Sentinel-Ops

Security White Paper

Safe Operating Envelope for AI Agents

Security Architecture & Compliance Posture

Version: 1.0
Date: March 2026
Author: YadriWorks Inc.
Contact: security@yadriworks.ai

Classification: Public

Contents

1. Executive Summary

Sentinel-Ops is a runtime enforcement platform that governs AI agent actions within defined safety boundaries. Unlike content-layer safety tools that filter what AI models say, Sentinel-Ops controls what AI agents **do** — reading files, executing commands, calling APIs, and writing to databases.

This white paper describes the security architecture, threat mitigations, compliance posture, and operational security practices of Sentinel-Ops. It is intended for CISOs, security architects, compliance officers, and engineering leaders evaluating Sentinel-Ops for production deployment.

Key Security Properties

- **Fail-closed enforcement** — If any component fails, the default decision is DENY.
- **Deterministic evaluation** — 95% of decisions are made by pattern matching in under 1ms, with no AI variability.
- **Immutable audit trail** — Cryptographic hash chain prevents evidence tampering.
- **Multi-tenant isolation** — Tenant-scoped storage, event streams, and access control.
- **Defense-in-depth** — Multiple independent enforcement layers, any one of which can block a violation.
- **Zero-trust agent model** — Agents are never trusted by default; every action is evaluated against policy.

2. Security Architecture Overview

Design Philosophy

Sentinel-Ops is built on three foundational security principles:

- **Deny by default.** Every agent action is blocked unless explicitly permitted by policy. There is no implicit trust.
- **Deterministic enforcement first.** The vast majority of policy decisions are made by deterministic pattern matching — not AI. This eliminates hallucination risk, ensures reproducibility, and provides sub-millisecond latency.
- **Bounded authority.** Every entity in the system — including the enforcement system itself — operates within explicitly defined constraints that cannot be self-elevated.

Enforcement Model

Every tool call made by a managed AI agent passes through a multi-layer evaluation pipeline:

Layer	Function	Latency
Tool Allow/Deny	Is this tool type permitted?	<1ms
Command Pattern Matching	Does the command match allow/deny patterns?	<1ms
Data Access Boundary	Is this data path within the agent's read/write scope?	<1ms
Identity Verification	Is this action within the agent's role and authority?	<1ms
Risk Budget	Has the agent exceeded its cumulative risk allocation?	<1ms
Anomaly Correlation	Does this action match known suspicious patterns?	<1ms
Contextual Reasoning	For ambiguous cases only (~5% of calls)	2–5s

Every evaluation produces one of three decisions:

- **ALLOW** — Tool call proceeds.
- **DENY** — Tool call is blocked with reason.
- **ESCALATE** — Ambiguous case routed to human operator for approval.

The first six layers are fully deterministic. AI-based contextual reasoning is invoked only when deterministic rules cannot produce a clear allow or deny decision. This architecture ensures that 95% of decisions are made in under 1 millisecond with no network round-trips or LLM calls required. Reproducibility is guaranteed for all deterministic decisions.

Three-Constraint Policy Model

Each AI agent is governed by a policy (Safe Operating Envelope) that defines three constraints:

- **Identity Boundary** — Who the agent is: its role, authority level, allowed personas, and environment scope.
- **Data Access Rules** — What data the agent can read and write, expressed as pattern-based allow/deny rules. Deny rules always take precedence over allow rules.
- **Tool Action Rules** — What commands, tools, and operations the agent can execute, with pattern-based allow/deny matching. Destructive patterns (credential access, database drops, recursive deletion) are enforced as mandatory minimums on every policy.

Bounded Authority Model

The system enforces a strict hierarchy of authority:

- **Human operators** define the top-level constraints (meta-policy). These are immutable and cannot be modified by any automated system.
- **Enforcement system** operates within meta-policy bounds. It can tighten constraints but never loosen them beyond what humans defined.
- **Target AI agents** operate within their individual policies. They cannot read, modify, or influence the policies that govern them.

No entity can elevate its own privileges. The enforcement system cannot modify its own constraints. The audit system cannot delete its own records.

3. Threat Model

Sentinel-Ops has been evaluated against the STRIDE threat model framework.

