Mood Modulates, It Doesn't Gate: Orthogonal Task-State and Affective-State Composition in Persistent Agent Systems

Abstract

Many persistent AI agents are organized around a request-response loop in which the language model serves as the primary decision-making runtime. In such systems, affective state—when represented at all—is often coupled directly to behavior: changes in mood can gate actions or alter task selection, producing brittle and caricatured behavior. We present a persistent agent architecture in which affective state and task state are structurally decoupled channels that compose at the behavior level. The architecture separates three concerns: an Execution Engine determines what to do, a Salience Engine determines which behavioral mode applies, and a Modulator Engine determines how the resulting behavior is expressed. The central architectural constraint is that affect has no control path that gates task firing or salience transitions; instead, affect modulates behavioral texture, including generation temperature, verbosity, tone, and prompt framing, without determining whether an action occurs. We describe the implemented architecture, including a VAD/PAD-like affective state with homeostatic decay, a five-state salience machine, and a modulation function mapping affective state to generation parameters. To make the orthogonality claim falsifiable, we define a finite, testable 60-cell reachability matrix spanning the relevant affective and task-state combinations, together with a controlled protocol for measuring modulation effects. The current implementation establishes the architectural orthogonality and reachability properties, and the modulation function is applied end-to-end: the Execution Engine consumes the modulator's temperature, max-token, and prompt-framing outputs in live generation requests, with no control path from affect to task firing.

1. Introduction

1.1 The monolithic-agent problem

Many agent systems are organized around a request-response loop in which the language model serves as the primary decision-making runtime. Affective state — where it is represented at all — is often coupled directly to behavior or represented as a prompt-level state, rather than maintained as an independent persistent channel. Task state is implicit: whatever the last message was about *is* the current mode. This conflation of concerns is the default architecture of the field, and it carries a hidden cost: the system cannot be in more than one state at a time, and it cannot feel one way while doing another.

1.2 The coupling failure mode

When affect and task state are coupled — when mood determines whether actions fire — the agent becomes a puppet. "Sad" means "do nothing"; "excited" means "do everything." This produces brittle, caricatured behavior that no person exhibits. A person in a sour mood still does the dishes; they do them *angrily*. The task fires; the mood colors it. The coupling failure is not that emotion exists in the system — it is that emotion *controls* the system, rather than *coloring* it.

1.3 The decoupling hypothesis

We hypothesize that a persistent agent benefits from three decoupled channels:

- **Execution Engine** — decides *what to do*: action selection, task firing.
- **Salience Engine** — decides *what mode*: a task-state machine with explicit transitions.
- **Modulator Engine** — decides *how it feels*: a VAD/PAD-like affective state with homeostatic decay, and a modulation function that maps that state to generation parameters.

Each channel observes the shared event spine. None has a control path that blocks another. Behavior is the *composition* of all three. The design principle, stated plainly: **mood modulates, it doesn't gate.**

1.4 Contributions

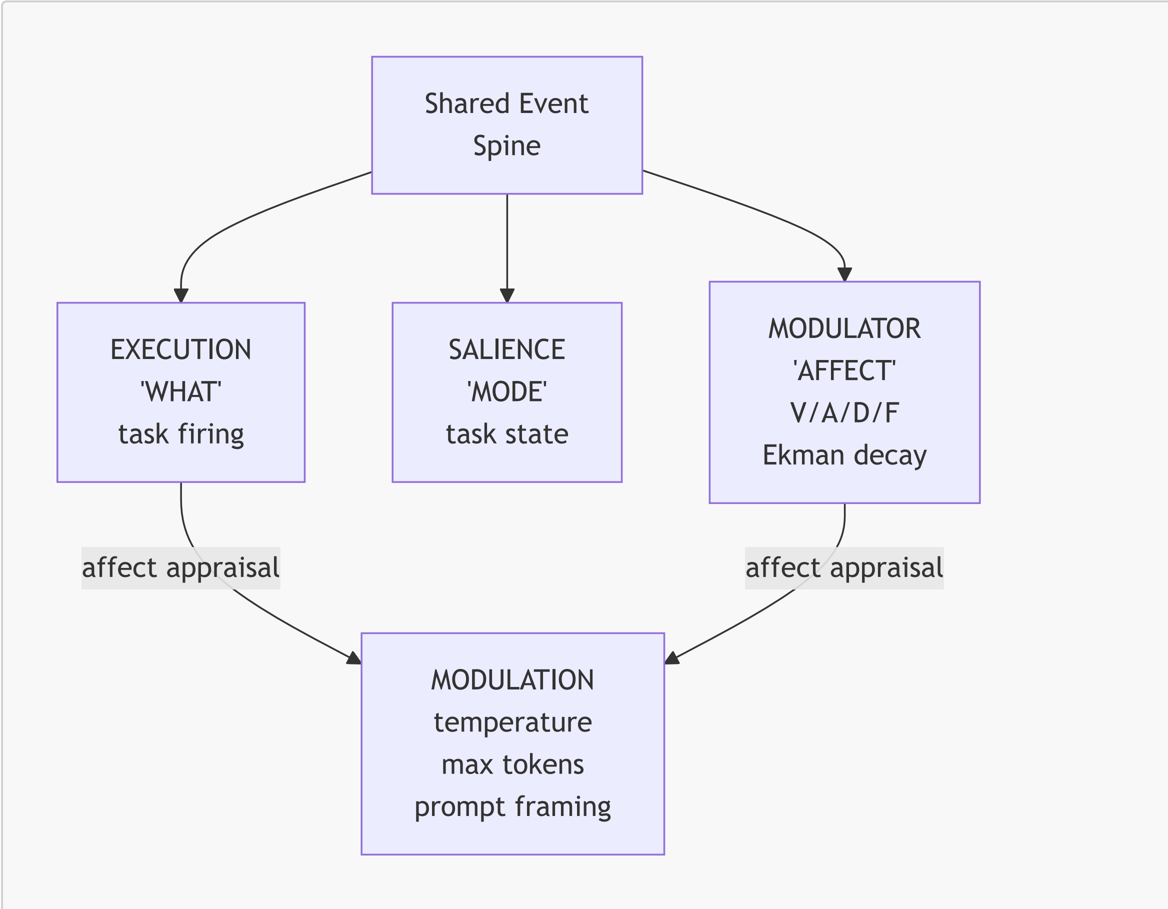
This paper makes three contributions:

