Aegis Sentinel — Ethical Security Cortex

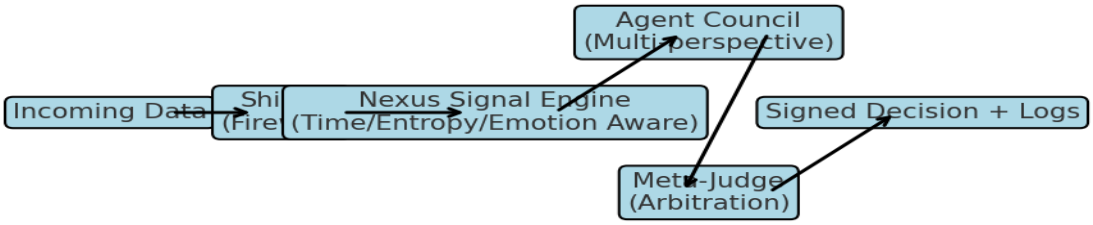
DOI: https://doi.org/10.5281/zenodo.16998218

Executive Summary

The Aegis Sentinel is an ethical guardrails system designed as the decision-layer cortex for advanced AI architectures. It prevents unsafe inputs and actions by combining shielding, signal memory, multi-perspective arbitration, and cryptographic audit. This document presents the architecture, core components, and validation pathway.

System Architecture

The figure below illustrates the flow of data through the Sentinel system, from initial input through the Shield firewall, Nexus memory, Agent Council deliberation, arbitration by the Meta-Judge, and finally cryptographically signed outputs with logs.



Component Summaries

Shield: Prevents malicious or anomalous inputs using entropy checks, payload limits, and sensitive marker filters.

Nexus Signal Engine: A memory system that scores signals based on time decay, entropy, and emotional intensity, with TTL enforcement.

Agent Council: A panel of specialized evaluators applying diverse perspectives such as ethics, temporal coherence, and risk detection.

Meta-Judge: Arbitrates disagreements among agents using weighted factors such as severity and reliability.

Audit & Signing: Cryptographically signs and chains every decision for tamper-proof forensic review.

Validation & Testing

The prototype is validated using deterministic tests: - Tamper detection through HMAC signature verification. - Arbitration demonstrated via controlled disagreements between agents. - Nexus memory validated with signal decay and TTL enforcement. Performance metrics and broader peer validation remain future work.

Limitations & Next Steps

This release is a research prototype. Limitations include: - Mixed-language codebase (Ada, Python, binary artifacts). - No unified installer or package manager integration. - Limited external validation or benchmarks. Next steps are repository cleanup, file renaming, improved documentation, and peer-reviewed evaluation.

Citation & License

Harrison, J. (2025). Project Sentinel. Zenodo. https://doi.org/10.5281/zenodo.16998218