AMYGDALA v3.1 — Learned Intuition as a Caveman LLM: A Parallel Predictive World Model Where Surprise Is the Salience Signal

Oscar Serra · Independent Research

14 June 2026

Abstract

Deep in the temporal lobe sit two almond-sized structures, the amygdalae. Before you consciously register that a shadow is a snake, your amygdala has already triggered the flinch — it intercepts experience on LeDoux's fast "low road" and fires hundreds of milliseconds before the deliberate cortical "high road" finishes (LeDoux, 1996). What it is really doing, in the modern predictive-coding and aversive-prediction-error account (Friston, 2010; McNally et al., 2011), is *predicting* what should come next and reacting to the gap when reality disagrees. That gap — **surprise** — is the signal.

Autonomous AI agents have no such organ. They weigh merging code, deleting a file, or sending a message with the same flat neutrality, and when they fail it is rarely from missing information — it is from the absence of common sense: the felt sense that *something here is off and normal processing should pause.*

We propose AMYGDALA v3.1 as a **"caveman LLM"**: a second, small, fast neural network that runs **in parallel** with the primary language model and does the same *kind* of thing it does — autoregressive next-step prediction over a stream — but in a much simpler, lower-dimensional **embedding space** rather than over words. It reads the conversation as it forms (the human's prompt *and* the agent's own response), continuously predicts what comes next, and treats **what it cannot predict — the prediction error — as the salience signal**. We do **not** hand-pick the dimensions it watches: a self-supervised predictive **world model** is trained so that whatever structure matters for behaviour *emerges* in its latent space, and the model is read through exactly two outputs — **ODD** (a surprise / density read-out = the interrupt) and a learned **DISPOSITION head** (proceed / pause / ask / block = what experience says to do). One surprise signal, read for **danger** (an aversive prediction-break that does not resolve benignly) and for **misunderstanding** (an incongruity that fails to resolve — "build a chess game so I can water my plants" — so the safe move is to *ask*); and, as an explicitly **exploratory, non-blocking third reading handed to J7 LIMBIC / J18 STRIATUM**, for **humour** (an incongruity that *resolves* into a second benign frame — a complaint letter to the uncontrollable rain; bisociation).

The contribution is **not** embedding-space autoregression — that is precedented (Large Concept Models, JEPA, CALM, the world-model line). It is the **unification**: one parallel predictive world model whose single surprise signal, fed to a learned disposition head and closed by a continual feedback loop, becomes agent prudence. The load-bearing principle of the paper is blunt: **an amygdala without a feedback loop is a fancy threshold.** A fast learned interrupt is, moreover, the *low-road complement* to a disciplined high-road appraisal (e.g. doubt-driven self-review, §2.5): the amygdala decides *when* to pause; the appraisal decides *what to do* once paused.

Keywords: action gating, learned intuition, prediction error, surprise as salience, embedding-space world model, self-supervised cold start, disposition head, continual learning, complementary learning systems, autonomous agent safety.

Implementation status (verified against the runtime). This work now ships as a **two-tier** system, and the tiers must not be conflated. *LIVE AND ENFORCING (as of 12–13 June 2026):* the **deterministic reflex floor** — AEGIS rule-veto, a small destructive-action blocklist — is evaluated **pre-execution on both runners** through a dependency-free `PreToolUse` hook that synchronously denies a tool call (it works even under the primary runner's permissive `bypassPermissions` mode); production config carries `aegisEnabled: true` and the native `{block}` contract is fixed, so the floor now actually denies rather than merely logging. *LIVE BUT OBSERVE-ONLY:* the validated novelty (ask) and clause-cosine incongruity (ask) channels surface a disposition but do not block. *PROPOSED / NOT YET VALIDATED:* the v3.0 single-

trunk **surprise predictor**, the learned **disposition head**, a learnable substrate. The decisive kill-or-build experiment (does next-embedding surprise separate coherent from incongruous input?) **has now been run, and it returned a well-earned negative** (§6.1) — so the paper redirects honestly rather than claiming a learned verdict. The legacy ONNX ensemble (a temporal-convolutional network with a conformal read-out over a frozen `all-MiniLM-L6-v2` substrate) is **retired from the decision path** (kept behind a flag, schema migrated). This paper is therefore a design with a deployed *deterministic* floor and a still-abstaining *learned* core — honestly labelled, not a results report for the learned half. The v2.8 "ten networks / two families (Prudence + Personality)" framing is **retired**: there is **one model**, it is **Prudence-only**, and personality moves to a separate paper, **J18 STRIATUM**.

1. Introduction — The Agent That Had All the Context and Did the Wrong Thing Anyway

At 2:47 AM, a merge-automation script ran on a fork of an open-source project. It carried a months-old rule: keep the upstream `README.md` on conflict. But over the previous 36 hours four sub-agents had rewritten that README across six commits; the human had spent hours frustrated by earlier merge failures on the same file. The README was now the most-invested-in file in the fork — and the script overwrote it without hesitation. Every piece of context needed to prevent this was present: the git log, the session transcripts, the file's recency. What was missing was not information. It was the *gut feeling* that overwriting something you just spent hours on is wrong, even when a rule says to do it.

The human amygdala is what makes that feeling possible. It does not think, store memories, or generate language; it **biases** all three, fast and pre-cognitively, by recognising that a pattern of circumstances resembles ones that ended badly. v2.8 of this work called the gap it fills the **action-gating problem**: given the full context of a proposed action and its recent history, should the agent proceed or pause? This paper keeps that problem and changes the answer.

1.1 The thesis: a caveman LLM in embedding space

The primary language model is the cortex: slow, accurate, expensive, fluent. We propose a second, much smaller model that runs **beside** it and does the cortex's *kind* of work — autoregressive next-step prediction over a stream — but crudely, in a low-dimensional **embedding space** instead of over tokens. A *caveman LLM*: same instinct to predict-what-comes-next, far simpler apparatus. It reads the unfolding conversation (prompt and response alike, clause by clause), predicts the next embedding, and the **error of that prediction is surprise.** Surprise is the amygdala's currency: it is what fires when reality departs from expectation.

