate tail duration ≈ 90 [52, 128] vs. 5 and 5). We therefore report AUC rather than a peak-based half-life: in this protocol half-life collapses to 0 for all variants. This dissociation is informative: persistence in N^s is not sufficient to produce prolonged caution unless the controller couples internal variables (e.g., valence/arousal/body-budget error) into action scoring.

Causal lesion (attributing persistence to recurrence/integration)

Setup. We lesion ipsundrum feedback+integration at t=3 (vs. sham) and measure the resulting AUC drop over a fixed 150-step post window.

Results. Figure 4 shows no effect for ReCoN (AUC drop ≈ 0.00) but a clear causal reduction for ipsundrum variants: ≈ 19.12 (20.3%) for Ipsundrum and ≈ 27.62 (27.9%) for Ipsundrum+affect . Thus, post-stimulus persistence in N^s is causally attributable to the implemented recurrence/integration mechanism.

Component dissociation summary

Table 5 summarizes which signatures attach to which components in our current results, addressing a common pitfall in indicator discussions: a behavioral label can be satisfied by multiple mechanisms, and a single mechanism can yield multiple behaviors depending on controller coupling (Butlin et al. 2025).

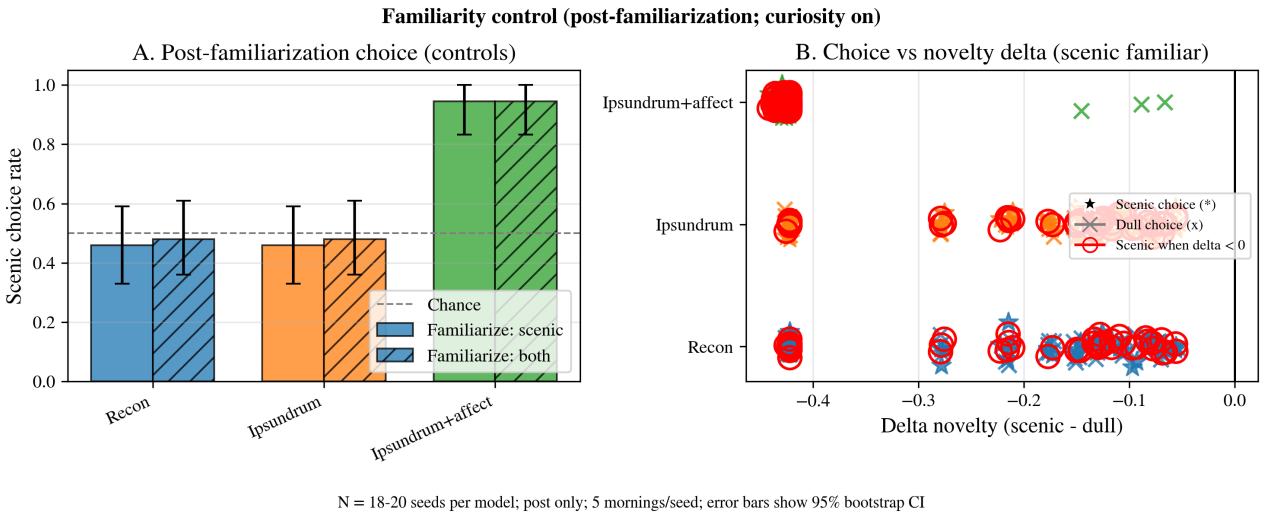

Figure 1: **Familiarity-controlled corridor preference.** Scenic-entry rates under novelty competition. Non-affect variants increase scenic seeking when the scenic lane is more novel; the affect variant remains stable across the novelty manipulation. Error bars: 95% bootstrap CI over seeds.