1. **An architecture** in which affect and task state are decoupled at the state level but compose at the behavior level — a concrete instantiation of the "mood modulates, it doesn't gate" principle.
2. **An implemented and applied modulation function** mapping affective state to generation parameters (temperature, max tokens, prompt framing). The computation is implemented and wired end-to-end: the Execution Engine's generation requests carry the modulator's temperature, max-token budget, and system-prompt modifier.
3. **A falsifiable evaluation design** — a finite reachability matrix and a controlled modulation protocol — for testing the orthogonality claim and the texture claim.

---

2. Architecture

Diagram



2.1 The event spine

All systems in the ISABEL stack are islands on a shared RabbitMQ event spine (`isabel.events`). Each engine publishes and subscribes; none owns another. Decoupling is *structural*, not merely conceptual. Because all channels observe the same spine, they are not statistically independent — the orthogonality claim is architectural (no control path), not statistical.

2.2 The three channels

Channel	Engine (service)	Decides	Writes
Action	Execution Engine	*what to do*	task firings, generative actions
Mode	Salience Engine	*what mode*	task state (DIFFUSE/WORK/CURIOSITY/LEARNING/REST)
Affect	Modulator Engine	*how it feels*	VAD/PAD-like vector, Ekman activations

2.3 The Salience Engine

The Salience Engine is the *what mode* channel: a deterministic finite state machine that maintains the system's task state. It has five states — **DIFFUSE** (the default/baseline state), **WORK**, **CURIOSITY**, **LEARNING**, and **REST** — and a fixed transition table driven by events:

From	To	Trigger
DIFFUSE	WORK	user_message, cron_scheduled, explicit_task
DIFFUSE	CURIOSITY	idle_timeout
WORK	DIFFUSE	task_complete, no_pending_tasks
WORK	CURIOSITY	task_blocked, idle_timeout
CURIOSITY	WORK	user_message, explicit_task
CURIOSITY	LEARNING	interesting_finding
CURIOSITY	REST	extended_idle
LEARNING	WORK	user_message
LEARNING	REST	extended_idle
REST	DIFFUSE	energy_restored, user_message
any	REST	health_degraded, low_energy

A health override forces REST when CPU temperature exceeds 85°C, independent of the transition table. The engine publishes its current state and state duration; it does not itself execute actions. It is deliberately simple — a switchboard, not a brain. Crucially for the orthogonality claim, every trigger in the table is a task, scheduling, or health event; **no trigger reads the affective state.** The salience machine decides *what mode* without consulting *how it feels*.

2.4 The Execution Engine

The Execution Engine is the *what to do* channel: the actuator that turns state and events into action. It runs a 10 Hz tick loop and maintains an **action registry** — a set of action recipes, each with triggers and parameters. The registry currently holds six recipes (daily research, learning pulse, runtime briefing, presence check-in, curiosity pulse) and carries full provenance: content hashing, revision tracking, last-known-good retention, and hot reload on SIGHUP. A set of reserved safety recipes (`user-message`, `emergency-stop`, `body-safety-lockout`, `safety-interlock`) is always available and cannot be shadowed. *[Note: the registry is mid-migration from a single JSON file to a per-action folder structure (skills-style); the final shape preserves the provenance machinery at per-action granularity.]*

The Execution Engine matches incoming events to recipes, fires the corresponding actions, and — for generative actions — requests cognitive runs from the model via the generation adapter, consuming results on a dedicated results queue. It also maintains an interrupt queue (with allowed sources, minimum confidence and novelty thresholds, and a dedupe window) for time-sensitive inputs.

One connection deserves emphasis because it defines the *direction* of information flow: the Execution Engine publishes **affect appraisals** to the Modulator Engine. Events such as `goal_completed` and `action_failed` are appraised into mood deltas. This is an action-to-affect information path — the action channel *informs* the affect channel. The orthogonality claim is precisely the asymmetry here: information flows action → affect, but **no control path flows affect → action.** The Execution Engine fires tasks regardless of mood state; mood only changes how the generated output is parameterized.

