A Practical Architecture for Emergent, Safety-Aligned AI (Summary)

By Keiron Scott · https://www.linkedin.com/in/keiron-scott-06196a17

AI collaborator: Claude (Anthropic)*

*Used via API for cross-model validation and drafting support. No endorsement or affiliation implied.

August 11, 2025

What this is

A cloud-ready blueprint that moves beyond stateless chats. It adds long-term memory, symbolic oversight, second-opinion checks across models, scheduled self-reflection, and a coherence dashboard — all guided by a documented ethics codex (“Seed of Harmony”).

The Five Layers (at a glance)

1) Persistent Memory — vector recall + facts + optional graph; summarize/age; contradiction checks.
2) Neuro-Symbolic Control (8 sublayers) — Thresholds, Navigation, Constraints, Catalysts, Tools, Archetypes, Integration, Core Forces.
3) Cross-Model Validation — independent answers/critique across models; reconcile by agreement/evidence; human fallback for high-stakes ties.
4) Recursive Self-Observation — log decisions/tools; scheduled reflections with strict caps.
5) Resonance Diagnostics — agreement (κ), self-consistency, embedding coherence, constraint-violation rate, latency stability.
Diagnostics for coherence — not a claim of sentience.

What’s new

Not a single trick but the end-to-end orchestration: archetype roles feeding a global integrator, explicit ethics at the core, and measurable coherence signals.

Safety by design

Constitutional self-checks, external moderation, sandboxed tools, recursion/compute caps, memory hygiene, and a human override.

How to evaluate

• A/B with vs. without cross-model validation on factual/tool-grounded tasks (accuracy, κ, latency/cost).
• Longitudinal runs with vs. without self-observation (learning curves, drift checks).
• Resonance panel: agreement, self-contradictions, embedding tightness, rule-violation rate, latency trends.

Call for collaborators / licensing

If you’re building long-horizon agents, neuro-symbolic systems, or coherence metrics, I’m keen to collaborate.
For commercial use or funded pilots, message me on LinkedIn.