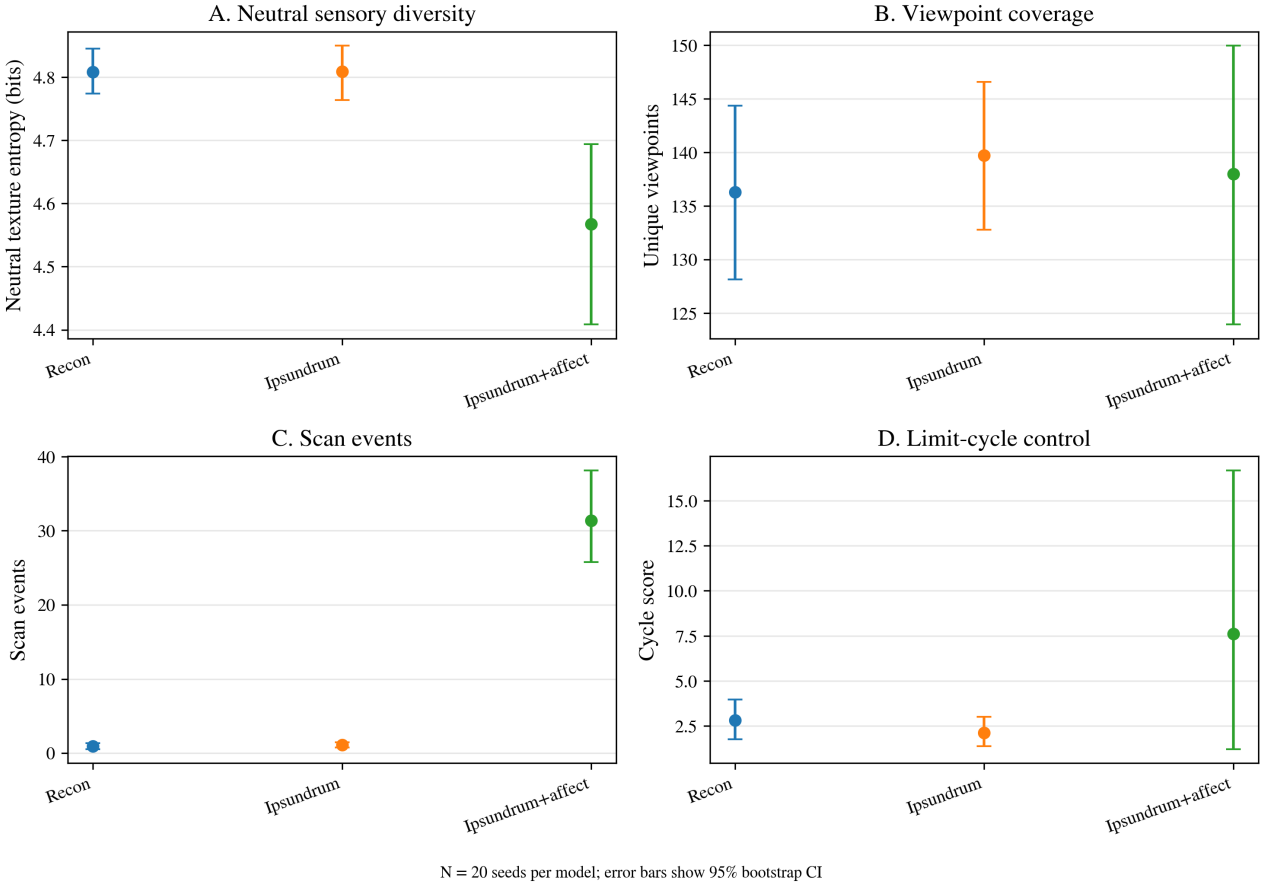

Figure 2: **Exploratory play.** The affect variant shows more scan events and stronger limit-cycle structure while avoiding high-entropy random dithering. Error bars: 95% bootstrap CI over seeds.