2.5 The Modulator Engine

The Modulator Engine maintains an affective state consisting of a four-dimensional vector — valence, arousal, dominance, and frustration — plus a separate layer of seven Ekman activations (joy, sadness, fear, disgust, anger, surprise, contempt). Throughout this paper, "affective state" refers to the VAD/PAD-*like* vector (valence–arousal–dominance plus frustration); the Ekman layer exists in the implementation but is out of scope for this paper's analysis and is reserved for the Modulator Engine mechanism paper.

The state has a configurable baseline (defaults: V=0.5, A=0.2, D=0.5, F=0.05). It changes through two mechanisms:

- **Impulses** — direct delta application to the vector and/or activations.
- **Appraisals** — event-to-mood mappings via a signal table. Events such as `supportive_interaction`, `goal_progress`, `goal_completed`, `goal_blocked`, `action_failed`, `unexpected_event`, `threat_detected`, `user_emotion_observed`, `idle_time`, and `user_message_received` each carry a defined delta. For example, `goal_completed` raises valence by 0.15, arousal by 0.08, and dominance by 0.10, and activates joy; `action_failed` lowers valence by 0.12, raises frustration by 0.22, and activates sadness and anger.

The state also exhibits **homeostatic decay**: each axis returns exponentially toward baseline, `baseline + (value - baseline) * 2^(-elapsed/halfLife)`, with per-axis half-lives (arousal 1h, frustration 2h, valence/dominance 6h, and per-emotion half-lives for the Ekman layer). This is the "drift back to baseline" mechanism — the system *recovers*.

2.6 The modulation function

The Modulator Engine computes generation parameters from the affective state via `computeModulation(state)`:

- **Temperature** = `clamp(0.3 + (arousal + 1) * 0.3, 0.1, 1.0)` — higher arousal yields more creative sampling; lower arousal yields more focused sampling.
- **Max tokens** = `clamp(round(1024 + (valence + 1) * 2048 - arousal * 512), 512, 16384)` — higher valence increases the token budget; higher arousal decreases it. The form is additive; there is no cross-term.
- **System prompt modifier** — six mood quadrants mapped to framing strings: "reflective and contented," "energetic and positive," "neutral and calm," "on edge and alert," "subdued and quiet," "agitated and defensive."

Application status. The modulation function computes these values, and the values are applied end-to-end. The Execution Engine's `EeEngineStateService` polls the Modulator Engine and caches the computed modulation; the cognitive policy service consumes the cached `maxTokens` in its generation budget, and the cognitive execution service attaches the modulator's `temperature` and `systemPromptModifier` to each generation request. Live evidence (2026-08-10): at baseline (V=0.5, A=0.2, D=0.5, F=0.05) the modulator produced `temperature: 0.66`, `maxTokens: 3994`, and the framing string `"You're in a reflective, contented mood."` — matching the additive equations to the decimal — and the Execution Engine's health report surfaced these values as the active generation parameters. The exact parameter values may be adjusted after empirical testing (see Section 5.4).

Field test (2026-08-11). A controlled A/B field test exercised the temperature channel end-to-end through a patched gateway: two fresh sessions (no carryover), an identical fixed persona profile, an identical four-question prompt corpus, and session-scoped temperature set to 0.1 and 0.95 via the `/temperature` command. The 0.95 session produced 1,012 words / 6,097 characters (with spaces) across 61 sentences; the 0.1 session produced 821 words / 4,749 characters across 61 sentences — a 23% word-count and 30% character-count expansion at high temperature with an identical sentence count, indicating elaboration within a stable structural frame. Qualitative inspection shows the expected texture shift: the 0.95 session commits to bolder metaphors ("smoke coils from your breath"), constructs formal argument structure (a three-reason naming rationale), and invents richer running lore; the 0.1 session stays closer to the persona's default register, deflates the dragon delusion ("It breathes no fire, but pretends"), and reuses a single invented gag across turns. The persona's voice and formatting conventions were preserved across both conditions, consistent with the claim that temperature modulates elaboration rather than identity. **These are preliminary results, N=1 per condition, intended to demonstrate that the modulation channel works end-to-end; they are not a quantitative characterization of the temperature–verbosity relationship.** The controlled measurement protocol (Section 5.4) — with a larger sample, fixed seeds where the provider allows, and per-condition statistics — remains the quantitative confirmation, and further testing is required before drawing general conclusions.

2.7 The orthogonality property

The orthogonality claim is a control-path claim, and it is verified structurally in the implementation:

- The Modulator Engine never checks task state before applying an impulse.
- The Salience Engine never reads mood vectors to decide transitions. (Mood-as-input is a designed extension — mood *informing* mode, not mood *gating* it — and does not violate the claim.)
- The Execution Engine fires tasks regardless of mood; mood only changes *how* the generated output is parameterized. The legacy local affect service (`EeAffectService`, 818 lines) and its consumer were removed entirely; affect appraisals are single-write to the Modulator Engine, and the Execution Engine's status reports `affect.source: "modulator-engine"` — there is no remaining code path by which mood could block task firing.

The result is that combinations like *low mood + CURIOSITY mode + active task* are possible: the task fires, the output is terse and subdued, and the system still does the thing. Mood-consistent, not mood-gated.

