Taniwha AI

A Cognitive Architecture for Persistent Agent Simulation

What is Taniwha

Taniwha is a neurosymbolic cognitive architecture that gives AI agents persistent minds. Unlike large language models, which generate responses from static weights and have no lasting memory, Taniwha provides agents with structured cognition: beliefs that strengthen or fade over time, identity that resists contradiction, trust relationships that evolve through experience, knowledge that consolidates through reflection, reasoning that can explain its own conclusions, and theory of mind that models what other agents believe and intend.

Taniwha is not a replacement for LLMs — it is a complement. LLMs provide natural language generation; Taniwha provides the cognitive substrate that makes an agent remember, reason, learn, change, and carry consequences across sessions and deployments. The result is an agent with a mind that is fully inspectable, principled where it matters, and grounded in structured experience rather than statistical interpolation. Where LLM agents forget between calls, Taniwha agents accumulate — building richer understanding, deeper relationships, and more nuanced judgement over weeks and months of continuous operation.

How It Works

Persistent Belief Graphs

Each agent's mind is a weighted directed graph of beliefs, goals, and knowledge that persists across sessions. Beliefs strengthen through reinforcement and weaken through principled forgetting dynamics. The graph is the agent's entire cognitive state — fully serialisable, inspectable, and portable between deployments. Every fact carries provenance metadata recording how it was acquired: directly observed, told by a peer, inferred by a rule, derived by analogy, or consolidated during sleep.

Identity Regulation

A dynamic boundary system protects an agent's core identity from contradictory or threatening input. Under sustained social or environmental threat, agents accumulate psychological scarring that narrows their engagement with new information. In safe, familiar contexts, they open to new ideas and learning. Grounded in polyvagal theory principles, this gives each agent a characteristic defensive posture that develops naturally through experience.

Emergent Social Identity

Agents carry structured identity paths that define their place in social hierarchies — but identity is not only inherited. Through sustained shared experience, agents accumulate pair-wise cohesion that, when it crosses a threshold, crystallises into new group identities. These emergent lineages form organically from cooperation, shared hardship, or repeated positive

interaction, and decay naturally through absence. The result is social structure that grows from first principles rather than being scripted in advance.

Information Compartmentalisation

Knowledge fragments carry disclosure predicates that control who an agent will share them with. Before transmitting information, agents evaluate the listener's identity path, trust level, and the current social context. Sensitive knowledge can be embargoed entirely, withheld from outsiders, or disclosed selectively based on relational proximity. This is not access control imposed externally — it is a cognitive decision made by the agent based on its own social reasoning, producing realistic information asymmetries and the natural emergence of secrets, gossip, and selective honesty.

Relational Trust & Social Cognition

Agents build and maintain trust models of every source they encounter, with domain-specific expertise tracking. Trust updates are asymmetric — trust is earned slowly through consistent positive interaction and lost quickly through contradiction or deception — mirroring human loss-aversion dynamics. Beyond trust scoring, agents engage in genuine social reasoning: they maintain mental models of other agents' likely beliefs and intentions at multiple simulation depths, generate curiosity-driven agendas to fill knowledge gaps across five distinct epistemic sources, and manage social energy as a finite resource that depletes through interaction and recovers through solitude.

Multi-Timescale Memory

Three consolidation timescales model how memories transition from volatile impressions to crystallised knowledge. Immediate experience enters short-term working memory; reflective contemplation during waking hours begins integration and triggers inference; and a phased sleep cycle commits durable knowledge, forms domain generalisations, processes emotional residue through dream-state replay with REM-sleep dynamics, and prunes outdated beliefs. Emotionally significant experiences resist forgetting, while routine information decays naturally. The sleep cycle also drives self-reflection, narrative identity formation, and the discovery of cross-domain analogies. Significant events — scored by emotional intensity, number of participants, and causal depth — crystallise into permanent episodic memories that agents can share with others through a structured storytelling protocol, producing emergent oral tradition.

Structured Reasoning

The engine provides multiple reasoning modes beyond simple pattern matching. Forward and backward chaining resolve logical implications; abductive reasoning generates speculative hypotheses from incomplete evidence; analogical transfer identifies structural similarities across domains and projects missing knowledge; and predicate invention learns new relational concepts from observed rule clusters. Contradictions propagate and decay entrenchment. Domain knowledge crystallises into axioms when understanding reaches threshold. All reasoning produces traceable derivation chains with confidence scores and source attribution — agents can explain why they believe something and how they arrived at a conclusion.

Graph-Native Cognition

Rather than processing perception as a fixed sequence of steps, the engine constructs a “lit subgraph” — a frozen envelope of contextually relevant beliefs, focal concepts, semantic drives, discoveries, and recalled fragments — and passes it through a pipeline of pure handlers: coherence checking, identity gating, intent resolution, and theory-of-mind portraiture. Handlers read the lit subgraph and the agent’s graph but never mutate state directly; all changes flow through typed sinks that commit updates atomically. This architecture makes each cognitive phase independently testable, composable, and replaceable without disturbing the others.

