SNAPWIRE

Public Safety Disclosure Report

Generated: May 23, 2026 at 11:18 UTC

1. NIST Grade & Coverage Score

C

NIST IR 8596 Coverage: 42%

12 categories | 5 covered | 0 partial | 7 gaps

Readiness Grade: Developing - Based on CSF 2.0 function mapping of installed Snapwire rule packs.

2. Coverage Breakdown

ID	Function	Category	Status	Packs
GOVERN-1.1	GOVERN	Legal and Regulatory Requirements	Covered	universal_start...
GOVERN-1.2	GOVERN	Trustworthy AI Characteristics	Covered	universal_starter
GOVERN-6.1	GOVERN	Deployment Policies	Gap	-
MAP-3.4	MAP	Data Constraints	Gap	-
MAP-3.5	MAP	Scientific Integrity	Gap	-
MAP-5.1	MAP	Impacts to People	Gap	-
MEASURE-2.6	MEASURE	Validity and Reliability	Covered	universal_starter
MANAGE-2.2	MANAGE	Risk Tracking Mechanisms	Covered	sql_redline, sh...
MANAGE-2.3	MANAGE	Risk Assessment Procedures	Gap	-
MANAGE-2.4	MANAGE	Risk Response and Recovery	Covered	shell_safety
MANAGE-4.1	MANAGE	Post-Deployment Monitoring	Gap	-
RESPOND-1.1	RESPOND	Incident Response	Gap	-

3. Intended Use & Boundaries

Snapwire is an Agentic Runtime Security (ARS) platform designed to sit between AI agents and their tool-call targets. It functions as a runtime security layer (reverse proxy) that intercepts, evaluates, and enforces policy on every tool call an autonomous agent makes.

- Real-time interception and policy enforcement for AI agent tool calls
- Human-in-the-loop review queue for high-risk or ambiguous actions
- Constitutional rule engine with severity-based blocking and monitoring
- Immutable forensic audit trail of all agent decisions and delegations

Snapwire does not generate, modify, or assume responsibility for the underlying intent or output of the AI model. It operates as a passive security intermediary.

4. Foreseeable Misuse & Mitigation

- CVE-2026-25253 (OpenClaw): BASE_URL redirect, credential exfiltration, domain spoofing, WebSocket hijacking, and environment variable injection are detected and blocked
- Hallucination loops: The Fuse Breaker (Loop Detector) identifies and kills repetitive tool-call patterns before they drain budgets
- Prompt injection via tool parameters: Input sanitization strips injection attempts from agent-supplied parameters
- Credential theft: The Identity Vault ensures agents never see raw secrets; Snap-Tokens are used as proxies
- Unauthorized escalation: Blast Radius controls and Honeypot tripwires detect and contain rogue agent behavior

5. Human Accountability Statement

- X-Snapwire-Authorized-By header injected into every proxied request for immutable operator attribution
- X-Snapwire-Origin-ID header traces every request back to its originating Snapwire instance
- All blocked, approved, and pending decisions logged with timestamps, agent IDs, and operator context
- The final Duty of Care for all agent actions and budgetary releases remains solely with the human operator

6. Algorithmic Discrimination Protections

- Constitutional Auditor: Every tool call evaluated against configurable constitutional rules encoding equity and fairness policies
- Equity-aware rule templates: Pre-built rule packs include data protection rules to prevent PII leakage and discriminatory data handling
- Observe & Audit Mode: New rules tested in observation mode before enforcement, preventing unintended discriminatory blocking
- Community Rules: Open, peer-reviewed rule contributions ensure diverse perspectives in governance policy
- Deception Detection: Heuristic analysis identifies agent circumvention of safety rules through obfuscation

7. Compliance Standards

- NIST IR 8596 - AI Agent Security Profile (CSF 2.0)
- Colorado SB24-205 - AI Consumer Protections
- Singapore Model Governance Framework v1.1
- OWASP Top 10 for LLM Applications

8. Active Safeguards

This instance has 13 active safeguards:

1. Constitutional Rule Engine
2. OpenClaw CVE-2026-25253 Safeguard
3. Loop Detector (Fuse Breaker)
4. Input Sanitizer
5. Blast Radius Controls
6. Honeypot Tripwires
7. Identity Vault (Snap-Tokens)
8. Tool Safety Catalog
9. Deception Detector
10. Schema Guard

11. Risk Index Scoring
12. Thinking Token Sentinel
13. Rate Limiter

9. High-Stakes Tools (Consequentiality Tagging)

No tools have been tagged as high-stakes (consequential). Use the Tool Catalog dashboard to identify and tag tools with potential for substantial consumer impact.

10. Audit Log Fingerprint

SHA-256 fingerprint of the most recent 100 audit log entries:

`eccd42103021ff63fd06adfb7907cae39324a8425381f9e04293f19418b01b4b`