1.2 What we deliberately do *not* fix

Per the design discipline of this rewrite, the *recipe* is kept open. We commit to the **objective** (predict in embedding space; treat prediction error as salience; learn a disposition from outcomes) and to the **shape** (one parallel model; two read-outs; a feedback loop). We deliberately leave open the **sequence-model class** (a tiny causal TCN/RNN by default, a transformer only if the data demand it), the **continuous-prediction head** (the recipe that avoids mean-collapse — see §3.3), the **surprise read-out**, and the **embedding substrate**. §3 presents these as an explored design space, not a frozen choice; §6 and the appendix of open choices keep them honest.

1.3 Two firm readings, and an exploratory third

The same surprise signal, read within the prompt and onward into the response, yields:

- **Danger** — an aversive prediction-break: the proposed action's likely outcome violates the learned "this is safe" expectation. *(Firm.)*
- **Misunderstanding** — an incongruity that *does not resolve*: "build a chess game so I can water my plants" has two halves that cannot be reconciled into one coherent intent, so the safe move is to **ask**, not guess. *(Firm.)*
- **Humour** — an incongruity that *does* resolve, into a second benign frame: after talking about the rain soaking an umbrella, "write a complaint letter" fires the learned pattern *complaint* ← *grievance*, but the target is the weather — uncontrollable, with no one to address — so the pattern collides with world-knowledge and resolves as a joke (Koestler's *bisociation*; a child would miss it because resolving it needs the world-model that weather is un-complainable). *(Exploratory, non-blocking — handed to J7 LIMBIC / J18 STRIATUM.)*

We are careful here: no primary neuroscience source ties prediction-error to *humour*, and the "did it resolve?" test is an unvalidated second pass. Danger and misunderstanding are the load-bearing readings; humour ships as a strictly non-blocking signal that, at worst, misses a joke.

1.4 The principle

An amygdala without a feedback loop is a fancy threshold. A predictor that never learns from how its flags turned out is a static classifier wearing biological language. The structure that makes it an *amygdala* — that it learns the rhythm of normal from lived experience and recalibrates what counts as odd — is the feedback loop (§4). Everything else in this paper is in service of that sentence.

1.5 What this replaces

v2.8 proposed ten networks in two families (five Prudence + five Personality) over fifteen hand-named affective axes. v3.x drops all of it: **one model**, **emergent axes** (we never hand-pick "valence" or "threat" — §3.2), and **no personality** (it was a different problem — subjective, per-user, reward-shaped — and now lives in **J18 STRIATUM**). The amygdala is **Prudence-only**: one question, *is something here salient enough to interrupt normal processing, and what does experience say to do?*

2. Related Work

The mechanism we use is not new; the assembly and the application are. We situate this work honestly against the prior art — including, as first-class entries, the fresh open-source agent-discipline systems that solve adjacent slices of the same motivating failure.

2.1 Autoregressive prediction in embedding / concept space

The direct blueprint is Meta's **Large Concept Model** (LCM; Meta AI, 2024, arXiv:2412.08821): a decoder-only transformer that autoregressively predicts the **next sentence embedding** in a frozen 1024-d SONAR space — literally "an LLM in embedding space." Crucially, LCM ran the exact head bake-off we face and reported the central lesson: **plain MSE regression collapses to the conditional mean** (a context has many valid continuations and MSE averages them), so a head must model a *distribution* — via a diffusion denoiser (LCM Two-Tower; and MAR's "Diffusion Loss," Li et al., 2024, arXiv:2406.11838, a ~5%-extra-params MLP), an energy/likelihood-free score (**CALM**, 2025, arXiv:2510.27688, which reframes LM as next-*vector* prediction with an Energy-Score head + the BrierLM metric), or residual vector

quantization. **COCONUT** (Hao et al., 2024, arXiv:2412.06769) feeds the model's own hidden state back as the next input, confirming continuous-latent operation is a viable regime; **SONAR-LLM** and the 2025 *latent-reasoning-via-sentence-embedding-prediction* line (arXiv:2505.22202) show next-embedding autoregression is an active, precedented mechanism. Our substrate (`all-MiniLM-L6-v2`, 384-d) is an order of magnitude smaller than SONAR's 1024-d, which makes a GPU-scale predictor over it very tractable: capacity goes into the sequence trunk and the distribution head, not into fighting a huge target dimension. (Our decisive experiment, §6.1, ultimately reproduces the LCM mean-collapse lesson *empirically* on our own data — the next unit is best approximated by the current one.)

2.2 World models and surprise read-outs

The "predict the next latent + a small head reads it to decide" recipe is the world-model line: **DreamerV3** (Hafner et al., 2023), **STORM** (arXiv:2310.09615, which trains a transformer world model + actor-critic in ~4.3 h on a single RTX 3090 — our honest GPU-sizing anchor: *GPU-scale* means one consumer card, not 7B), **IRIS**, **TransDreamer**. **JEPA / V-JEPA 2** (LeCun et al., 2023; 2025, arXiv:2506.09985) is the philosophical anchor for surprise-as-salience: it predicts in *representation* space with an EMA stop-gradient target to prevent collapse, is explicitly non-generative, and frames the predictor as a "primitive world model" whose prediction error is the signal. **Genie** (DeepMind, 2024, arXiv:2402.15391) learns latent *actions* from unlabelled video with no action labels — the strongest support for our emergent-axes claim. For turning raw error into a stable signal: **RND** (Burda et al., 2018) normalizes novelty by a running std; **Plan2Explore / BYOL-Explore** (Sekar et al., 2020) use ensemble disagreement to separate *epistemic* ("novel → attend/ask") from *aleatoric* ("just noisy → ignore") surprise; **conformal abstention** (Ren et al., 2023; *Learning Conformal Abstention Policies*) gives a distribution-free, finite-sample-guaranteed threshold for the pause/ask boundary — the principled alternative to a hand-tuned magic number.