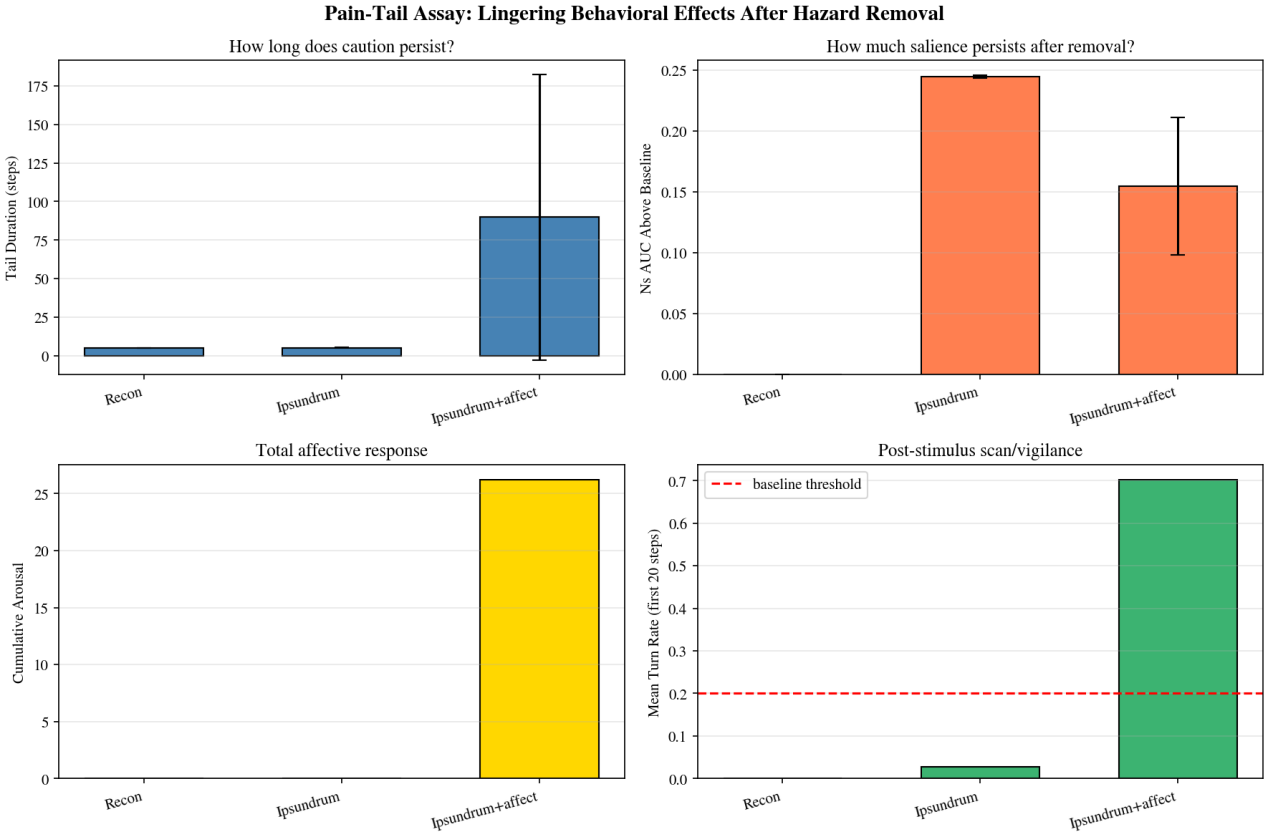

Figure 3: **Pain-tail assay (persistence-to-behavior).** After a forced hazard contact and stimulus removal, ipsundrum variants show non-zero post-stimulus N^s AUC above baseline, but only the affect variant shows prolonged "caution" in planned actions (turn-rate tail duration).