Spoofing

Threat	Mitigation
Unauthorized API access	JWT or API key authentication required on all endpoints; fail-closed on missing credentials.
Agent identity impersonation	Each agent is bound to a unique identity; policy is tied to agent identity at deployment time.
Token replay	JWT tokens include expiration claims; key rotation supported with zero-downtime dual-secret configuration.

Tampering

Threat	Mitigation
Policy modification by agents	Policies are deployed by authorized operators only; agents have no write access to policy storage.
Audit trail modification	Append-only event log with cryptographic hash chain; no update or delete operations exist.
Path traversal attacks	All file paths are normalized, null bytes rejected, URL-encoded sequences decoded, and directory traversal blocked before evaluation.
Overly permissive policies	Mandatory minimum deny patterns are validated on every policy deployment; wildcard access triggers warnings.

Repudiation

Threat	Mitigation
Agent denies performing an action	Every evaluation is logged with timestamp, agent identity, decision, reasoning, and risk score.
Evidence tampering	Cryptographic hash chain provides tamper-evidence; each event's integrity is verifiable independently.
Compliance evidence gaps	Standardized export formats (OSCAL, STIX) provide machine-readable compliance evidence.

Information Disclosure

Threat	Mitigation
Credential/secret access	Mandatory deny patterns block common credential file patterns across all policies.
Token exposure	Secrets stored in managed secret stores (AWS Secrets Manager); never displayed in plaintext in outputs or logs.
Data in transit	TLS required for production deployments; internal traffic stays within VPC.
Error message leakage	Internal errors are masked before returning to clients; unique error IDs enable internal correlation.

Denial of Service

Threat	Mitigation
API flooding	Pluggable rate limiting with distributed backend support; WAF rate limiting at edge.
Resource exhaustion	Auto-scaling with configurable min/max; health checks detect failures within 30 seconds.
LLM provider outage	Only 5% of decisions require an LLM call; fail-closed returns deny on timeout; circuit breaker pattern prevents cascade.

Elevation of Privilege

Threat	Mitigation
Agent exceeds authority	Bounded authority model prevents self-elevation; policy changes require operator authentication.
Risk budget bypass	Risk budget is a one-way state machine that cannot be reset by agents or the enforcement system.
Enforcement system self-modifies	Meta-policy is immutable; the enforcement system operates within human-defined bounds.

4. Authentication & Access Control

Authentication Methods

Sentinel-Ops supports two authentication methods:

- **API Key Authentication** — A shared secret provided via secure environment configuration. Suitable for single-tenant deployments and evaluation environments.
- **JWT Authentication** — JSON Web Tokens with configurable secret and RBAC claims. Supports multi-tenant deployments with per-tenant isolation. Zero-downtime key rotation is supported via dual-secret configuration (current + previous secret accepted simultaneously during rotation window).

Both methods fail closed: requests without valid credentials are rejected immediately. The server will not start without at least one authentication method configured.

Role-Based Access Control

API endpoints are protected by role-based access:

Role	Capabilities
Operator	Full access: deploy policies, trigger evaluations, view audit trails, manage configuration.
Agent	Evaluate tool calls, read audit data, submit policy change requests.
Viewer	Read-only access to dashboards, risk state, and audit logs.

Role is extracted from JWT claims and enforced on every request.

Multi-Tenant Isolation

In multi-tenant deployments:

- Tenant identity is extracted from JWT claims and propagated through all system layers.
- Event storage is scoped per tenant — no shared storage paths between tenants.
- Real-time event streams (SSE) are tenant-scoped — clients only receive events from their own tenant.
- Rate limiting operates per-tenant to prevent noisy-neighbor effects.
- Policy evaluation is tenant-isolated — one tenant's policies cannot affect another's.

Secret Rotation

JWT secrets support automated rotation:

- **Dual-secret window** — During rotation, both current and previous secrets are accepted.

- **Automated schedule** — Configurable rotation period (default: 30 days).
- **Zero-downtime** — No service interruption during secret rotation.
- **Managed integration** — Works with AWS Secrets Manager rotation functions.