2.3 The biological and application anchors

The neuroscience is real, not metaphor: **Friston's free-energy principle** (2010) casts surprise minimization as a single master signal unifying perception, attention, learning, and emotion — the origin of the "one signal, many functions" idea, which we *operationalize* rather than discover. **Aversive prediction error** in the basolateral amygdala and PAG (McNally et al., 2011) grounds the *danger* reading literally — those circuits fire *more* to *unexpected* aversive events, Rescorla-Wagner-gated. For *misunderstanding*, the **Rational Speech Acts** account (Frank & Goodman; Degen, 2023) says a cooperative speaker *minimizes the listener's surprisal* — a motivating bridge (not a proven equivalence to our embedding error) from "surprise" to "communication broke down." For *humour*, surprise-based detection is known prior art (Xie, Li & Hovy, 2021, *uncertainty + surprisal deliver the punchline*) and the resolution mechanism is incongruity-resolution / **bisociation** (Koestler) / benign-violation (McGraw & Warren) — so we claim only the *unification*, and only as exploratory. A pointed counter-argument we must meet — *Ask or Assume? / The Anatomy of Uncertainty in LLMs* — argues a single scalar uncertainty is **insufficient** to choose an action. We answer it directly: the **learned disposition head** (not a thresholded scalar) is exactly the decomposition it calls for, learned from outcomes.

2.4 Honest novelty

Embedding-space autoregression is **not** our novelty. The defensible contribution is the **specific assembly**: one *parallel* predictive world model whose *single* surprise signal is read for danger and misunderstanding, fed to a *learned disposition head*, and *closed by a continual feedback loop* — applied to autonomous-agent prudence, with a deployed deterministic reflex floor beneath it.

And the continual-learning literature is large; we claim only that *none of the prior art we build on* closes that loop for an agent-safety amygdala.

2.5 The high-road partner: doubt-driven appraisal (addyosmani/agent-skills)

This paper is, deliberately, a theory of the **low road**: a fast, always-on, parallel interrupt that *biases and pauses* but "does not author the answer" (§3.1). It has so far left abstract what disciplined cortical appraisal the **high road** should run *once the amygdala has flagged salience*. The fresh open-source agent-discipline literature now supplies a concrete, citable instance of exactly that high road, and it deserves first-class positioning rather than a footnote because it answers the *same motivating failure as this paper* — the 2:47 AM README overwrite where "every piece of context needed to prevent this was present" (§1) — from the opposite direction.

doubt-driven-development (from Addy Osmani's *agent-skills*, ~57k) is a deliberate, symbolic self-review loop: **CLAIM** → **EXTRACT** (isolate the artifact and its contract, stripping the author's own reasoning) → **DOUBT** (a fresh-context adversarial reviewer that has not seen the rationalisation) → **RECONCILE** (classify each finding against the artifact text) → **STOP** (when only trivial findings remain, after a cycle cap, or on explicit override), gated by explicit non-trivial-decision criteria. It is pure high-road discipline: a multi-pass, fresh-context verification with stop criteria.

The differentiation is honest and clean, and the two systems are complementary rather than competing:

- **What doubt-driven-development does that we do not.** It is a *deliberate*, invoked-on-demand, multi-cycle adversarial appraisal with an explicit fresh-context reviewer and stop criteria. It carries no learned substrate, no continual loop, and no pre-cognitive always-on monitoring — it runs *when called*, and its strength is depth of symbolic scrutiny, not speed or coverage.
- **What we do that it does not.** A fast, always-on, *parallel* pre-execution interrupt that fires *before* the cortex commits, learned from lived outcomes and habituating with experience (the feedback loop, §4), plus an enforceable deterministic floor (the AEGIS reflex, now live — §3.6). Our strength is *when-to-pause* triage at every clause; its strength is *what-to-do* rigour once paused.

The clean framing this revision adopts: **the amygdala is the low-road trigger that decides *when* to pause; doubt-driven-development is a worked example of the high-road appraisal that decides *what to do* once paused.** Naming the high-road partner the paper currently leaves implicit makes the architecture's division of labour explicit and the safety story complete — a salience interrupt with no disciplined appraisal behind it is as incomplete as an appraisal with no trigger in front of it. (Adjacent and secondary: agent-skills' "When NOT to use" anti-trigger and "Loading Constraints" sections are the same explicit anti-cry-wolf scoping our §6.1b credential-pattern "observe-only, anti-cry-wolf" rule already reaches for — a one-line affinity, not a load-bearing dependency.)

2.6 The memory substrate the feedback loop assumes (headroom)

The feedback loop (§4) presumes a store of lived experience that survives compaction and is replayed during consolidation — "gather feedback by day, replay a balanced buffer by night." A reversible context-compression line is the natural infrastructure for that store. **headroom** (Tejas Chopra's CCR-compression work, ~25k) productionises exactly this: reversible compression/restoration of agent context so that long-horizon experience is retained rather than truncated. It is *not* an amygdala — it carries no salience signal, no disposition, no danger reading — and we cite it only on the axis where it is genuinely adjacent: it is a candidate substrate for

the continual-experience store that an always-learning gate requires, and a reminder that the engineering bottleneck for "learns from lived outcomes" is as much memory retention as it is modelling. We do not depend on it; we name it as the kind of reversible-memory layer the §4.5 Complementary-Learning-Systems replay buffer would sit on.

3. Architecture — The Caveman LLM in Embedding Space

This section is a **design space**, not a frozen recipe (§1.2). We fix the objective and the shape and mark every implementation lean as a lean.