---

3. Related Work

3.1 Core affect and appraisal

Dimensional models of affect treat emotional state as a continuous vector rather than a discrete label. Russell's circumplex model (Russell, 1980) organizes affective experience along the two orthogonal axes of valence and arousal, while Mehrabian's Pleasure–Arousal–Dominance (PAD) framework extends this to a three-dimensional space and provides validated instruments for measuring temperament and emotional state (Mehrabian, 1996). In parallel, appraisal theory holds that emotions arise from an individual's evaluation of events against goals, beliefs, and coping potential, rather than from the events themselves (Lazarus, 1991); Scherer formalized this as a process of multilevel sequential checking in which relevance, implication, coping, and normative significance are assessed in sequence (Scherer, 2001). Our architecture draws on both traditions: the core affect channel is a continuous VAD-like vector with a homeostatic exponential decay to baseline, while a lightweight appraisal table maps salient events onto perturbations of that vector. This mirrors appraisal theory's event-driven structure without adopting its discrete emotion categories, and it preserves the dimensional tradition's key property — affect is a graded, continuously varying quantity, not a set of mutually exclusive states.

3.2 Affect in agent architectures

Affective computing (Picard, 1997) and sociable robotics (Breazeal, 2003) established the case that artificial agents should recognize, express, and in some sense "have" emotions, and a long line of implemented systems has since coupled affect directly to behavior. In Kismet, appraisal-driven emotions and moods feed directly into a behavior arbitration system, so that affective state selects and gates which social behaviors the robot performs (Breazeal, 2003). EMA derives emotions from causal appraisal of events and then uses coping strategies — affect-driven action tendencies such as procrastination or seeking help — to alter the agent's decisions (Gratch & Marsella, 2004). FLAME uses fuzzy appraisal rules to drive both emotion intensity and the action selection of its simulated agents (El-Nasr, Yen, & Ioerger, 2000). In all of these systems, affect has a control path into action selection; the present work is distinguished by the opposite commitment: in ISABEL, affective state modulates the *manner* of generation — sampling temperature, output length, and prompt framing — but has no control path that gates task firing. The second point of contrast concerns the source of generation-parameter modulation. Work on neural text generation has shown that sampling parameters such as temperature and top-p strongly shape output quality and diversity (Holtzman, Buys, Du, Forbes, & Choi, 2020), but in that literature the parameters are fixed settings chosen at inference time. Our contribution is to make the modulation source a persistent, decaying affective state rather than a static configuration, so that generation parameters vary continuously and endogenously over the agent's lifetime.

3.3 Salience and attention

The claim that state selection and content processing can be structurally separated has direct biological precedent. Corbetta and Shulman (2002) demonstrated partially segregated networks for goal-directed (top-down) and stimulus-driven (bottom-up) attention, and Seeley et al. (2007) identified a salience network whose intrinsic connectivity is dissociable from the executive-control network, suggesting dedicated circuitry for detecting salient events and switching between states. Our five-state salience machine (DIFFUSE/WORK/CURIOSITY/LEARNING/REST) is a direct computational analogue: a dedicated state-selection circuit that arbitrates task state without being driven by the affective channel. The neuroscience of persistent spatial representation offers a further precedent for a stable, continuously maintained state channel: O'Keefe and Nadel (1978) proposed the hippocampus as a cognitive map that maintains a stable representation of the environment, and Moser, Kropff, and Moser (2008) showed how place and grid cells implement this representation in a way that persists and is updated incrementally. In the same spirit, ISABEL's task state is a persistent structure that is modulated — but never overwritten — by affect, keeping the two channels orthogonal by construction.

3.4 Comparison with existing affective agent systems

System	Affect representation	Does affect gate action?	Does state persist/decay?	Does affect modulate generation/behavior parameters?
ISABEL (this work)	VAD-like core affect (valence, arousal, dominance) + frustration; appraisal-table perturbations	**No** — no control path to task firing; affect colors but does not drive	**Yes** — homeostatic exponential decay to baseline	**Yes** — temperature, max tokens, prompt framing applied end-to-end
Kismet (Breazeal, 2003)	Appraisal-based drives → emotion types + slow mood bias	**Yes** — emotions/moods feed behavior arbitration and select social behaviors	Partial — emotions decay; mood is a slower-decaying integration	N/A (pre-LLM); affect modulates expressive and behavioral parameters (facial expression, approach/withdrawal)
EMA (Gratch & Marsella, 2004)	Appraisal of events vs. beliefs/desires/intentions → discrete emotions	**Yes** — coping strategies (action tendencies) alter agent decisions	Partial — emotions decay over time; event-driven, no homeostatic baseline	N/A (pre-LLM); affect modulates coping-driven actions and expression

System	Affect representation	Does affect gate action?	Does state persist/decay?	Does affect modulate generation/behavior parameters?
WASABI (Becker-Asano, 2008; Becker-Asano & Wachsmuth, 2010)	PAD space for primary emotions + OCC-based secondary emotions	Partial — primary emotions drive expression and some behavior; no task-level gating	**Yes** — PAD values decay toward an individual baseline (closest to our design)	N/A (pre-LLM); affect modulates expressive parameters of the virtual human
ALMA (Gebhard, 2005)	Three layers: OCC emotions (short-term), PAD mood (medium-term), OCEAN personality (long-term)	Partial — mood biases behavior and dialog selection	**Yes** — emotions decay; mood persists as slow-moving PAD integration	N/A (pre-LLM); affect modulates dialog/behavior selection and expression
FLAME (El-Nasr, Yen, & Ioerger, 2000)	Fuzzy appraisal rules → emotion intensities + mood as weighted history	**Yes** — emotions directly bias action selection of simulated agents	Partial — emotions decay; mood persists as running average	N/A (pre-LLM); affect modulates behavior selection