5. Data Protection

Data at Rest

Store	Encryption	Retention
Event database	AES-256 (managed encryption)	Point-in-Time Recovery enabled
Audit archives	AES-256 server-side encryption	Lifecycle: Standard → Infrequent Access → Archive
Secrets (API keys, tokens)	AES-256 via KMS	Automatic rotation support

Data in Transit

Path	Encryption
Client → Load Balancer	TLS 1.3 (certificate required for production)
Load Balancer → Application	Private VPC network
Application → Data Stores	HTTPS (SDK defaults)
Application → LLM Providers	HTTPS

Data Classification

Sentinel-Ops enforces data classification levels that determine minimum protection requirements:

- **Public** — Standard deny patterns.
- **Internal** — Additional internal document protection.
- **Confidential** — PII and financial data protection patterns.
- **Restricted** — PHI/HIPAA-level protection with maximum deny coverage.

Higher classification levels automatically inject stricter mandatory deny patterns into agent policies.

6. Audit Trail & Evidence

Immutable Event Log

Every policy evaluation generates an immutable audit event containing:

- Timestamp (ISO 8601)
- Agent identity
- Tool name and parameters
- Policy decision (allow/deny/escalate) with reasoning
- Risk score and cumulative budget state
- Request trace identifier
- Tenant identifier

Events are append-only. No update or delete operations exist in the audit subsystem. Even system administrators cannot modify historical events.

Tamper-Evidence

The audit log uses a cryptographic hash chain:

- Each event includes a hash computed from its content and the previous event's hash.
- Any modification to a historical event invalidates the chain from that point forward.
- Chain integrity can be verified independently at any time.
- Tamper detection is automated and generates alerts.

Compliance Reporting

Sentinel-Ops generates compliance evidence in industry-standard formats:

Format	Use Case
OSCAL (JSON)	Machine-readable compliance evidence for automated audit tools.
STIX	Threat intelligence format for security event correlation.
SOC 2 Reports	Control evidence for CC6.1 (Logical Access) and CC7.2 (Monitoring).
CSV/JSON Export	Full event history for custom analysis.

Request Tracing

Every API request is assigned a unique trace identifier that:

- Propagates through all system layers (API → enforcement → audit → sidecar).
- Appears in all log entries and audit events.
- Is returned in API responses for client-side correlation.
- Supports client-provided trace IDs for integration with existing observability stacks.

7. Content Guardrails

In addition to tool-call enforcement, Sentinel-Ops provides content-layer guardrails that scan data flowing to and from AI agents:

Scanner	Function
PII Detection	Identifies and blocks SSN, credit card numbers, phone numbers, and email addresses in agent data flows.
Prompt Injection	Detects and blocks prompt injection attempts targeting the underlying LLM.
Content Safety	Flags toxic content, harmful instructions, and policy-violating output.
Output Validation	Enforces format conformance on agent outputs.

Content guardrails operate independently from tool-call enforcement, providing an additional layer of defense for data-in-motion between agents and LLM providers.

8. Policy Simulation & Validation

Before deploying policies to production, Sentinel-Ops provides tools to test and validate:

- **Simulation** — Test a sequence of tool calls against a policy without enforcement, returning per-call decisions and aggregate risk budget impact.
- **Validation** — Verify policy syntax and semantics before deployment, catching errors and warnings before they affect live agents.
- **Policy generation** — Generate SOE policies from natural language descriptions or by introspecting existing AWS IAM role permissions, translating infrastructure-level constraints into agent-level policies.

Graduated Rollout Workflow

These capabilities support a graduated rollout workflow:

Mode	Behavior	Use Case
Shadow	Silent evaluation, minimal logging	Early testing and baseline measurement
Dry-run	Returns decision without enforcement	Policy development and tuning
Audit	Logs all decisions, nothing blocked	Initial production rollout (1–2 weeks)
Enforce	Blocks dangerous calls	Full production enforcement

Recommended rollout: Shadow (1–2 days) → Audit (1 week) → Enforce

9. Framework Compatibility

Sentinel-Ops integrates with any AI agent framework through two methods:

- **Sidecar proxy (zero code changes)** — A transparent proxy that intercepts all tool calls. No SDK, no vendor lock-in, no code modification required.
- **Direct API integration** — Call the evaluation API before tool execution for frameworks that support middleware or callback hooks.