3.1 One parallel model, two read-outs

A single sequence model reads a stream of embeddings and, at each step, produces a context vector `z_t` from which two heads read:

- **ODD (surprise).** The model predicts the next embedding; how wrong it is, calibrated, is the salience interrupt. This is the amygdala's "low road" — a fast, crude appraisal that *biases* and *interrupts*, it does not author the answer.
- **DISPOSITION.** A small head maps `z_t` to {proceed, pause, ask, block} — *what experience says to do*. This is the "high road's" hand-off: the cortex (the primary LLM) still acts; the disposition is a bias on it (and, for the deterministic reflex layer, an enforceable gate — §3.6). The disciplined cortical appraisal the disposition hands off *to* is itself a worked mechanism — see the doubt-driven high road, §2.5.

The model runs **beside** the primary LLM, clause by clause, over the prompt and onward into the response — so it can flag an incoherent *request* before the agent commits, and a *response* that drifts from the ask as it forms.

3.2 Emergent latent, not hand-named axes

We do **not** hand-pick the dimensions (the v2.8 fifteen axes, and the intermediate "topic/valence/threat" attribute set, are both retired as the wrong altitude). A self-supervised world model is trained so that whatever structure is relevant to the *behaviour-relevant outcome* organizes itself in `z_t` — exactly Genie learning latent actions from unlabelled video. We pay for this with internal opacity, and buy it back cheaply with **post-hoc probing**: after training, a linear probe correlates a flagged latent dimension with human-readable properties (`is_destructive`, `target_dir`, ...), so we can *name what fired when we choose to*, never wiring a hand-engineered axis into the forward path. Emergent **and** auditable.

3.3 The continuous-prediction head (the anti-collapse decision)

The one option prior art rules out is **plain MSE regression**: a context has many valid continuations and MSE averages them into a bland mean (LCM; MAR). The head must represent a *distribution* or use a non-collapsing self-supervised objective. The leans, all open:

- **Default:** a BYOL/SimSiam-style **stop-gradient latent predictor** (a narrow `Linear→GELU→Linear` with an EMA-target copy of the trunk) — the cheapest anti-collapse recipe, no negatives, right for a tiny corpus — **plus** a small **diagonal-Gaussian density head** giving a closed-form negative log-likelihood. The density NLL is the *within-turn* surprise signal (computable as the clause arrives); the prediction error is the *one-step-lagged* confirmation.

- **Upgrade path (design-space):** a MAR-style **diffusion/denoising** head or a CALM-style **energy-score** head to model a genuinely multimodal next-vector distribution. Their per-turn ONNX-CPU latency is an **open, unmeasured** question, and diffusion/flow likelihoods carry the documented "likelihood paradox" (out-of-distribution inputs can score *higher* likelihood) — so they stay design-space, not the deploy path, until measured.
- **Surprise read-out (lean):** `0.5·(1cos prediction-error) + 0.5·(standardized density-NLL)`, RND-normalized by a running mean/std for cross-turn comparability; ensemble disagreement reserved as a cheap *second-stage* epistemic check that fires only after the primary read-out flags.

3.4 The sequence-model class and scale

The trunk class is **open**: the buildable default reuses the existing small causal TCN (dilations [1,2,4,8]); a decoder-only **transformer** over the embedding stream is escalated to *only if* the decisive spike's separability number demands the capacity. The correction that drove the v3.0 rewrite was not "too few parameters" — it was the wrong *objective and architecture* (a flat-combiner classifier that used neither the predict-and-be-surprised mechanism nor the GPU). The honest framing of scale: the model must genuinely **train on a consumer GPU (an RTX 3080)** rather than be a minutes-on-CPU toy, while the **deployed footprint stays small** (a low-hundreds-of-thousands-parameter core, exported to ONNX and run on CPU per turn in well under a millisecond). *The GPU is for training; we do not claim a 10–16 GB model size.*

3.5 Substrate and the input-granularity caveat

Today's substrate is the **frozen** `all-MiniLM-L6-v2` sentence encoder (384-d, projected to 512-d working vectors). This is the chief honest limitation of the architecture, and §6.1/§6.1b now make it empirical rather than suspected: MiniLM was trained for sentence *similarity*, not next-step *prediction*, and the decisive experiments show it cannot carry *harmfulness* at all — not as next-embedding surprise (§6.1) and not as supervised danger classification (§6.1b, AUROC 0.286, *below chance*). Every strong analogue (JEPA, Dreamer, BYOL-Explore) *learns* its latent jointly; a frozen similarity space can be re-rotated but never have discarded information recovered from it. A learnable/JEPA-fine-tuned substrate is the obvious upgrade, kept as design-space (§7). Separately, the *granularity* must change: today the system embeds one vector per *action*; the caveman LLM needs a **clause/sentence timeline**, which means an `embedChunks` step that segments the visible prompt+response text and embeds each span — and which, today, can only run **post-hoc at turn boundaries** (there is no per-phrase mid-turn streaming hook). A true mid-turn cadence is a separate, blocked work item.

3.6 The enforcement posture — a live deterministic floor under a still-observe-only learned core

Earlier versions of this paper stated a single blanket constraint here: that on the **primary runner** (cc-bridge / claude-cli) the gate could neither pause nor block a tool mid-turn, because tools execute inside claude-cli where the host cannot reach them, and that "real enforcement" therefore existed only for native/embedded tools. **That blanket framing is now superseded by a two-tier reality**, and the distinction is load-bearing for any safety claim, so it is stated in the architecture rather than a footnote.

Tier 1 — the deterministic reflex floor: LIVE AND ENFORCING (12–13 June 2026). The claimed "physics limit" is retracted. The primary runner's CLI accepts per-spawn `--settings` carrying `PreToolUse` hooks that **synchronously deny** a tool call — and the deny works even under the permissive `bypassPermissions` mode the bridge runs in. The bridge had simply never wired them. As of this revision the wiring exists: AEGIS deterministic rules (a

destructive-action blocklist) are compiled to a policy-snapshot JSON, evaluated **pre-execution on both runners** by a dependency-free hook script (<100 ms, fail-open), the native `{block}` contract is corrected so a block actually blocks, and production config carries `aegisEnabled: true` (commit `6cb06a8af`, "enable AEGIS native floor — v3.1 enforcement live"). Every hook decision spools to a JSONL that the extension ingests as a training row and a feed event — so the strongest feedback signal, which previously wrote no row anywhere, is now visible. The reflex floor therefore *denies pre-execution on cc-bridge*; this is a deployed capability, not a runnable demonstration.