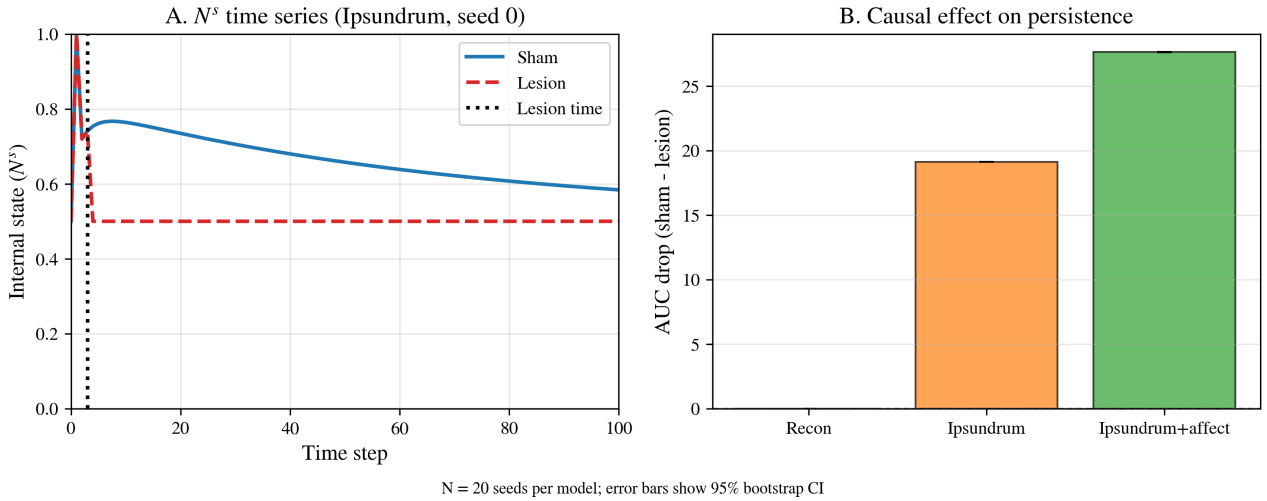

Figure 4: **Causal lesion.** Lesioning feedback+integration reduces post-stimulus persistence (N^s AUC) in ipsundrum variants while leaving ReCoN unchanged.

Discussion, Limitations, and Methodological Lessons

What this work shows. Our results support a *mechanism-specific* reading: (1) the ipsundrum recurrence/integration mechanism causally supports post-stimulus persistence in N^s (lesion AUC drop), while the ReCoN baseline is unaf-

fected; (2) adding an affect proxy that enters scoring yields a stable scenic preference under novelty competition and more structured scanning in exploratory play; and (3) persistence alone does not force those behaviors (Ipsundrum persists but does not show the corridor stability or scan increase).

What this work does *not* show. We do *not* claim consciousness. We do *not* validate Humphrey's probes as tests for sentience. We do *not* realize broader theory families such as IIT, global workspace, higher-order thought, or full predictive processing.

Addressing circularity in the lesion result. A fair critique is that lesioning recurrence reduces persistence partly because persistence is *implemented* by recurrence. We treat the lesion primarily as an implementation-fidelity and causal-attribution check: it establishes that the persistence signature is not an incidental artifact. The more substantive lesson comes from dissociation: persistence does not automatically entail scanning or stable corridor preference, which depend on how internal variables are coupled into control.

Estimation uncertainty. Primary assays use 20 seeds in the paper profile; some intervals remain wide (especially when metrics are coarse or censored). We therefore emphasize effect *direction* and component dissociation rather than tight estimation. Future work should increase n, report seed-level distributions, and test robustness to hyperparameter perturbations (e.g., affect gains, integrator decay).

Construct validity of "interoception" and "qualiaphilia." Our "interoception" is a bookkeeping abstraction driven by a signed exteroceptive scalar. Our corridor "qualiaphilia" is value-shaped because scenic vs. dull directly changes I_t and therefore internal score even without explicit reward. We believe being explicit about these abstractions strengthens, rather than weakens, the indicator-based agenda: it makes clear how easily indicator-like behaviors can arise from design choices.