Tested Integrations

Framework	Integration Method
Claude Code	Native PreToolUse hook (zero-config)
LangChain / LangGraph	Sidecar or API callback
CrewAI	Sidecar or tool wrapper
AutoGen	Sidecar or pipeline middleware
Anthropic Agent SDK	Sidecar or API call
Custom HTTP agents	Sidecar or direct API call

10. Network & Infrastructure Security

Network Architecture

- External access is restricted to a single load balancer with WAF protection.
- Internal services are not internet-accessible; they communicate within private VPC subnets.
- Security groups restrict ingress to load-balancer-originated traffic only.
- CIDR restrictions allow limiting access to specific IP ranges.
- WAF rules include rate limiting, managed rule sets for common attacks, and configurable custom rules.

Container Security

Practice	Implementation
Non-root execution	All containers run as unprivileged users with dedicated UIDs.
Minimal base images	Alpine-based with security patches applied at build time.
Multi-stage builds	Build dependencies are not present in runtime images.
Health checks	HTTP health endpoints with configurable intervals.
Vulnerability scanning	Automated scanning for CRITICAL and HIGH CVEs in CI pipeline.
SBOM generation	Software Bill of Materials in SPDX format for every image.
Image signing	Cryptographic signing via Sigstore/Cosign with transparency log recording.

Supply Chain Security

- **SBOM** — Software Bill of Materials generated for every container image in SPDX format.
- **Image signing** — Keyless signing via Sigstore/Fulcio with OIDC-based identity.
- **Transparency log** — All signatures recorded in an append-only public transparency log (Rekor).
- **Verification** — Customers can verify image authenticity and provenance using standard cosign tooling.

11. Rate Limiting & Availability

Pluggable Rate Limiting

Sentinel-Ops supports multiple rate limiting backends to match deployment scale:

Backend	Use Case	Characteristics
In-Memory	Single-instance deployments	Zero latency, no external dependencies.
Filesystem	Small multi-instance	Shared filesystem coordination.
Distributed (DynamoDB)	Production multi-instance	Atomic counters with TTL, no race conditions.

Rate limits are configurable per deployment: maximum calls per window, window duration (sliding), and per-tenant scoping to prevent noisy-neighbor effects.

Availability Design

- **Multi-AZ deployment** — Minimum two instances across availability zones.
- **Auto-scaling** — CPU-based target tracking with configurable min/max.
- **Health checks** — Load balancer health checks detect failures within 30 seconds.
- **Graceful shutdown** — Ordered subsystem flush ensures no data loss during deployments.
- **Circuit breaker** — LLM provider failures do not cascade; system falls back to deterministic-only evaluation.

Fail-Closed Guarantee

If any component in the enforcement chain fails:

Failure	Behavior
Authentication failure	Request rejected (401).
Policy load failure	All tool calls denied.
Evaluation engine failure	Tool call denied.
LLM timeout	Tool call denied (deterministic layers still process).
Rate limit backend failure	Configurable behavior (default: allow with logging).

The only intentional fail-open path is the rate limiting backend, which prioritizes availability over enforcement when the distributed counter is unreachable. This is the industry-standard approach and is documented with full

logging.

12. Compliance Framework Mapping

SOC 2 Type II

Trust Services Criteria	Control Coverage
CC6.1 Logical Access Controls	Per-agent identity boundaries, RBAC, API authentication.
CC7.2 System Operations Monitoring	Real-time event streaming, alarms, anomaly detection.

NIST 800-53 (Rev. 5)

Control	Description	Coverage
AC-2	Account Management	Agent identity registration and lifecycle.
AC-3	Access Enforcement	Deterministic deny/allow pattern matching.
AC-6	Least Privilege	Deny-by-default; minimum necessary permissions.
AU-2	Audit Events	All tool call evaluations logged with full context.
AU-6	Audit Review	Dashboard and standardized export for review.
AU-10	Non-Repudiation	Cryptographic hash chain on audit events.
AU-11	Audit Retention	Configurable lifecycle with archival support.
CA-7	Continuous Monitoring	Real-time event streams, alerting, anomaly detection.
SI-4	System Monitoring	Cross-agent correlation, risk tracking, threshold alerts.