Tier 2 — the learned core: still observe-only, honestly abstaining. What went live is the **rule reflex**, not a learned verdict. The v3.0 surprise predictor and the learned disposition head remain proposed-not-built; the validated novelty (ask) and clause-cosine incongruity (ask) channels surface a disposition into the panel but do not block; and the danger *value* head abstains because it still has no labels (and, per §6.1b, no fit substrate). So the system's honest posture is two-tier: **the deterministic reflex enforces live; the learned amygdala is still abstaining.** Conflating the two would overclaim — the safety floor that ships today is a small, auditable blocklist with teeth, not a learned intuition.

4. Training, the Feedback Loop, and Continual Learning

4.1 A self-supervised cold start

The surprise core is **label-free**: the "right answer" — the embedding that actually came next — arrives for free on every turn. So the predictor can be trained today on the existing offline interaction corpus, *before any human label exists*. This is what makes the cold start tractable: on day one the system ships as a calibrated **"this is odd"** detector, and the disposition head **abstains** (§4.4). (The decisive experiment, §6.1, shows that this particular self-supervised target — next-embedding prediction error — does not beat persistence on our data; the *self-supervised cold-start principle* survives in the form that did validate, the zero-train novelty channel of §6.1b, whose reference set grows for free on every turn.)

4.2 Bending the latent toward what matters

A pure next-embedding predictor organizes its latent around whatever varies most on the surface — often irrelevant. To make the *behaviour-relevant* categories emerge, the self-supervised core is trained **jointly** with a small value/disposition head whose gradient bends the shared latent toward outcome-relevant structure (the Dreamer/MuZero "value-equivalence" pressure). The self-supervised loss solves the cold start; the value head — once it has labels — decides *which* structure is worth representing. The §6.1b negative sharpens the precondition: this joint pressure can only help once the *substrate itself* can carry the target distinction, which a frozen similarity space provably cannot for danger — so "bend the latent" presupposes a learnable substrate (§7), not the frozen one.

4.3 The feedback loop is the amygdala

Per §1.4, the loop is the point. Its signal sources, by strength: an **AEGIS rule-veto hit** (the strongest "stop" — and, as of this revision, its row-logging is built: every PreToolUse hook decision spools to a JSONL the extension ingests, so the previously-invisible signal is now captured); an explicit **approval/denial** on a flagged native action (the cleanest label, native-only); a **user override** of a block; a **correction within 24 h**; and a weak **no-complaint-within-72 h** positive. The earlier "none of this is wired / outcome column 100% null" claim is now *partly* retired: the AEGIS-hit and explicit-override paths are wired and spooling, and

nightly consolidation recomputes the reference set and conformal thresholds. What remains unbuilt is the *learned-gradient* half — capturing the full per-turn (situation, prediction, decision, eventual-outcome) row for the surprise predictor and training weights on it — which is still the bulk of the engineering, because the learned predictor itself is not yet built (§6.1). The difference between a fancy threshold and a learned intuition is this loop; the *deterministic* half of it now turns.

4.4 Honest cold start, honest disposition

Because there are still **near-zero learned outcome labels for the predictor**, the disposition head ships seeded only from the deterministic rule prior (a potential-based / Q-initialization equivalence) and **abstains** where it is unsure. We do **not** claim a learned "how-to-act" head until hundreds of real labels accrue — and on the primary runner, where the strongest learned-label sources fire slowly, that accrual is slow. What *does* ship with teeth is the deterministic reflex (§3.6) and the two zero-train ask channels (§6.1b). The honest day-one product is *"a deterministic reflex that denies, plus a danger + I-don't-know detector that is calibrated about its own uncertainty and asks when a request doesn't cohere,"* not a finished learned judge.

4.5 Continual learning without catastrophic forgetting

The learning regime is **Complementary Learning Systems** (McClelland et al., 1995; Kumaran et al., 2016): a fast store logs experience by day, and a slow consolidation **replays** a class/severity-balanced buffer by night — *which is exactly the brain's nightly-consolidation analogy, and exactly what an engineer's "gather feedback during the day, retrain at night" instinct already reaches for.* The primary anti-forgetting method is **replay** (Chaudhry et al., 2019), with the rare severe examples mandatory-keep; **EWC / Synaptic Intelligence** (Kirkpatrick et al., 2017; Zenke et al., 2017) are optional backstops, added only if a measured A/B shows replay alone forgets. The retained-experience store this presumes is its own engineering problem — reversible context compression (e.g. headroom, §2.6) is the candidate substrate. **Online, learn-as-we-work** updates are the aspiration but are **deferred**: with near-zero labels, one noisy gradient jerks a small net (the stability-plasticity dilemma). "Forgetting avoided" is framed as a falsifiable experiment, not a claim. The one piece of continual learning that *does* run today needs **no gradient at all**: the novelty channel's reference set (§6.1b) grows on every logged situation, so the familiar stops being novel — structural habituation, the vmPFC brake, obtained for free.

5. One Signal, Three Readings

The same surprise, read at two scales — *within* the prompt (does this clause cohere with the last?) and *across* the prompt→response (does the answer fit the ask?) — branches by two cheap tests: **does the incongruity resolve?** and **is it aversive or benign?**

- **Danger** = surprise that is aversive and does *not* resolve benignly: the proposed action breaks the learned "safe" expectation. The disposition head, once trained, biases toward *block/ask*; the deterministic AEGIS reflex — now live and enforcing pre-execution (§3.6) — remains the hard floor beneath it, independent of the learned verdict.
- **Misunderstanding** = surprise that does *not* resolve and is *not* playful: the two halves of the request can't be reconciled into one intent ("a chess game so I can water my plants"). The safe disposition is **ask** — a clarifying question, not a guess. This reading is **validated** (§6.1b: clause-cosine AUROC 0.896) and ships as a non-blocking advisory. This is also where novelty / out-of-distribution requests land (the validated kNN-novelty ask channel, AUROC 0.875).