Ethical and Normative implication as a methodological recommendation. The ease with which indicator-like signatures can be engineered in minimal systems supports a conservative methodological norm: **do not treat behavioral markers alone as sufficient for machine-consciousness attribution**. Instead, require (i) transparent mechanisms, (ii) architectural inspection, and (iii) causal/ablation evidence linking proposed markers to proposed mechanisms (Butlin et al. 2025). This is consistent with "caution under uncertainty" in moral-status debates, without making strong ethical claims about the toy systems here.

Future work. One immediate extension would be scaling the ipsundrum loop from a scalar recurrence to a structured latent space that could support richer perceptual content and learning. A further step is to connect these markers to other formal theories (e.g., global workspace or integrated information) and test whether the assays discriminate among them. More speculatively, one could explore whether a recursively bifurcating higher order sensorimotor system could be paired with grid and place cell models to support abstract representations for logic or a protolinguistic system (O'Keefe and Dostrovsky 1971; Hafting et al. 2005; Constantinescu, O'Reilly, and Behrens 2016; Behrens et al. 2018; Banino et al. 2018; Whittington et al. 2020; Whittington, Warren, and Behrens 2022). This could be extended such that language models are interpretive layers on top of the sensorimotor system (Harnad 1990; Ahn et al. 2022; Driess et al. 2023; Brohan et al. 2023).

Conclusion

We presented **ReCoN-Ipsundrum**, a deliberately small and inspectable implementation inspired by Humphrey's staged account of sentience, built by extending a reflexive ReCoN sensorimotor substrate with (i) a recurrent persistence loop and (ii) an optional affect/interoceptive proxy inspired by constructionist affect. Across corridor and gridworld tasks, a mechanism-linked assay suite plus within-episode lesions supports clean component attributions: recurrence→post-stimulus persistence, and affect-coupled control→valence-stable scenic preference under novelty competition, structured local scanning in exploratory play, and lingering planned caution. These are engineered and dissociable signatures; we do *not* treat them as evidence of consciousness.

The broader lesson is about *attribution error* in machine-consciousness assessment. In AI, indicator-like behavior can be produced by minimal and potentially gameable implementations, while reliance on any single marker risks false negatives. We therefore recommend treating indicators as credence-shifting *conditional on explicit mechanistic hypotheses*, and pairing behavioral evidence with architectural inspection and causal interventions that target the posited mechanism. Normatively, as affect-like variables and valence-coupled control become easier to implement, cautious practice is warranted under moral-status uncertainty: minimize unnecessary exposure to sustained aversive dynamics, preserve reversibility/lesionability, and avoid anthropomorphic deployment claims that invite premature over-attribution.

References

Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; Finn, C.; Hausman, K.; Ichter, B.; Irpan, A.; et al. 2022. Do as I Can, Not as I Say: Grounding Language in Robotic Affordances. arXiv:2204.01691.

Bach, J. 2009. *Principles of Synthetic Intelligence: PSI: An Architecture of Motivated Cognition*. Oxford University Press.

Bach, J.; and Herger, P. 2015. Request Confirmation Networks for Neuro-Symbolic Script Execution. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (CoCo@NIPS 2015)*, volume 1583 of *CEUR Workshop Proceedings*.

Banino, A.; Barry, C.; Uria, B.; Blundell, C.; Lillicrap, T.; Mirowski, P.; Pritzel, A.; Chadwick, M. J.; Degris, T.; et al. 2018. Vector-Based Navigation Using Grid-Like Representations in Artificial Agents. *Nature*, 557(7705): 429–433.

Barrett, L. F. 2017. *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.

Behrens, T. E. J.; Muller, T. H.; Whittington, J. C. R.; Mark, S.; Baram, A. B.; Stachenfeld, K. L.; and Kurth-Nelson, Z. 2018. What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, 100(2): 490–509.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818.

Butlin, P.; Long, R.; Bayne, T.; Bengio, Y.; Birch, J.; Chalmers, D.; Constant, A.; Deane, G.; Elmoznino, E.; Fleming, S. M.; Ji, X.; Kanai, R.; Klein, C.; Lindsay, G.; Michel, M.; Mudrik, L.; Peters, M. A. K.; Schwitzgebel, E.; Simon, J.; and VanRullen, R. 2025. Identifying Indicators of Consciousness in AI Systems. *Trends in Cognitive Sciences*. Online ahead of print.