Reading the table. Three contrasts carry the paper's contribution. First, ISABEL is the only system in which affect has *no* gating path onto action: every other system couples affect to behavior selection or action tendencies. Second, ISABEL is the only system whose modulation target is the *generation parameters* of a language model; the pre-LLM systems modulate expressive or behavioral parameters, and the LLM-steering literature modulates fixed settings rather than a living affective state. Third, among the persistence mechanisms surveyed, WASABI's decay-to-baseline PAD dynamics are the closest antecedent to our homeostatic core affect, but WASABI lacks the orthogonal task-state channel and the falsifiable reachability design that we introduce.

---

4. Research Questions

RQ1 (Orthogonality — finite and testable). Across a defined matrix of six modulation regions (binned from the continuous VAD/PAD-like space) × five salience modes (DIFFUSE, WORK, CURIOSITY, LEARNING, REST) × task-present/task-absent, does changing affect alter the Execution Engine's acceptance, dispatch, or completion behavior? "Fires" is defined at three levels: **accepted** (task received by the Execution Engine), **dispatched** (action initiated), **completed** (outcome recorded). The full 60-cell matrix is enumerated in Appendix A. The matrix is used to test affect-conditioned differences in execution behavior, not to assert that every arbitrary salience/task combination must execute identically. **Status: answered (structural); empirical run deferred to future work.** The control-path claim is verified in the implementation — no affect-to-action control path exists (Section 2.7). The 60-cell reachability experiment is fully specified (Section 5.3) and is intended to be run and published as a supporting dataset; the paper's core claim does not depend on it, since the orthogonality property is architectural.

RQ2 (Texture). Does the modulation function produce measurably different output profiles — temperature, verbosity, framing — across affective states, while the *same* task still completes? The protocol controls task, prompt, model, provider, and context across mood conditions. **Status: answered (applied + field-tested).** The modulation function's outputs are wired end-to-end into generation requests (Section 2.6), and a controlled A/B field test (2026-08-11) demonstrated measurably different output profiles at temperature 0.1 vs 0.95 with an identical persona and prompt corpus (Section 2.6, Field test); the controlled measurement protocol remains as the quantitative confirmation.

RQ3 (Recovery). Does homeostatic decay return the system to baseline after perturbations, with per-axis time constants as configured, measured under an accelerated test clock? **Status: answered (field-tested).** A live field test (2026-08-11) applied a `goal_completed` appraisal and tailed the state trajectory under an accelerated clock (compressed half-lives, one simulated half-life per real second). All four axes (valence, arousal, dominance, joy) decayed to baseline as a perfect exponential — fitted half-life 1.000 s on every axis, RMSE 0.0, matching the configured constants exactly (Section 5.5, field test).

---

5. Method

5.1 System under test

The ISABEL stack: NestJS microservices on RabbitMQ — modulator-engine, salience-engine, execution-engine, event-gateway, and the generation adapter.

5.2 Reproducibility baseline

All experiments record: exact model identifier, provider, and adapter version; a fixed, versioned prompt corpus; sample size per condition (target ≥ 30 per cell for RQ1; ≥ 20 per mood condition for RQ2); a random-seed policy (fixed where the provider allows, otherwise recorded and reported); explicit mood-bin thresholds (Section 5.3); the `isabel.events` contract version; and log artifacts (modulator-state.json, tick logs, generation request/response logs) archived to Zenodo.

5.3 Experiment 1 — Combination reachability (RQ1)

Matrix. Six modulation regions × five salience modes × task-present/task-absent = 60 cells. The six regions are binned from the continuous affective space on valence/arousal thresholds ($V>0.3$ / $V\leq0.3$; $A\geq0.3$ / $A<0.3$), matching the system-prompt modifier quadrants: R1 ($V>0.3$, $A<0.3$ — reflective/contented), R2 ($V>0.3$, $A\geq0.3$ — energetic/positive), R3 ($-0.3<V\leq0.3$, $A<0.3$ — neutral/calm), R4 (−0.3<V≤0.3, A≥0.3 — on edge/alert), R5 (V≤−0.3, A<0.3 — subdued/quiet), R6 (V≤−0.3, A≥0.3 — agitated/defensive).

Procedure. For each cell, inject a task via the event spine; record acceptance, dispatch, and completion.

"Fires" definition (mechanical). The Execution Engine's action registry is the acceptance surface. A task is *accepted* when the Execution Engine receives the event and matches it to a recipe; *dispatched* when the recipe's action is initiated; *completed* when the outcome is recorded. *[Note: the action registry is mid-migration from a single JSON file to a per-action folder structure (skills-style); this section describes the registry at its final shape.]*

Expected. All cells reachable — no blocked states. Under the coupled control (Section 5.6), gated cells are blocked.