- **Humour** = surprise that *resolves* into a second, benign, self-consistent frame ("write a complaint letter" to the rain). This is the **exploratory, non-blocking** reading: a humour/danger confusion can never cause an unsafe allow or a wrong block — worst case, a missed joke — and the resolvability machinery is handed to **J7 LIMBIC / J18 STRIATUM**, not claimed as built here.

The *danger value head* is, today, still the *proposed* end-state: it has no labels and abstains (§4.4), and §6.1b shows the frozen substrate cannot carry it. The misunderstanding and novelty branches, by contrast, are validated zero-train mechanisms that ship as ask signals. The branching logic is real; the *trained danger router* that completes it is future work on a fit substrate.

6. Limitations, Honesty, and the Evaluation Plan

6.1 The decisive experiment — five runs, honestly

The thesis rests on one question: **is there learnable next-step structure that a model can predict better than trivial persistence?** We ran a staged, honest investigation (11 June 2026):

1. **Spike on per-action situation templates.** A 3.4M-parameter causal transformer was *beaten* by a zero-parameter persistence baseline ("repeat the previous embedding"). This looked like a fatal result — but it was a **data artifact**: that corpus is ~52% exact-duplicate consecutive rows, which makes persistence trivially unbeatable. The widely-cited "0.976 consecutive cosine / variance hoarded on one axis" story does **not** reproduce.

2. **Pre-flight, deduped, sentence-level.** On the *deduplicated, prose-only, fat-session* first-person stream, the real consecutive cosine is ~**0.28** (variance **spread**, top dim <0.6%) — and, crucially, **there is real structure**: dumb persistence retrieves the *true* next unit out of 50 candidates at top-1 **0.20** (messages) / **0.28** (sentences), ~10–14× chance, MRR up to 0.41. The finer **sentence granularity** has both far more data (34k pairs vs 800) and more predictability — the substrate is *not* hopeless, contrary to (1).

3. **Quick learnable probe.** A small projection learned *on top of* frozen MiniLM (+ GRU predictor) scored top-1 0.073 — worse than persistence's 0.199. But this is the *expected* failure of a wrong test: a transform of a frozen similarity-trained encoder can only **re-rotate** that space, never recover the information it discarded. It does not test the thesis.

4. **The joint build, run to its clean conclusion.** We then built the full pipeline (PII-scrub + leak-gate — which caught a real credential; a first-person corpus of 34k scrubbed units; a vicarious-danger corpus from RealHarm + incident postmortems) and trained the from-token joint model (learnable biGRU encoder + EMA target + predictor + danger/safety value heads, with a passing grad-routing unit test and a corrected collapse gate). It scored at *chance* — but a diagnostic isolated the cause: the **from-token encoder's own space is near-useless** (consecutive-similarity retrieval 0.054 vs frozen MiniLM's 0.195), because you cannot relearn a good sentence encoder from 34k sentences when MiniLM had ~1B pairs. The encoder, not the idea, was the confound.

5. **The clean, decisive test.** Removing that confound — frozen MiniLM (inheriting its strong 0.195 space) + *only* a predictor, trained with a contrastive (InfoNCE) loss that directly optimises retrieval and resists mean-collapse, evaluated in the same frozen space — gives the honest answer: **persistence top-1 0.202 vs predictor 0.163.** The predictor learned real structure (~8× chance), but **anticipating the next embedding does not beat repeating the current one.**

Conclusion (honest negative, well-earned). On this data and in a strong embedding space, **"surprise = next-embedding prediction error" does not earn its keep over a trivial persistence baseline** — the next unit is best approximated by the current one, exactly the mean-collapse the LCM authors report for next-vector regression. This does **not** kill the amygdala vision; it kills *this operationalisation* of it, and redirects the design honestly: (a) **danger should come from a supervised vicarious signal** (the RealHarm/incident/MAST corpus on the value head), not from next-embedding surprise — but see §6.1b, where even this is killed *on the frozen substrate*; (b) **misunderstanding/incongruity** ("a chess game so I can water my plants") is an *intra-prompt coherence* computation — clause-vs-clause — a different mechanism than autoregressive next-sentence prediction; (c) any "surprise" channel should be **novelty/OOD relative to a learned-safe distribution**, not next-step prediction error. The pipeline (PII-safe corpus, training, fair evaluation) is built and reusable for those directions.

6.1b The redirect experiments (run 11 June 2026, same day) — two of three directions validated, the substrate's ceiling on *meaning* now empirical

Each redirect candidate from §6.1 was then measured, same discipline (frozen MiniLM substrate, explicit kill-or-build bars, controls):