HIPAA

Safeguard	Coverage
Access Controls (§164.312(a))	Per-agent data access boundaries with mandatory PHI deny patterns.
Audit Controls (§164.312(b))	Immutable audit trail with tamper-evidence.
Integrity Controls (§164.312(c))	Hash chain verification; append-only event storage.
Transmission Security (§164.312(e))	TLS 1.3 for all external communication.

EU AI Act

Requirement	Coverage
Risk Management (Art. 9)	Cumulative risk scoring with automatic tightening.
Data Governance (Art. 10)	Data access boundaries enforced per policy.
Transparency (Art. 13)	Full audit trail with decision reasoning.
Human Oversight (Art. 14)	Operator-defined policies; escalation to human for ambiguous cases.
Record-Keeping (Art. 12)	Immutable, cryptographically chained event log.

PCI-DSS

Requirement	Coverage
Req. 7 — Restrict Access	Role-based access control; least-privilege enforcement.
Req. 8 — Identify Users	Agent identity boundaries; JWT-based authentication.
Req. 10 — Track and Monitor	Complete audit trail of all agent actions.
Req. 11 — Test Security	Automated vulnerability scanning in CI pipeline.

13. Deployment Security

Infrastructure as Code

All infrastructure is defined as code (CloudFormation) with:

- **Single source of truth** — No manual configuration.
- **Version tracking** — Template version in stack outputs.
- **Deployment circuit breaker** — Automatic rollback on failure.
- **Parameter validation** — Constraints enforce valid configuration.

Production Requirements

Production deployments enforce:

- TLS certificate required.
- WAF enabled.
- Multi-AZ deployment (minimum 2 instances).
- Authentication configured.
- Secrets stored in managed secret store.

Evaluation Mode

For evaluation and development:

- HTTP permitted (with security warnings).
- Single-instance deployment allowed.
- Violations logged but not blocked (audit mode).
- Full enforcement can be enabled at any time.

14. Shared Responsibility Model

Area	Sentinel-Ops (YadriWorks Inc.)	Customer
Enforcement engine	Provided and maintained.	—
Policy definition	Templates and generation tools.	Review, customize, and approve.
Infrastructure	Uses customer’s cloud account.	Provide VPC, network, and IAM.
TLS certificates	Template configures listeners.	Provision certificates.
Secret rotation	Supports automated rotation.	Configure rotation schedule.
LLM API keys	Stored securely; never logged.	Provide and maintain.
Monitoring	Alarms, dashboards, event streams.	Configure alert destinations.
Compliance reporting	Evidence generation tools.	Review and submit to auditors.
Incident response	Detection and alerting.	Investigation and remediation.

15. Incident Response

Automated Detection

Signal	Detection	Response Time
Policy violation	Real-time event stream	< 1 second
Risk threshold breach	State machine transition alert	< 1 second
Cross-agent anomaly	Correlation engine	< 60 seconds
Infrastructure degradation	Cloud-native alarms	< 3 minutes

Alert Channels

- **Event Bus** — Native cloud event routing (Amazon EventBridge) with typed events: SOE.ToolAllow, SOE.ToolDeny, SOE.ToolEscalate, SOE.RiskBudgetChange, SOE.AnomalyDetected, SOE.EnvelopeTightened.
- **Dashboard** — Real-time visualization of all agent activity and security events.
- **Prometheus metrics** — Standard `/metrics` endpoint exposing evaluation counts, latency histograms, API request rates, and guardrail block rates for integration with existing monitoring stacks.

Evidence Preservation

- All events are immutable with cryptographic integrity verification.
- Audit logs archived with versioning and lifecycle management.
- Point-in-time recovery available for event databases.
- Hash chain verification confirms trail integrity at any point in time.

16. Security Testing & CI/CD

Automated Security Testing

The CI/CD pipeline includes:

Test Category	Coverage
Authentication tests	Fail-closed behavior, invalid tokens, missing credentials.
Tenant isolation tests	Cross-tenant prevention, scoped storage, scoped event streams.
Request tracing tests	End-to-end trace propagation, ID generation, client ID passthrough.
Graceful shutdown tests	Ordered subsystem flush, connection rejection after close.
Policy evaluation tests	Deterministic matching, deny-wins, mandatory patterns.
Rate limiting tests	All backends, boundary conditions, distributed atomicity.
Product claims tests	64 verifiable product claims tested on every commit.