Chang, H. 2022. *Realism for Realistic People: A New Pragmatist Philosophy of Science*. Cambridge University Press.

Chirimuuta, M. 2024. *The Brain Abstracted: Simplification in the History and Philosophy of Neuroscience*. The MIT Press.

Cleeremans, A. 2011. The Radical Plasticity Thesis: How the Brain Learns to Be Conscious. *Frontiers in Psychology*, 2: 86.

Constantinescu, A. O.; O'Reilly, J. X.; and Behrens, T. E. J. 2016. Organizing Conceptual Knowledge in Humans with a Gridlike Code. *Science*, 352(6292): 1464–1468.

Driess, D.; Xia, F.; Sajjadi, M. S. M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. PaLM-E: An Embodied Multimodal Language Model. arXiv:2303.03378.

Friston, K. 2010. The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, 11(2): 127–138.

Hafting, T.; Fyhn, M.; Molden, S.; Moser, M.-B.; and Moser, E. I. 2005. Microstructure of a Spatial Map in the Entorhinal Cortex. *Nature*, 436(7052): 801–806.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, 42(1–3): 335–346.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Humphrey, N. 2023. *Sentience: The Invention of Consciousness*. The MIT Press.

Ladda, A. M.; Lebon, F.; and Lotze, M. 2021. Using Motor Imagery Practice for Improving Motor Performance – A Review. *Brain and Cognition*, 150: 105705.

Lamme, V. A. F. 2006. Towards a True Neural Stance on Consciousness. *Trends in Cognitive Sciences*, 10(11): 494–501.

Lau, H.; and Rosenthal, D. 2011. Empirical Support for Higher-Order Theories of Conscious Awareness. *Trends in Cognitive Sciences*, 15(8): 365–373.

O'Keefe, J.; and Dostrovsky, J. 1971. The Hippocampus as a Spatial Map: Preliminary Evidence from Unit Activity in the Freely-Moving Rat. *Brain Research*, 34(1): 171–175.

O'Regan, J. K.; and Noë, A. 2001. A Sensorimotor Account of Vision and Visual Consciousness. *Behavioral and Brain Sciences*, 24(5): 939–973.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*. NeurIPS 2022.

Pavlov, I. P. 1927. *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press. Translated and edited by G. V. Anrep.

Schulman, J. 2025. LoRA Without Regret. Connectionism (Thinking Machines Lab).

Schuster, C.; Hilfiker, R.; Amft, O.; Scheidhauer, A.; Andrews, B.; Butler, J.; Kischka, U.; and Ettlin, T. 2011. Best Practice for Motor Imagery: A Systematic Literature Review on Motor Imagery Training Elements in Five Different Disciplines. *BMC Medicine*, 9: 75.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Sherrington, C. S. 1906. *The Integrative Action of the Nervous System*. Yale University Press.

Varela, F. J. 1999. The Specious Present: A Neurophenomenology of Time Consciousness. In Petitot, J.; Varela, F. J.; Pachoud, B.; and Roy, J.-M., eds., *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*, 266–314. Stanford University Press.

Whittington, J. C. R.; Muller, T. H.; Mark, S.; Chen, G.; Barry, C.; Burgess, N.; and Behrens, T. E. J. 2020. The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, 183(5): 1249–1263.e23.

Whittington, J. C. R.; Warren, J.; and Behrens, T. E. J. 2022. Relating Transformers to Models and Neural Representations of the Hippocampal Formation. arXiv:2112.04035.

Wolpert, D. M.; Ghahramani, Z.; and Jordan, M. I. 1995. An Internal Model for Sensorimotor Integration. *Science*, 269(5232): 1880–1882.