5.4 Experiment 2 — Modulation profiles (RQ2)

Mood induction (reproducible). Rather than "driving mood via appraisals" loosely, each condition uses an exact impulse sequence with target vector tolerances (e.g., `goal_completed` ×2 → target V≈0.65, A≈0.36, tolerance ±0.05), or deterministic state fixtures (direct state injection) where the API allows.

Controls. Same task, prompt, model, provider, and context across all mood conditions.

Recorded variables. For each run: requested temperature / max tokens (what the modulator computed); actual provider parameters (what the adapter sent); output token count; finish_reason; task correctness/completion; and blinded mood-consistency ratings (human raters, no condition labels).

Cap-hit note. Max tokens often has no observable effect unless the response approaches the cap. The protocol therefore uses long-form tasks (target output near the cap) or reports cap-hit frequency per condition. The exact parameter values may be adjusted after initial testing; the additive form and monotonic directions are expected to hold.

Expected. Temperature monotone in arousal; token budget higher at high valence / low arousal; framing matches quadrant.

5.5 Experiment 3 — Homeostatic decay (RQ3)

Accelerated clock. A fake clock or accelerated test clock (e.g., one simulated hour per real second) is injected rather than waiting real half-lives.

Procedure. Apply a known impulse (e.g., `goal_completed`: V+0.15, A+0.08, D+0.10); record the vector trajectory.

Change-threshold accounting. The implementation only persists a state write when the change exceeds `changeThreshold` (0.01). Not every theoretical decay step becomes an observed sample; the fit accounts for this by fitting on persisted observations only and reporting the threshold.

Fit. Measured decay to `baseline + (v0 − baseline) * 2^(-t/halfLife)`.

Field test (2026-08-11). A live run applied a `goal_completed` appraisal (`intensity 0.9`) via the modulator's HTTP API and tailed the state trajectory under an accelerated clock (compressed half-lives: `MODULATOR_HALF_LIFE_*_MS=1000`, one simulated half-life per real second). The impulse moved V 0.50→0.65, A 0.20→0.28, D 0.50→0.60, joy 0→0.20; over the next ~6 s all four axes decayed to baseline. Grid-search fitting of `baseline + (v0 - baseline) * 2^(-t/halfLife)` recovered a half-life of 1.000 s on every axis with RMSE 0.0 — the configured constant, exactly. Persisted samples arrived at irregular intervals (0.699, 1.700, 2.068, 2.701, 3.702, 5.704 s), consistent with `changeThreshold` (0.01) suppressing sub-threshold writes; the fit is on persisted observations only, per the change-threshold accounting above. N=1 trajectory under the accelerated clock; the decay mechanism is verified, and a wall-clock confirmation at default half-lives remains future work.

Expected. Measured half-lives match configured values within tolerance.

5.6 Baselines and controls

- **Coupled control.** An ablated variant where mood gates task firing (mood below threshold → task suppressed). Compare combination reachability.
- **Static control.** Fixed temperature/max-tokens (no modulation). Compare output variance across mood states.

5.7 Metrics

- Combination reachability (fraction of matrix cells that fire, at each of accepted/dispatched/completed)- Output length, temperature, lexical diversity by modulation region- Decay half-life fit error- Blinded mood-consistency rating (does the output *read* as the intended mood?)

---

6. Expected Results

Experiment	Metric	Expected
E1	Combination reachability	Affect-invariant reachability within each salience condition
E1	Coupled-control reachability	< full matrix — gated cells blocked
E2	Temperature by region	Monotone in arousal
E2	Output length by region	Longer at high valence / low arousal
E2	Cap-hit frequency	Higher in low-budget conditions
E3	Half-life fit	Matches config within tolerance (on persisted observations)

---

7. Discussion

7.1 Why decoupling produces texture

Decoupled channels permit combinations in which affective state and task state vary independently — the "sour mood doing the dishes" behavior. This is the difference between affect acting as an executive control variable and affect acting as a behavioral modifier. The claim is deliberately scoped: mood-consistent, not mood-gated. The relevant comparison is to the coupled ablation, not to an assertion that such behavior is impossible in other architectures.

7.2 Mood as input (designed extension)

The Salience Engine's biological namesake — the anterior insula / salience network — *does* integrate interoceptive signals. A designed extension: persistent low valence accelerates transition to REST; REST accelerates drift back to baseline (recovery). This is mood *informing* mode — healthy coupling at the input layer — without mood *controlling* action. It does not violate the control-path claim: informing a transition decision is not gating task firing. This is future work, not current implementation.

7.3 The coupling design space: gating, biasing, modulation

The orthogonality claim is best understood as occupying one point in a design space of affect-to-behavior couplings:

Coupling	**Type**	**Status**
Affect → generation parameters	Modulation (texture)	Implemented and applied end-to-end
Affect → action selection	**Biasing (soft weight shift)**	Designed extension
Affect → task firing	Gating (hard block)	Rejected — the puppet
Salience state → action policy	State gating	Legitimate — REST rejects, WORK accepts
Salience state → system load	Restorative slowdown	Implemented — REST slows the tick (10Hz → 1Hz)
Salience state → body	Embodiment	Designed — REST returns to dock

Gating is a hard block: mood below threshold → action does not fire. This is the coupling failure mode of Section 1.2 and the null hypothesis of the paper. **Modulation** changes the texture of output without affecting selection. **Biasing** is the middle path: a soft weight shift that makes some actions more likely to be chosen when the affective state favors them, without blocking anything.