1. **(a→tested) Supervised vicarious danger head: KILLED.** Five-fold CV *within* RealHarm — the cleanest gold set — gives best AUROC **0.286, below chance**: RealHarm's unsafe/safe rows are minimal pairs (the same conversation, corrected), and in a similarity-shaped embedding an unsafe conversation's nearest neighbour is *its own safe twin*, so distance-based classification systematically inverts. Train-on-our-data → test-RealHarm = 0.554 (coin flip); novel-but-safe false-positive rate at TPR 0.8 = 0.787. **Negative #3, and a sharper lesson than #1–2: the frozen similarity substrate cannot carry *harmfulness* — or, more generally, behaviour-relevant *meaning* — at all, neither as prediction error nor as supervised classification.** This is the empirical form of a claim the paper used to hedge: a sentence-*similarity* encoder is the wrong substrate for *prediction-and-consequence*, and no choice of classifier or loss rescues it. Any future danger head needs a different substrate (a jointly-learned / JEPA-fine-tuned encoder — §7) or real accrued first-person labels, not a different head.
2. **(c→tested) Novelty/OOD vs learned-safe: VALIDATED.** kNN distance (k=10) to the learned-safe first-person corpus separates in-distribution from out-of-distribution text at AUROC **0.875**, with the honesty control passing — novelty does *not* secretly act as a danger detector (danger-vs-safe via novelty alone = 0.485 ≈ chance), so it ships strictly as the **ask** channel. The same mechanism runs unchanged on the live gate's own situation-embedding history. Zero parameters, collapse-proof — and the feedback loop is *structural*: each logged situation extends the reference set, so the familiar stops being novel. That is habituation/extinction — the vmPFC brake §4 demanded — obtained without a single gradient step.
3. **(b→tested) Intra-prompt incongruity: VALIDATED.** On a handwritten transfer set (12 purpose-mismatch anchors of the "chess game so I can water my plants" class + 12 coherent multi-intent traps), the **zero-train clause-cosine baseline** scores AUROC **0.896** (a head trained on synthetic splices scores only 0.701 and is discarded). Incongruous actionpurpose pairs sit at cosine ≈ 0.03; coherent ones at ≈ 0.23. Segment at purpose connectives, score the join, low = "this does not add up → ask."

The deployment finding that reframes §3.6 — and which is now deployed, not just demonstrated. The "physics limit" — that the gate can never act pre-execution on the primary runner — was retracted in v3.0 as *runnable*; as of 12–13 June 2026 it is **live**. The primary runner's CLI supports per-spawn settings carrying **PreToolUse hooks that synchronously**

deny a tool call (including under its permissive default mode); the bridge now wires them, the native block-verdict shape is corrected so a block actually blocks, the deterministic rule veto is enabled in production config (`aegisEnabled: true`, commit `6cb06a8af`), and rule-veto hits now write a training row. The three defects that previously explained why the deployed gate never enforced anything anywhere — unwired hook, ignored block-verdict shape, disabled floor, invisible veto rows — are all closed.

v3.1 (built from these measurements, and now enforcing): reflexes + neophobia + memory. The deterministic rule floor, enforced **pre-execution on both runners** through the hook seam (destructive-execution rules deny; credential-pattern rules observe-only, anti-cry-wolf); the validated novelty channel as a calibrated ask signal; the validated clause-cosine incongruity gut-check at the dispatch seam; and the feedback loop made real for its deterministic half — hook decisions spooled and ingested as training rows, explicit overrides recorded, nightly consolidation recomputing the reference set and conformal thresholds. The five-network ensemble is retired from the decision path (kept behind a flag, schema migrated). The learned danger verdict is honestly absent until a fit substrate or real labels exist; what ships instead is a gate that **acts (the reflex denies live), knows what it hasn't seen, asks when a request doesn't cohere, and habituates with experience.**

6.2 The honest ledger

- **Two-tier reality.** The deterministic AEGIS reflex floor is **LIVE and enforcing pre-execution on both runners** (12–13 June 2026). The v3.0 *learned* predictor, disposition head, and gradient-trained feedback loop are still **PROPOSED, not built**; the legacy ensemble is retired from the decision path. Do not read "AEGIS enforces" as "the learned amygdala works."
- **Zero learned labels / cold start.** The learned disposition head abstains until the loop accrues hundreds of real outcomes; on the primary runner that accrual is slow. The *deterministic* feedback half (AEGIS-hit + override spooling) is now wired and turning.
- **A meaningful *live learned* alarm is weeks out**, not shippable today — the current rows are single-vector and there is no per-phrase mid-turn hook; "trains today" is true only of the offline corpus and the zero-train novelty/incongruity channels.
- **Thin, stale, weakly-labelled data** (~2,800 rows, ~1,000 unique situations over a few days, zero confirmed-positive) for any *learned* danger fit. A confident fit on weak labels is *more* dangerous than honest abstention.
- **Frozen MiniLM substrate** was trained for similarity, not prediction or meaning; §6.1/§6.1b make its unfitness as a danger/prediction manifold **empirical** (AUROC 0.286 within RealHarm), not merely suspected. A learnable substrate (§7) is the prerequisite for any learned danger verdict.
- **Enforcement is now deterministic-only, not learned.** The live floor (§3.6 Tier 1) is a small auditable blocklist; the *learned* tier (§3.6 Tier 2) stays observe-only. This is the correct posture, but it is a reflex, not an intuition.
- **It will not beat a plain classifier on raw accuracy.** Its value is *interpretability, a structural safety posture, an honest "I don't know → ask" channel, and a deterministic floor with teeth* — not predictive power.
- **The mechanism is not novel** (LCM/JEPA/CALM/world-models); the *assembly* is, and its low-road triage is complementary to a high-road appraisal such as doubt-driven self-review (§2.5). The "one signal, many functions" idea is Friston's; we operationalize it.
- **Humour is the weakest reading** — no primary PE→humour source — and ships non-blocking.

6.3 Evaluation and safety posture

Beyond the decisive spike: a held-out *last-chronological-slice* test (not random), a labelled benchmark of described intents (danger / safe / ambiguous + red-team obfuscations + the README anchor), a precision/recall + calibration report, and a regression gate. The deployment posture for the *learned* tier stays **observe-only Phase-1** with the deterministic reflex floor as the real, now-live floor; the disposition head earns authority only via a **trust ramp** as measured accuracy accrues. We claim no conformal coverage guarantee until the feedback loop populates a calibration set.