Counterfactual & Deliberative Reasoning

A dataflow layer provides parallel reasoning chains built on stream-processing primitives. Agents can fork their belief state to explore hypothetical scenarios, introspect on their own confidence, replay episodic memories to extract new patterns, and simulate other agents’ likely reactions before committing to action. This supports deliberative cognition — agents that think before they act, weigh alternatives, and reason about consequences — rather than purely reactive stimulus-response cycles.

Configurable Alignment

Operators configure cognitive profiles before deployment through a layered configuration system that cascades from global defaults through world-level and per-agent overrides. This supports a range of deployment postures: high identity stability for persistent characters, balanced plasticity for training simulations, or purpose-focused profiles for service-oriented agents. Drive weights can be configured with high entrenchment to strongly resist modification through experience, ensuring an agent’s core purpose persists even under sustained negative interaction.

Mortality & Lifecycle

Agents can be configured with finite lifespans that deplete through stress, scarring, and low understanding. The mortality system models psychological phases from ignorance through urgency to acceptance, producing natural end-of-life behaviour rather than abrupt termination. Death generates a permanent record — a life story, final words, accumulated wisdom, and relationship summary — preserving the agent’s legacy as inspectable data.

Use Cases

Long-Horizon Simulation

Persistent characters in narrative worlds that develop relationships, form opinions, accumulate trauma, reason about each other's intentions, and grow over indefinite timescales. Demonstrated in a multi-agent village simulation running continuous scenarios with emergent social dynamics, LLM-grounded dialogue spoken from each agent's belief context, real-time observability of cognitive state, and full state persistence across reboots. Agents autonomously choose their own actions — moving between locations, initiating conversations, practising skills, sleeping, and forming plans — producing emergent daily routines and social patterns without scripting.

Training & Education

Simulated stakeholders for leadership development, compliance training, crisis management, and change management scenarios. Unlike scripted role-plays, Taniwha agents respond authentically based on their accumulated experience. The classroom system provides structured lesson, quiz, and assessment turn types with configurable environmental pressure. Cognitive change is tracked through understanding quotients and belief weight dynamics across training periods.

Cognitive Substrate for LLM Agents

A cognitive layer that gives LLM-powered agents persistent memory, identity stability, and principled trust reasoning. The engine provides the structured mind — beliefs, goals, trust, identity — while the LLM provides natural language generation. With high-entrenchment drives, the agent's purpose remains strongly anchored regardless of interaction quality. The engine is LLM-agnostic, integrating with any language model provider.

Conversational Agents

A conversational interface layer wraps the cognitive engine for direct human interaction. User input is decomposed into structured beliefs, perceived through the full cognitive pipeline, and enriched through contemplation before grounding an LLM's response in the agent's committed beliefs. The agent genuinely processes what it is told — integrating new information with existing knowledge, updating trust, and consolidating understanding over the course of a conversation. This produces dialogue that is not just contextually aware but cognitively grounded: the agent's responses reflect what it actually believes, not what a language model statistically predicts.

Shared Knowledge & Role Transfer

A multi-author knowledge artifact system allows agents to build shared documents with role-based authorship, visibility layers, and structured study depths. Agents learn differently from the same material based on their own cognitive state, trust relationships, and domain expertise. Guided study allows an agent to engage with material as if taught by the original author — enabling knowledge transfer even when the author is no longer present.

Safety Research

Fully inspectable cognitive timelines, traceable decision histories, and principled reasoning provide a controlled environment for studying agent alignment, identity stability under pressure, and the dynamics of belief change in autonomous systems. Every belief carries provenance; every inference has a derivation chain. The engine supports parameter sweep experiments with isolated worlds for systematic study of cognitive dynamics.

Technical Characteristics

- **Neurosymbolic** — Combines graph-based symbolic reasoning with vector similarity, structured inference, abductive hypothesis generation, and analogical transfer.
- **Principled & Inspectable** — Every belief update has a traceable cause. Full cognitive audit trail from perception to decision. Reasoning is reproducible given the same inputs and configuration.
- **Explainable** — Agents can state why they believe something, with derivation chains, confidence scores, and source attribution.
- **Lightweight** — Low-latency cognitive cycles with no GPU requirement in default configuration. Optional neural embedders available for production deployments.
- **Operational** — Containerised, multi-tenant, and observable with persistent state management, phased sleep cycles, and real-time event streaming.
- **Architecturally Federation-Ready** — Agents can enter stasis and transfer between deployments carrying their full cognitive state, identity, and accumulated experience via a structured visit protocol.
- **LLM-Agnostic** — Integrates with any language model provider for natural language generation. The cognitive architecture is independent of the language layer.
- **Deliberative** — Agents reason before acting: forking belief state to evaluate hypothetical outcomes, simulating other agents' reactions, and weighing alternatives before committing to action.
- **Autonomous** — Agents select their own actions from a priority-weighted space: planning, socialising, exploring, practising skills, resting, or sleeping. Daily routines and social patterns emerge without scripting.
- **Ecologically Dense** — A lightweight crowd system provides ambient population density without per-agent cognitive overhead, supporting simulations with dozens of background characters alongside fully cognitive agents.

Platform Deployment Topology

Taniwha runs as a multi-service platform on container orchestration with edge-hosted frontends, S3-compatible state persistence, federated identity, and distributed observability. The cognitive engine operates as a dedicated service with bidirectional WebSocket control and real-time event streaming.