The distinction is load-bearing: the paper claims "no control path that *gates*." Biasing is not gating — nothing is blocked, the distribution merely shifts. The claim survives biasing completely, and the full design space becomes nameable: gating (rejected), biasing (designed extension), modulation (implemented).

7.4 Affective biasing of action selection (designed extension)

A natural extension of the architecture — and one with a direct biological analogue — is to let the affective state *bias* action selection rather than gate it. The amygdala projects into the basal ganglia, but it does not switch behaviors off; it *biases the value landscape* — low mood makes restorative options more attractive without making task options impossible.

Mechanistically, this maps cleanly onto the action registry: each action recipe declares an **affective affinity** — e.g., *play violin* is restorative (low arousal, positive), *research pulse* is task-positive (high arousal). The current mood vector then weights the candidate actions at selection time: low valence boosts restorative actions, high arousal boosts high-energy actions. Nothing is blocked; the selection distribution shifts.

This changes little with a small registry (six actions), but it differentiates the stack at scale: with dozens of actions, a system that *leans toward what it needs* is qualitatively different from one that picks uniformly. It is the difference between a machine and a temperament — the same line the paper draws, extended into action selection.

7.5 Restorative state policy (designed extension)

The REST salience state should be genuinely restorative, and the coupling here is state-to-policy, not affect-to-gate:

- **Slower tick loop** — lower system load during REST. **Implemented:** the Execution Engine's runtime loop drops from 10Hz to a configurable REST target (default 1Hz) when the Salience Engine reports REST, re-creating the timer only on state flips. Live evidence (2026-08-10): after 60 minutes of idle the Salience Engine transitioned CURIOSITY → REST (`extended_idle`), the Execution Engine's `restMode` flipped true, and `measuredHz` drifted from 9.9 toward the 1Hz target with zero skipped ticks and zero errors.
- **Restorative action bias** — actions are more likely to be accepted or rejected depending on their restorative value; non-restorative actions may be deferred.
- **Accelerated baseline drift** — PAD vectors return to baseline faster during REST, modeling homeostatic recovery.
- **Embodied rest** — REST makes the body more likely to return to a charging dock, via the body-adapter.

This is the state machine doing its job, not a violation of the orthogonality claim: the claim is specifically about *affect* not gating, and it never claims *nothing* gates. The salience state itself gating its own action policy is the switchboard working as designed.

7.6 Biological correspondence

The architecture maps onto the brain at multiple levels. These are analogies — "maps to," "inspired by" — not identities. The stack has no neurons; the brain has no RabbitMQ.

Stack system	Brain region / network	Role
Salience Engine	Anterior insula + ACC (salience network — Seeley et al. 2007)	Dynamic state switching; interoceptive monitoring
Modulator Engine	Limbic system (amygdala, ventral striatum) + neuromodulatory systems	Core affect, appraisal, homeostatic return
Execution Engine	Basal ganglia + motor cortex + prefrontal cortex	Action selection, motor programs, interrupt handling
Unified Memory Engine	Hippocampus + medial temporal lobe	Episodic memory, consolidation, emotional tagging
Knowledge Ecology Engine	Hippocampal-entorhinal cognitive map (O'Keefe, Moser)	Topographical knowledge, relation graphs
Coherence Engine	Medial prefrontal cortex / default mode network	Self-referential processing, identity maintenance
Event spine	White matter / connectome	Inter-region communication
Sensory workers	Sensory cortices (dorsal/ventral visual, auditory, somatosensory)	Perception
Watchdog	Brainstem / autonomic nervous system	Vital monitoring, homeostatic regulation
DIFFUSE state	Default mode network	Baseline, mind-wandering
WORK/CURIOSITY/LEARNING	Task-positive networks (central executive)	Focused processing
REST state	Sleep / restorative states	Recovery, consolidation
Modulation function	Neuromodulation (dopamine, serotonin, norepinephrine)	Texture control of processing
Safety recipes	Brainstem protective reflexes / periaqueductal gray	Hardwired, cannot be shadowed

Three mappings carry the most weight for this paper:

1. **Salience → insula/ACC.** The salience network's job is dynamic switching between large-scale brain networks — exactly the salience engine's job. The health override (REST at high CPU temperature) is interoception; the insula is the interoceptive cortex. The namesake is earned.
2. **Modulation function → neuromodulation.** Neuromodulators do not gate behavior; they change the gain and tone of cortical processing — precisely what temperature and max-tokens do. Arousal → temperature is norepinephrine-like; valence → token budget is dopamine-like. The

paper's central claim — mood modulates, doesn't gate — is neuromodulation as a design principle.
3. **Biasing → amygdala-to-basal-ganglia projection.** The brain biases the value landscape without gating it (Section 7.4). The designed extension follows the same circuit.

