HAIA Policy Document 6

Overwatch: The Cognitive Monitoring Shield for GOPEL

A Working Paper to Support AI Provider Plurality

Basil C. Puglisi, MPA | May 2026 | Proof of Concept v2.4

What Is Overwatch?

Overwatch is a cognitive monitoring shield built to protect GOPEL (Governance Orchestrator Policy Enforcement Layer) infrastructure. GOPEL enforces AI governance through deterministic, non-cognitive operations: hashing, signing, encrypting, verifying. These operations are immune to prompt injection because they contain no cognitive processing. But deterministic enforcement can only enforce rules defined in advance. Novel attacks, emergent behaviors, and contextual anomalies require cognitive detection. Overwatch provides that detection. It sits outside GOPEL, observes everything GOPEL processes, and identifies threats that deterministic rules cannot catch. Overwatch never writes into GOPEL, never modifies the audit chain, and never alters enforcement decisions. Even a fully compromised Overwatch cannot tamper with the governance record.

Why It Was Built

I built Overwatch in direct response to security findings from seven independent AI platform reviews of GOPEL (ChatGPT, MiniMax, Grok, Gemini, Kimi, Claude, DeepSeek). Over 60 findings across 14 versioned releases identified threats that deterministic enforcement alone could not address: prompt injection, supply chain compromise, behavioral drift, semantic manipulation, and insider threat patterns requiring cognitive analysis for detection. During that process, I identified the semantic manipulation gap, where individually clean prompts collectively steer a model toward an unauthorized outcome. That discovery became the v2.3 Trajectory Gatekeeper. In v2.4, I built the Factics Calibration loop with tamper-evident, HMAC-SHA256 signed, self-tuning thresholds that converge rather than oscillate.

What It Does

Capability	Function
Intent Analysis	Scans prompts for injection patterns, role manipulation, and social engineering before they reach the AI model.
Context Inspection	Detects embedded directives including obfuscated base64/hex payloads and trust tier violations.
Output Evaluation	Detects behavioral drift, confused deputy conditions, and unauthorized state changes in model responses.
Structural Verification	Verifies GOPEL source code integrity against signed deployment manifests from outside the trust boundary.
Escalation Management	Five severity levels (NOMINAL through HALT) with dual-mode RAI/AIG operating model.
Random Audit	Cryptographic probabilistic sampling. Any transaction may be audited. Probability ratchets up with findings.
Adaptation (Factics)	Confirmed attacks become detection rules. Confirmed false positives refine thresholds. Continuous improvement.
Trajectory Gatekeeper (v2.3)	Watches three soft signals (scope drift, privilege gradient, coherence decay) across conversation windows. Fires CBG trigger when 2+ signals converge. Catches semantic manipulation that pattern matching cannot detect.
Factics Calibration (v2.4)	Tamper-evident HMAC-SHA256 signed CalibrationState. Asymmetric threshold adjustment (10% tighten on confirmed attack, 5% loosen on confirmed false positive). Self-tuning thresholds that converge to useful sensitivity without oscillating.

Proof of Concept Status

Overwatch v2.4: 16 modules, 390 tests across 15 test suites (100% passing), 14 versioned releases, 0 cross-imports with GOPEL verified by static analysis. GOPEL v0.6.2: 11 modules, 218 tests, 11 security fixes. Total: 27 source modules, 0 shared code. Source: github.com/basilpuglisi/HAIA/tree/basilpuglisi-overwatch/overwatch