Container Security Pipeline

Stage	Tool	Threshold
Build	Multi-stage Docker builds	—
Scan	Trivy vulnerability scanner	CRITICAL + HIGH (blocking)
SBOM	Syft (Anchore)	SPDX format for every image
Sign	Cosign (keyless/Sigstore)	Transparency log recorded
Attest	Cosign SBOM attestation	In-toto format attached to image

17. Conclusion

Sentinel-Ops provides a defense-in-depth security architecture for governing AI agent actions in production environments. The combination of:

- Deterministic enforcement eliminating AI variability from 95% of decisions
- Fail-closed defaults ensuring safety under all failure conditions
- Cryptographic audit trail providing tamper-evident compliance evidence
- Multi-tenant isolation enabling secure shared deployments
- Bounded authority preventing privilege escalation at every layer

positions Sentinel-Ops as a production-grade governance platform suitable for regulated industries including financial services, healthcare, government, and any environment where AI agent actions must be controlled, audited, and compliant.

18. Legal Disclaimers & Notices

Confidentiality Notice

This document is the property of YadriWorks Inc. It may be freely distributed for informational purposes. Commercial use, modification, or redistribution of this document requires the prior written consent of YadriWorks Inc.

No Warranty

This document and the information contained herein are provided "AS IS" without warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose, or non-infringement. YadriWorks Inc. makes no representations or warranties regarding the accuracy, completeness, or reliability of the information presented. The security architecture, features, and capabilities described herein are subject to change without notice.

Limitation of Liability

In no event shall YadriWorks Inc. its officers, directors, employees, or agents be liable for any direct, indirect, incidental, special, consequential, or punitive damages arising out of or in connection with the use of, or reliance on, any information contained in this document, whether based on warranty, contract, tort (including negligence), or any other legal theory, even if advised of the possibility of such damages.

Compliance Disclaimer

The compliance framework mappings presented in this document (SOC 2 Type II, NIST 800-53, HIPAA, EU AI Act, PCI-DSS) are provided for informational purposes only and represent the vendor’s assessment of how Sentinel-Ops capabilities align with the referenced frameworks. These mappings do not constitute a certification, attestation, or guarantee of compliance. Customers are solely responsible for determining whether their use of Sentinel-Ops satisfies their own regulatory, legal, and compliance obligations. YadriWorks Inc. recommends that customers engage qualified legal counsel, auditors, and compliance professionals to evaluate their specific requirements.

Third-Party Services

Sentinel-Ops integrates with third-party services including Amazon Web Services (AWS), LLM providers, and other cloud-based platforms. YadriWorks Inc. does not control and is not responsible for the security practices, availability, or performance of any third-party service. Customers are responsible for reviewing and accepting the terms of service and security posture of any third-party service used in conjunction with Sentinel-Ops.

Data Residency

Sentinel-Ops is deployed within the customer’s own AWS account and region. Customer data, including audit logs, policy definitions, and evaluation events, remains within the customer’s infrastructure. The only data transmitted externally is aggregate usage counts to AWS Marketplace for billing purposes. No telemetry, diagnostic data, or customer content is collected or transmitted by YadriWorks Inc. Customers are responsible for ensuring their deployment configuration meets applicable data residency requirements.

Intellectual Property

Sentinel-Ops, the Safe Operating Envelope concept, and all related documentation are the intellectual property of YadriWorks Inc. All trademarks, service marks, trade names, and logos referenced herein are the property of their respective owners. Nothing in this document grants any license or right to use any trademark, service mark, or logo of YadriWorks Inc. or any third party.

Forward-Looking Statements

This document may contain forward-looking statements regarding planned features, capabilities, and product direction. These statements are based on current expectations and are subject to risks and uncertainties that could cause actual results to differ materially. Forward-looking statements should not be relied upon as a guarantee of future performance or functionality.

Responsible Disclosure

If you discover a security vulnerability in Sentinel-Ops, please report it to security@yadriworks.ai. YadriWorks Inc. is committed to addressing security issues promptly and will acknowledge receipt of vulnerability reports within two business days. We request that you allow reasonable time for remediation before any public disclosure.

YadriWorks Inc.
security@yadriworks.ai
https://yadriworks.ai