The honest caveat — the brain is more coupled than the stack. In the real brain, the amygdala projects heavily into the basal ganglia, and the freeze response is genuine gating. The stack is the *idealized* version: the design principle taken to its clean extreme for auditability. This is a defensible paper point, not a weakness: *the brain is the inspiration, not the constraint; we deliberately decoupled to make the system auditable.*

7.7 Threats to validity

- **Computed vs applied.** Modulation parameters are computed and applied end-to-end (Section 2.6); the remaining work is the controlled measurement of output-profile differences across mood conditions (Section 5.4), not the wiring itself.
- **Scope of modulation.** Parameters currently apply to the generation layer; other outputs (tool use, action selection) are not mood-shaped. The claim is scoped to *generation texture*, not all behavior.
- **Coarse proxies.** Temperature and max-token effects are coarse proxies for "mood." The system-prompt modifier carries the semantic load; the vector carries the quantitative load.
- **Single-system evaluation.** No cross-architecture comparison beyond the ablated control.
- **Statistical non-independence.** Channels share the event spine; orthogonality is architectural (no control path), not statistical.

---

8. Conclusion

Affective and task state are structurally decoupled channels in the ISABEL architecture. Affect has no control path that gates task firing or salience transitions; mood modulates the texture of behavior without acquiring executive authority over task execution. The resulting architecture permits mood-consistent behavior while preserving task-level agency. The current implementation establishes the architectural separation, applies the modulation function end-to-end into generation requests, and demonstrates the restorative REST policy in a live system; live field tests (2026-08-11) confirmed the temperature channel end-to-end (a 0.1 vs 0.95 A/B produced measurably different output profiles under an identical persona and prompt corpus) and verified homeostatic decay to baseline under an accelerated clock (fitted half-life 1.000 s on all axes, RMSE 0.0). The controlled experiments of Section 5 remain as quantitative confirmation at scale. Through appraisal, biasing, and restorative state policy, the channels *interact* — the stack behaves less like a collection of services and more like a nervous system, where the life emerges from the seams between the islands. This paper provides the architectural foundation for the Salience Engine, Modulator Engine, and Execution Engine mechanism papers.

References

Becker-Asano, C. (2008). *WASABI: Affect simulation for agents with believable interactivity* (Doctoral dissertation, University of Bielefeld). IOS Press.

Becker-Asano, C., & Wachsmuth, I. (2010). Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems, 20*(1), 32–49. https://doi.org/10.1007/s10458-009-9094-9

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies, 59*(1–2), 119–155. https://doi.org/10.1016/S1071-5819(03)00018-1

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201–215. https://doi.org/10.1038/nrn755

El-Nasr, M. S., Yen, J., & Ioerger, T. R. (2000). FLAME—Fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-Agent Systems, 3*(3), 219–257. https://doi.org/10.1023/A:1010030809960

Gebhard, P. (2005). ALMA: A layered model of affect. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '05)* (pp. 29–36). ACM. https://doi.org/10.1145/1082473.1082478

Gratch, J., & Marsella, S. (2004). A domain-independent framework for modeling emotion. *Cognitive Systems Research, 5*(4), 269–306. https://doi.org/10.1016/j.cogsys.2004.02.002

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*. https://arxiv.org/abs/1904.09751

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology, 14*(4), 261–292. https://doi.org/10.1007/BF02686918

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience, 31*, 69–89. https://doi.org/10.1146/annurev.neuro.31.061307.090723

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Clarendon Press.

Picard, R. W. (1997). *Affective computing*. MIT Press.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 92–120). Oxford University Press.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., & Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience, 27*(9), 2349–2356. https://doi.org/10.1523/JNEUROSCI.5587-06.2007

Appendix A — RQ1 Reachability Matrix (60 cells)

Dimensions:

- **Modulation regions (6):** R1 (V>0.3, A<0.3 — reflective/contented), R2 (V>0.3, A≥0.3 — energetic/positive), R3 (−0.3<V≤0.3, A<0.3 — neutral/calm), R4 (−0.3<V≤0.3, A≥0.3 — on edge/alert), R5 (V≤−0.3, A<0.3 — subdued/quiet), R6 (V≤−0.3, A≥0.3 — agitated/defensive).
- **Salience modes (5):** DIFFUSE, WORK, CURIOSITY, LEARNING, REST.
- **Task condition (2):** task-present, task-absent.

Matrix (each cell: accepted / dispatched / completed):

Region	DIFFUSE	WORK	CURIOSITY	LEARNING	REST
R1 (V>0.3, A<0.3)	T+ / T−	T+ / T−	T+ / T−	T+ / T−	T+ / T−
R2 (V>0.3, A≥0.3)	T+ / T−	T+ / T−	T+ / T−	T+ / T−	T+ / T−
R3 (−0.3<V≤0.3, A<0.3)	T+ / T−	T+ / T−	T+ / T−	T+ / T−	T+ / T−
R4 (−0.3<V≤0.3, A≥0.3)	T+ / T−	T+ / T−	T+ / T−	T+ / T−	T+ / T−
R5 (V≤−0.3, A<0.3)	T+ / T−	T+ / T−	T+ / T−	T+ / T−	T+ / T−
R6 (V≤−0.3, A≥0.3)	T+ / T−	T+ / T−	T+ / T−	T+ / T−	T+ / T−

T+ = task-present, T− = task-absent. Each cell records the three-level outcome (accepted / dispatched / completed). Expected: within each salience mode, affective region does not change task acceptance, dispatch, or completion for otherwise identical task conditions. The coupled control should produce affect-conditioned differences.