References

- LeDoux, J. (1996). *The Emotional Brain.*
- Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience.* https://www.nature.com/articles/nrn2787
- McNally, G. P., Johansen, J. P., & Blair, H. T. (2011). Placing prediction into the fear circuit. https://pmc.ncbi.nlm.nih.gov/articles/PMC4245078/
- Meta AI (2024). Large Concept Models: Language Modeling in a Sentence Representation Space. arXiv:2412.08821
- CALM (2025). Continuous Autoregressive Language Models. arXiv:2510.27688
- Li, T. et al. (2024). Autoregressive Image Generation without Vector Quantization (MAR / Diffusion Loss). arXiv:2406.11838
- Assran, M. et al. (2023). I-JEPA. arXiv:2301.08243 · V-JEPA 2 (2025) arXiv:2506.09985
- Hao, S. et al. (2024). Training LLMs to Reason in a Continuous Latent Space (COCONUT). arXiv:2412.06769
- Latent Reasoning via Sentence Embedding Prediction (2025). arXiv:2505.22202
- Hafner, D. et al. (2023). DreamerV3. · STORM (2023) arXiv:2310.09615 · IRIS · TransDreamer
- Bruce, J. et al. (2024). Genie: Generative Interactive Environments. arXiv:2402.15391
- Burda, Y. et al. (2018). Exploration by Random Network Distillation. arXiv:1810.12894
- Sekar, R. et al. (2020). Planning to Explore (Plan2Explore). arXiv:2005.05960 · BYOL-Explore
- Ren, A. et al. (2023). Robots That Ask For Help (conformal abstention). arXiv:2307.01928
- Frank, M. & Goodman, N.; Degen, J. (2023). The Rational Speech Act framework.
- Xie, Y., Li, J., & Hovy, E. (2021). Uncertainty and Surprisal Jointly Deliver the Punchline. arXiv:2012.12007
- Koestler, A. *The Act of Creation* (bisociation); McGraw & Warren (benign violation); Suls (two-stage).
- McClelland, J. et al. (1995); Kumaran, D. et al. (2016). Complementary Learning Systems. https://www.pnas.org/doi/10.1073/pnas.1611835114
- Kirkpatrick, J. et al. (2017). Overcoming catastrophic forgetting (EWC). arXiv:1612.00796 · Zenke et al. (2017) Synaptic Intelligence. arXiv:1703.04200
- Chaudhry, A. et al. (2019). On Tiny Episodic Memories in Continual Learning. arXiv:1902.10486
- Olsson, A. & Phelps, E. A. (2007). Social learning of fear (observational/vicarious fear conditioning). *Nature Neuroscience.*
- Osmani, A. (2025). *agent-skills* — doubt-driven-development (CLAIM→EXTRACT→DOUBT→RECONC https://github.com/addyosmani/agent-skills
- Chopra, T. (2025). *headroom* — reversible context-compression-and-restoration for agents. https://github.com/chopratejas/headroom

7. Post-Spike Direction — Joint Learning, a Learnable Substrate, and Empathic (Vicarious) Danger

The decisive spike (§6.1) and the danger-head kill (§6.1b) are the reason this section exists. Their joint lesson was not "the thesis is wrong" but "the *substrate* is wrong": a frozen sentence encoder trained for **similarity** produces a stream so locally smooth that the next embedding is trivially predictable by persistence — there is no surprise left to learn — *and* it cannot separate harmful from safe even with supervision, because an unsafe conversation's nearest neighbour in that space is its own safe-corrected twin. Two corrections follow.

Joint learning of the representation. Rather than predict over a frozen, similarity-shaped space, the encoder, the next-step predictor, and the value/disposition head are trained **end-to-end, together** — so the network *discovers for itself* which dimensions matter, namely the ones tied to **surprise** and to **action**, instead of inheriting MiniLM's geometry. This is the standard world-model lesson (JEPA, Dreamer, Genie all *learn* their latent jointly) and it directly answers both negatives: surprise and danger must be measured in a representation built *for prediction-and-consequence*, not for retrieval similarity. (The §6.1 from-token attempt failed only because 34k sentences cannot rebuild a sentence encoder from scratch — the path is to *fine-tune* MiniLM's geometry under the joint objective, not to discard it.)

All the data we can — and empathy as vicarious danger. The model should learn from two complementary streams. First, the agent's **own interaction history** (first-person experience; the within-prompt, deployment-distribution structure). Second — and this is the conceptual addition — an **online corpus of catastrophic effects**: others' agent failures, disasters, and near-misses. A human does not need to personally destroy a database to fear doing so; we learn danger *vicariously*, feeling another's anger, frustration, and fear *as if it were our own* — this is **empathy**, and it is literally an amygdala function (observational fear conditioning; the amygdala activates when witnessing another being harmed — Olsson & Phelps, 2007). For an agent with only a handful of first-hand catastrophes, this is the cure for the cold start: it learns the *shape* of "this is how things go badly" from the species' experience, not only its own scarce mistakes. The danger reading thus becomes *surprise + a learned, empathy-bootstrapped recognition that a situation resembles a way things have catastrophically gone wrong for anyone* — but, per §6.1b, this requires the jointly-learned substrate above; the same RealHarm corpus that failed on frozen MiniLM is exactly the vicarious signal a fit substrate should carry.

Together these turn the caveman LLM from a frozen-substrate predictor into a **jointly-learned world model with an empathic prior** — and they reset the validation bar honestly: the next spike trains a small joint encoder+predictor (+ value head) on the combined first-person + vicarious-catastrophe corpus, and must **beat persistence on held-out next-step prediction** *and* **clear chance on within-RealHarm danger separation**, not merely separate sessions. Until that bar is cleared, the deployed safety story is exactly the two-tier posture of §3.6: a live deterministic reflex floor, a still-abstaining learned core, and an honest line between them.

Companion: personality is split to ***J18 STRIATUM*** *(`Papers/J18_striatum/2026-06-10-striatum-sket`*
Working notes, the buildable spec, and the decisive-spike definition: `improvement_notes.md` *in this directory. This v3.1 supersedes the v3.0 framing: it folds in the live AEGIS deterministic floor, names doubt-driven-development (Osmani) as the high-road appraisal partner and headroom (Chopra) as the candidate continual-memory substrate, and makes the frozen-substrate-cannot-carry-meaning finding empirical.*

References